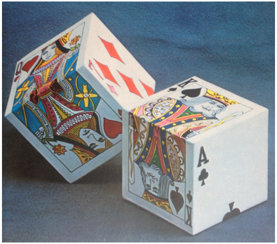
Fundamentos Estatísticos para Ciência dos Dados
Introdução ao curso

Renato Martins Assunção

DCC, UFMG - 2017



Diferença entre probabilidade e estatística

- Probabilidade: um ramo da matemática pura.

Diferença entre probabilidade e estatística

- Probabilidade: um ramo da matemática pura.
- Ela permite fazer cálculos matemáticos sobre fenômenos aleatórios.

Diferença entre probabilidade e estatística

- Probabilidade: um ramo da matemática pura.
- Ela permite fazer cálculos matemáticos sobre fenômenos aleatórios.
- Não precisa de dados estatísticos.

Diferença entre probabilidade e estatística

- Probabilidade: um ramo da matemática pura.
- Ela permite fazer cálculos matemáticos sobre fenômenos aleatórios.
- Não precisa de dados estatísticos.
- Como funciona: estabeleça um modelo probabilístico.

Diferença entre probabilidade e estatística

- Probabilidade: um ramo da matemática pura.
- Ela permite fazer cálculos matemáticos sobre fenômenos aleatórios.
- Não precisa de dados estatísticos.
- Como funciona: estabeleça um modelo probabilístico.
- A seguir, calcule probabilidades de eventos de interesse.

Sequência mais longa

- Jogue moeda para cima 100 vezes

Sequência mais longa

- Jogue moeda para cima 100 vezes
- Qual a chance de observar uma sequência de 8 ou mais caras em seguida?

Sequência mais longa

- Jogue moeda para cima 100 vezes
- Qual a chance de observar uma sequência de 8 ou mais caras em seguida?
- É um cálculo matemático.
- Não precisa realizar o experimento físico para calcular as chances.

Sequência mais longa

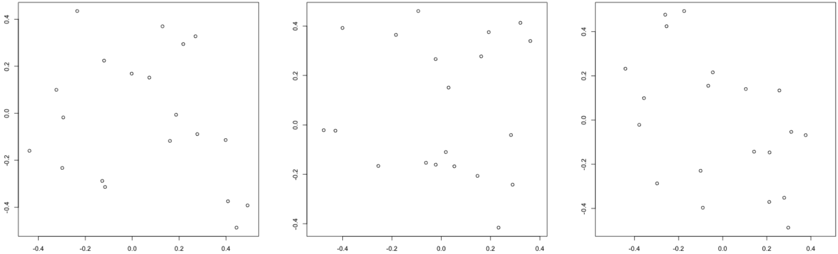
- Jogue moeda para cima 100 vezes
- Qual a chance de observar uma sequência de 8 ou mais caras em seguida?
- É um cálculo matemático.
- Não precisa realizar o experimento físico para calcular as chances.
- Mas para quê isto?
- Hot hand em esportes.
- Pontos sucessivos para time A ou B como se fossem cara-coroa de uma moeda.
- Mas probabilidade de A varia ao longo do jogo: às vezes, fica quente.
- Isto é verdade? Até que ponto esta variação pode ocorrer?

Probabilidade: modelo espacial 1

- n pontos são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.

Probabilidade: modelo espacial 1

- n pontos são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.
- Veja três realizações independentes deste experimento.



Probabilidade: modelo espacial 1

- Qual a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$?

Probabilidade: modelo espacial 1

- Qual a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$?
- Se $r \approx 0$, $\mathbb{P}_1(r)$ deve ser próxima de 1.

Probabilidade: modelo espacial 1

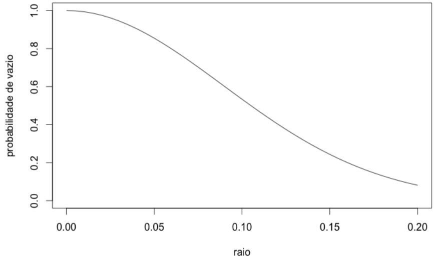
- Qual a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$?
- Se $r \approx 0$, $\mathbb{P}_1(r)$ deve ser próxima de 1.
- Quando r aumenta, $\mathbb{P}_1(r)$ deve decrescer para zero.

Probabilidade: modelo espacial 1

- Qual a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$?
- Se $r \approx 0$, $\mathbb{P}_1(r)$ deve ser próxima de 1.
- Quando r aumenta, $\mathbb{P}_1(r)$ deve decrescer para zero.
- Pode-se mostrar que $\mathbb{P}_1(r)$ é aproximadamente igual a $\exp(-n\pi r^2)$.

Probabilidade: modelo espacial 1

- Qual a probabilidade $\mathbb{P}_1(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0,0)$?
- Se $r \approx 0$, $\mathbb{P}_1(r)$ deve ser próxima de 1.
- Quando r aumenta, $\mathbb{P}_1(r)$ deve decrescer para zero.
- Pode-se mostrar que $\mathbb{P}_1(r)$ é aproximadamente igual a $\exp(-n\pi r^2)$. Gráfico com $n = 20$.



Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.
- A figura com três configurações de pontos é apenas ilustrativa.

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.
- A figura com três configurações de pontos é apenas ilustrativa. Ela não foi usada no cálculo de $\mathbb{P}_1(r)$.

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.
- A figura com três configurações de pontos é apenas ilustrativa. Ela não foi usada no cálculo de $\mathbb{P}_1(r)$.
- Para este modelo de pontos aleatórios, várias outras probabilidades podem ser calculadas.

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.
- A figura com três configurações de pontos é apenas ilustrativa. Ela não foi usada no cálculo de $\mathbb{P}_1(r)$.
- Para este modelo de pontos aleatórios, várias outras probabilidades podem ser calculadas.
- Por exemplo, qual a probabilidade de que existam pelo menos 2 pontos numa certa região de área α?

Probabilidade: modelo espacial 1

- A probabilidade $\mathbb{P}_1(r)$ é calculada SEM DADOS.
- É um cálculo matemático.
- A figura com três configurações de pontos é apenas ilustrativa. Ela não foi usada no cálculo de $\mathbb{P}_1(r)$.
- Para este modelo de pontos aleatórios, várias outras probabilidades podem ser calculadas.
- Por exemplo, qual a probabilidade de que existam pelo menos 2 pontos numa certa região de área α?
- É aproximadamente
$$1 - e^{-n\alpha}(1 + n\alpha)$$
- Não é preciso coletar nenhum dado para fazer estes cálculos.

Mudando o modelo probabilístico

- Outro modelo probabilístico para geração de pontos no quadrado leva a resultados bem diferentes no cálculo de probabilidade.

Mudando o modelo probabilístico

- Outro modelo probabilístico para geração de pontos no quadrado leva a resultados bem diferentes no cálculo de probabilidade.
- Por exemplo, suponha que apenas 5 pontos-pais são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.

Mudando o modelo probabilístico

- Outro modelo probabilístico para geração de pontos no quadrado leva a resultados bem diferentes no cálculo de probabilidade.
- Por exemplo, suponha que apenas 5 pontos-pais são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.
- A seguir, cada ponto-pai gera 4 pontos-filhos de forma que temos 20 pontos-filho no final.

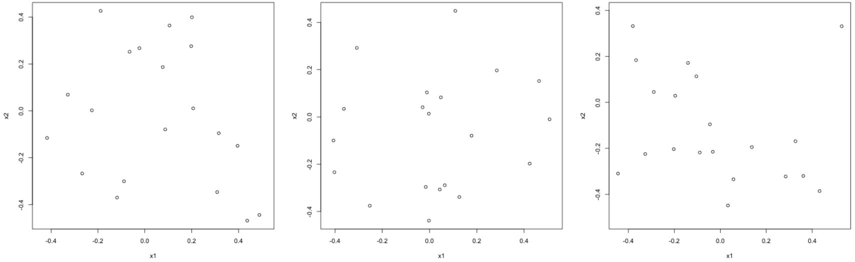
Mudando o modelo probabilístico

- Outro modelo probabilístico para geração de pontos no quadrado leva a resultados bem diferentes no cálculo de probabilidade.
- Por exemplo, suponha que apenas 5 pontos-pais são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.
- A seguir, cada ponto-pai gera 4 pontos-filhos de forma que temos 20 pontos-filho no final.
- Os filhos espalham-se ao acaso em torno dos pais até uma distância máxima de 0.1

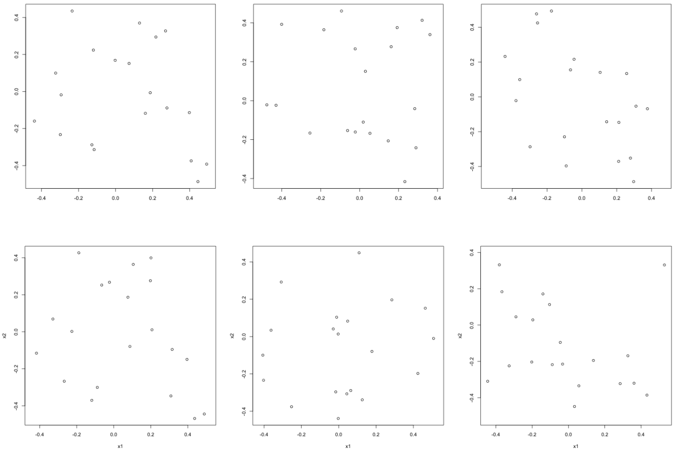
Mudando o modelo probabilístico

- Outro modelo probabilístico para geração de pontos no quadrado leva a resultados bem diferentes no cálculo de probabilidade.
- Por exemplo, suponha que apenas 5 pontos-pais são jogados completamente ao acaso no quadrado de área 1 centrado na origem $(0, 0)$.
- A seguir, cada ponto-pai gera 4 pontos-filhos de forma que temos 20 pontos-filho no final.
- Os filhos espalham-se ao acaso em torno dos pais até uma distância máxima de 0.1
- Considere o padrão espacial dos pontos compostos apenas pelos filhos.

Probabilidade: modelo espacial 2



Modelo 1 (1^a linha) e modelo 2 (2^a linha)

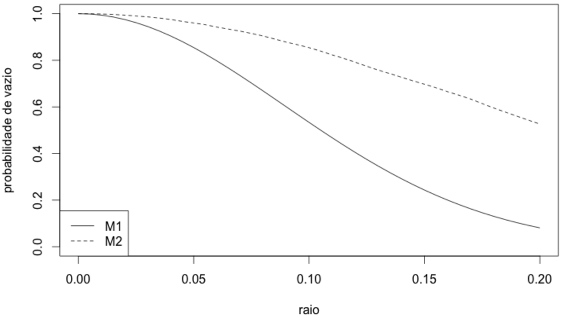


Probabilidade: modelo espacial 2

- No modelo 2, qual a probabilidade $\mathbb{P}_2(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0,0)$?

Probabilidade: modelo espacial 2

- No modelo 2, qual a probabilidade $\mathbb{P}_2(r)$ de que não exista nenhum ponto num raio r em torno da origem $(0, 0)$?
- Como no modelo 1, temos $\mathbb{P}_2(r) \approx 1$ e diminuindo para zero quando o raio r aumenta.
- Mas ela faz isto de forma bem diferente nos dois modelos.



Estatística: dados, dados, dados...

- Estatística: um ramo da matemática aplicada.

Estatística: dados, dados, dados...

- Estatística: um ramo da matemática aplicada.
- Precisa de dados estatísticos. O que fazemos com esses dados?

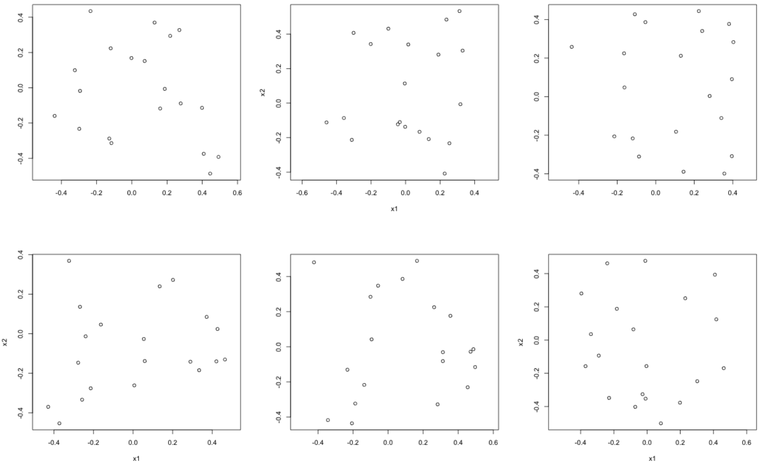
Estatística: dados, dados, dados...

- Estatística: um ramo da matemática aplicada.
- Precisa de dados estatísticos. O que fazemos com esses dados?
- Procuramos inferir qual foi o modelo probabilístico que gerou os dados observados.

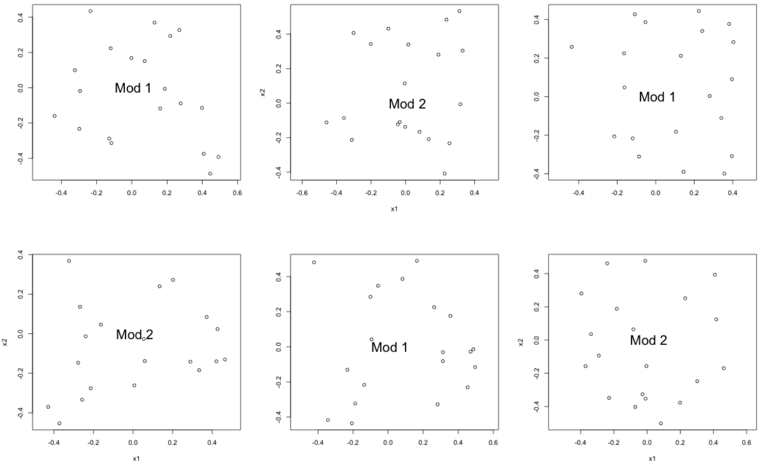
Estatística: dados, dados, dados...

- Estatística: um ramo da matemática aplicada.
- Precisa de dados estatísticos. O que fazemos com esses dados?
- Procuramos inferir qual foi o modelo probabilístico que gerou os dados observados.
- Qual dos dois modelos gerou cada um dos seis plots a seguir?

Qual modelo gerou cada plot?



Identificando os modelos



Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
  - Por meio do cálculo de probabilidades,

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
  - Por meio do cálculo de probabilidades,
  - sem usar dados estatísticos,

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
    - Por meio do cálculo de probabilidades,
    - sem usar dados estatísticos,
    - com matemaática pura (ou por simulação Monte Carlo)

Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
  - Por meio do cálculo de probabilidades,
  - sem usar dados estatísticos,
  - com matemaática pura (ou por simulação Monte Carlo)
  - obtive limites (m, M) tais que,

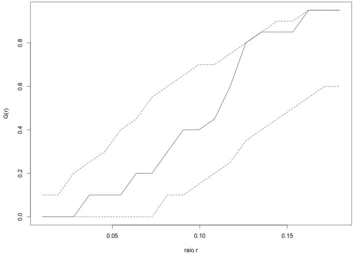
Um teste estatístico para discriminar entre os modelos

- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
    - Por meio do cálculo de probabilidades,
    - sem usar dados estatísticos,
    - com matemaática pura (ou por simulação Monte Carlo)
    - obtive limites (m, M) tais que,
    - se os dados vierem de fato do modelo 1,

Um teste estatístico para discriminar entre os modelos

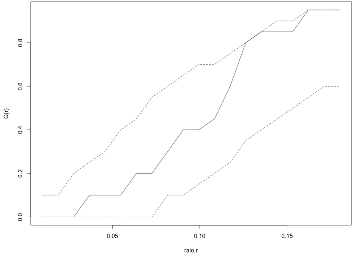
- Para cada ponto, achei a distância até o seu ponto vizinho mais próximo.
- Para vários raios r, contei a proporção de pontos que tiveram distância menor que r.
- Por exemplo, para $r = 0.10$, obtive a proporção G de pontos observados que tiveram seu vizinho mais próximo a uma distância menor que 0.10.
- Então:
  - Por meio do cálculo de probabilidades,
  - sem usar dados estatísticos,
  - com matemaática pura (ou por simulação Monte Carlo)
  - obtive limites (m, M) tais que,
  - se os dados vierem de fato do modelo 1,
  - o valor de G deveria estar entre m e M com probabilidade muito alta.
  - Se estiver fora dos limites, o modelo 2 deve ser o correto.

Um teste estatístico para discriminar entre os modelos



Eixo horizontal: raio r.

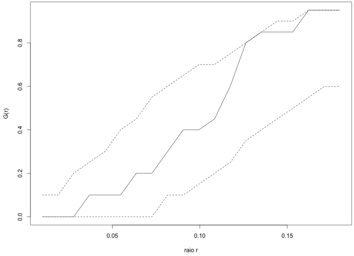
Um teste estatístico para discriminar entre os modelos



Eixo horizontal: raio r.

Eixo vertical: $G(r) =$ probab da distância ao vizinho mais próximo ser menor que r.

Um teste estatístico para discriminar entre os modelos

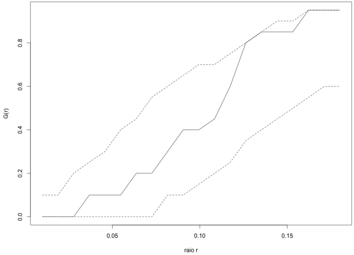


Eixo horizontal: raio r.

Eixo vertical: $G(r) =$ probab da distância ao vizinho mais próximo ser menor que r.

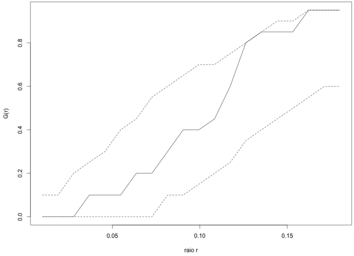
Linhas tracejadas:
limites para $G(r)$ versus raio r
CASO MODELO 1 SEJA CORRETO.

Um teste estatístico para discriminar entre os modelos



Linhas tracejadas foram obtidos
com cálculo de probabilidades, sem dados.

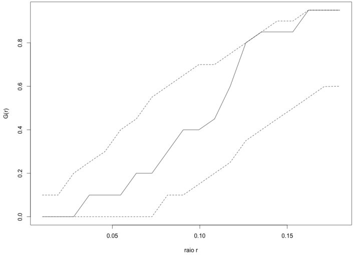
Um teste estatístico para discriminar entre os modelos



Linhas tracejadas foram obtidos
com cálculo de probabilidades, sem dados.

Curva contínua: proporção $G(r)$
calculada com os dados estatísticos.

Um teste estatístico para discriminar entre os modelos



Linhas tracejadas foram obtidos
com cálculo de probabilidades, sem dados.

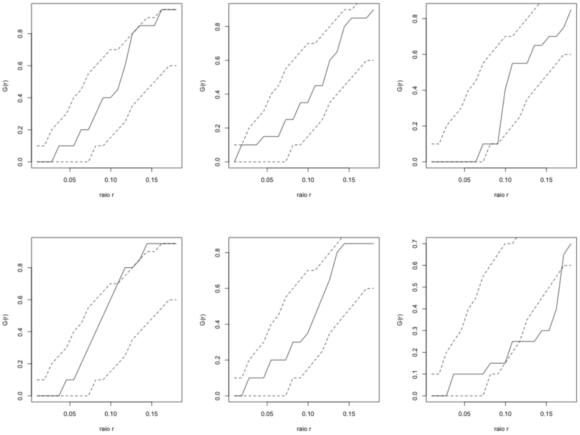
Curva contínua: proporção $G(r)$
calculada com os dados estatísticos.

Se ficar dentro dos limites,
fique com o modelo 1.

Se sair fora dos limites,
fique com o modelo 2.

Decisão: poucos erros

Decisão errada apenas no plot $(1, 2)$ cujos dados são do modelo 2.



Resumo

- **Probabilidade:** a partir de um modelo probabilístico, calcula matematicamente a probabilidade de diversos eventos (ou dados).

Resumo

- **Probabilidade:** a partir de um modelo probabilístico, calcula matematicamente a probabilidade de diversos eventos (ou dados).
- Não precisa ter nenhum dado estatístico para isto.

Resumo

- **Probabilidade:** a partir de um modelo probabilístico, calcula matematicamente a probabilidade de diversos eventos (ou dados).
- Não precisa ter nenhum dado estatístico para isto.
- **Estatística:** possui dados observados, uma tabela de números.

Resumo

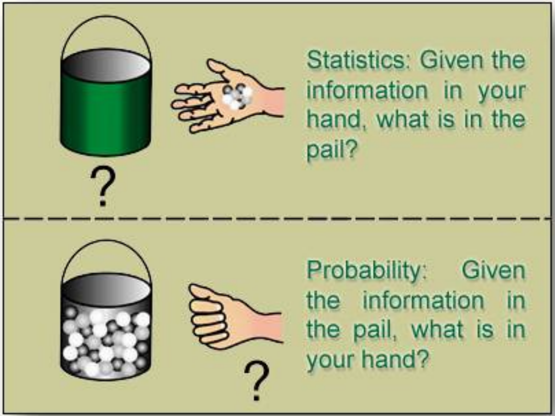
- **Probabilidade:** a partir de um modelo probabilístico, calcula matematicamente a probabilidade de diversos eventos (ou dados).
- Não precisa ter nenhum dado estatístico para isto.
- **Estatística:** possui dados observados, uma tabela de números.
- Deseja descobrir qual foi o modelo probabilístico que gerou estes dados.

Resumo

- **Probabilidade:** a partir de um modelo probabilístico, calcula matematicamente a probabilidade de diversos eventos (ou dados).
- Não precisa ter nenhum dado estatístico para isto.
- **Estatística:** possui dados observados, uma tabela de números.
- Deseja descobrir qual foi o modelo probabilístico que gerou estes dados.

Estatística versus probabilidade em imagens

Extraído de `http://herdingcats.typepad.com/my_weblog/`



Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.
- Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.
- Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado.
- Precisamos de um modelo de probabilidade para fazer estes cálculos.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.
- Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado.
- Precisamos de um modelo de probabilidade para fazer estes cálculos.
- Existem muitos (infinitos) modelos possíveis.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.
- Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado.
- Precisamos de um modelo de probabilidade para fazer estes cálculos.
- Existem muitos (infinitos) modelos possíveis.
- Alguns são melhores que outros pois conseguem prever melhor que cada cliente vai fazer.

Risco de crédito: dados e modelo probabilístico

- Clientes solicitam crédito ou tomam empréstimo em bancos.
- Bancos querem saber, para cada cliente, se ele vai pagar de volta dentro do prazo.
- Um modelo de risco de crédito avalia a probabilidade disso ocorrer DADO que o cliente possui certos atributos.
- Se a probabilidade for baixa, ele é um risco potencial e o crédito deveria ser negado.
- Precisamos de um modelo de probabilidade para fazer estes cálculos.
- Existem muitos (infinitos) modelos possíveis.
- Alguns são melhores que outros pois conseguem prever melhor que cada cliente vai fazer.
- Quais os dados para identificar um modelo desses?

Risco de crédito: dados típicos

- Dados de 1000 clientes de um banco que pegaram empréstimo no passado.

Risco de crédito: dados típicos

- Dados de 1000 clientes de um banco que pegaram empréstimo no passado.
- Para cada cliente, anota-se uma resposta binária Y.

Risco de crédito: dados típicos

- Dados de 1000 clientes de um banco que pegaram empréstimo no passado.
- Para cada cliente, anota-se uma resposta binária Y.
- $Y = 1$ se pagou de volta no devido tempo.

Risco de crédito: dados típicos

- Dados de 1000 clientes de um banco que pegaram empréstimo no passado.
- Para cada cliente, anota-se uma resposta binária Y.
- $Y = 1$ se pagou de volta no devido tempo.
- $Y = 0$ case contrário.

Risco de crédito: dados típicos

- Dados de 1000 clientes de um banco que pegaram empréstimo no passado.
- Para cada cliente, anota-se uma resposta binária Y.
- $Y = 1$ se pagou de volta no devido tempo.
- $Y = 0$ case contrário.
- Além disso, temos 20 atributos que podem influenciar o comportamento dos clientes.

Risco de crédito: dados típicos

- Balance of current account
- For how long has been a client (in months)
- Payment of previous credits: *no previous credits/paid back all previous credits*; *hesitant payment of previous credits*; *problematic running account*.
- Purpose of credit: *new car*; *used car*; *items of furniture*; *vacation*; etc.
- Amount of credit.
- Value of savings or stocks.
- For how has been employed by current employer (in years).
- Installment in % of available income
- Marital Status
- Sex
- Age, etc.

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.
- Se esta sub-amostra não for muito pequena ...

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.
- Se esta sub-amostra não for muito pequena ... (digamos, maior que 100 indivíduos) ...

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.
- Se esta sub-amostra não for muito pequena ... (digamos, maior que 100 indivíduos) ...
- Dentre os indivíduos dessa sub-amostra, obtenha a proporção dos que não pagaram o crédito.

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.
- Se esta sub-amostra não for muito pequena ... (digamos, maior que 100 indivíduos) ...
- Dentre os indivíduos dessa sub-amostra, obtenha a proporção dos que não pagaram o crédito.
- Esta proporção é aproximadamente a probabilidade de não-pagamento.

Inferindo diretamente a partir dos dados

- Precisamos mesmo de um modelo probabilístico?
- Nos dias de big data, os dados não respondem tudo?
- Afinal, podemos fazer cálculos diretos e simples a partir dos dados diretamente.
- Por exemplo, qual a probabilidade de um cliente com mais de 60 anos e saldo médio maior que 5 mil reais não pagar o crédito?
- Separe a sub-amostra de clientes com mais de 60 anos e saldo maior que 5 mil.
- Se esta sub-amostra não for muito pequena ... (digamos, maior que 100 indivíduos) ...
- Dentre os indivíduos dessa sub-amostra, obtenha a proporção dos que não pagaram o crédito.
- Esta proporção é aproximadamente a probabilidade de não-pagamento.
- Muito simples, apenas contagem no banco de dados.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.
- Isto dá 3276800, ou mais de 3 milhões de indivíduos na base de dados.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.
- Isto dá 3276800, ou mais de 3 milhões de indivíduos na base de dados.
- Será difícil obter uma base de dados relativamente recentes desta forma para este problema.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.
- Isto dá 3276800, ou mais de 3 milhões de indivíduos na base de dados.
- Será difícil obter uma base de dados relativamente recentes desta forma para este problema.
- Suponha que não exista na base de dados NENHUM indivíduo com idade x, saldo y, etc.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.
- Isto dá 3276800, ou mais de 3 milhões de indivíduos na base de dados.
- Será difícil obter uma base de dados relativamente recentes desta forma para este problema.
- Suponha que não exista na base de dados NENHUM indivíduo com idade x, saldo y, etc.
- Ou quem sabe existam apenas 3 indivíduos com estes atributos.

Nem sempre é tão simples

- O cliente tem muitos atributos, não apenas idade e saldo mëdio.
- Para cada cliente, temos mais de 15 atributos.
- Se cada atributo possui apenas dois valores possíveis, temos $2^{15} = 32768$ configurações diferentes de atributos para os clientes.
- Em cada uma dessas 32 mil configurações possíveis, queremos a probabilidade de não pagamento.
- Precisamos de pelo menos uns 100 indivíduos em cada configuração para estimar a probabilidade.
- Isto dá 3276800, ou mais de 3 milhões de indivíduos na base de dados.
- Será difícil obter uma base de dados relativamente recentes desta forma para este problema.
- Suponha que não exista na base de dados NENHUM indivíduo com idade x, saldo y, etc.
- Ou quem sabe existam apenas 3 indivíduos com estes atributos.
- Como estimar bem a probabilidade de não pagamento de um novo cliente com estes atributos?

Nem sempre é tão simples

- Perdas financeiras associadas com tufões em Taiwan.

Nem sempre é tão simples

- Perdas financeiras associadas com tufões em Taiwan.
- Qual a probabilidade de ocorrer um tufão causando perda maior que 4 milhões nos próximos 10 anos?
- Zero?

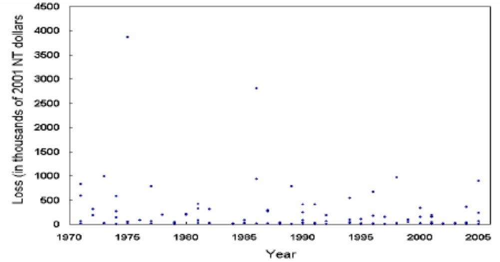


Figure 1. Scatter plot of Taiwan typhoon rice loss

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.
- Suponha que o experimento precisa fornecer uma estimativa um anos após o início do estudo.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.
- Suponha que o experimento precisa fornecer uma estimativa um anos após o início do estudo.
- Um ano após o estudo, 50% dos pacientes faleceram (e portanto sabe-se o valor de T para estes indivíduos).

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.
- Suponha que o experimento precisa fornecer uma estimativa um anos após o início do estudo.
- Um ano após o estudo, 50% dos pacientes faleceram (e portanto sabe-se o valor de T para estes indivíduos).
- Mas 50% ainda não faleceram e não se conhece T para estes outros indivíduos.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.
- Suponha que o experimento precisa fornecer uma estimativa um anos após o início do estudo.
- Um ano após o estudo, 50% dos pacientes faleceram (e portanto sabe-se o valor de T para estes indivíduos).
- Mas 50% ainda não faleceram e não se conhece T para estes outros indivíduos.
- A média dos valores conhecidos vai tender a subestimar o valor esperado de sobrevida.

Mais um exemplo

- Dados $T_1, T_2, \ldots, T_n$: o tempo de sobrevida de n pacientes submetidos a um novo tratamento médico.
- Deseja-se estimar o tempo esperado $\mathbb{E}(T)$ de sobrevida após o tratamento.
- Simples: tire a média aritmética dos n tempos observados.
- Suponha que o experimento precisa fornecer uma estimativa um anos após o início do estudo.
- Um ano após o estudo, 50% dos pacientes faleceram (e portanto sabe-se o valor de T para estes indivíduos).
- Mas 50% ainda não faleceram e não se conhece T para estes outros indivíduos.
- A média dos valores conhecidos vai tender a subestimar o valor esperado de sobrevida.
- Como fazer neste caso?

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.
- **Modelo Estatístico Conceitual:** Uma distribuição de probabilidade **hipotética** descrevendo como os dados observados **poderiam** ter sido gerados.

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.
- **Modelo Estatístico Conceitual:** Uma distribuição de probabilidade **hipotética** descrevendo como os dados observados **poderiam** ter sido gerados.
- A modelagem é a concepção de um arcabouço matemático capaz de gerar os dados.

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.
- **Modelo Estatístico Conceitual:** Uma distribuição de probabilidade **hipotética** descrevendo como os dados observados **poderiam** ter sido gerados.
- A modelagem é a concepção de um arcabouço matemático capaz de gerar os dados.
- Os dados que nos interessam não são determinísticos.

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.
- **Modelo Estatístico Conceitual:** Uma distribuição de probabilidade **hipotética** descrevendo como os dados observados **poderiam** ter sido gerados.
- A modelagem é a concepção de um arcabouço matemático capaz de gerar os dados.
- Os dados que nos interessam não são determinísticos.
- Assim esse modelo matemático geralmente é um modelo probabilístico ou estocástico.

Modelos conceituais

- Precisamos de um *modelo estatístico conceitual*.
- **Modelo Estatístico Conceitual:** Uma distribuição de probabilidade **hipotética** descrevendo como os dados observados **poderiam** ter sido gerados.
- A modelagem é a concepção de um arcabouço matemático capaz de gerar os dados.
- Os dados que nos interessam não são determinísticos.
- Assim esse modelo matemático geralmente é um modelo probabilístico ou estocástico.
- Vamos listar algumas das propriedades desejadas de um bom modelo estatístico.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.
- O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.
- O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade.
- Um mecanismo plausível pode sugerir intervenções ou ações que alterem a realidade de alguma maneira desejada (prevenindo doenças e fraudes, por exemplo).

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.
- O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade.
- Um mecanismo plausível pode sugerir intervenções ou ações que alterem a realidade de alguma maneira desejada (prevenindo doenças e fraudes, por exemplo).
- Finalmente, o modelo deve ser facilmente manipulável matematicamente e conceitualmente.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.
- O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade.
- Um mecanismo plausível pode sugerir intervenções ou ações que alterem a realidade de alguma maneira desejada (prevenindo doenças e fraudes, por exemplo).
- Finalmente, o modelo deve ser facilmente manipulável matematicamente e conceitualmente.
- Precisamos fazer cálculos de probabilidade com o modelo. Se ele for muito complexo, não seremos capazes disso.

Propriedades desejadas de um modelo probabilístico

- O modelo probabilístico deve ser capaz de simular dados com características estatísticas semelhantes a aquelas observadas na realidade.
- Por exemplo, deve ser capaz de predizer mais ou menos bem eventos que realmente ocorrem na realidade.
- O modelo propõe um mecanismo plausível, que corresponde em algum sentido ao que realmente acontece na realidade.
- Um mecanismo plausível pode sugerir intervenções ou ações que alterem a realidade de alguma maneira desejada (prevenindo doenças e fraudes, por exemplo).
- Finalmente, o modelo deve ser facilmente manipulável matematicamente e conceitualmente.
- Precisamos fazer cálculos de probabilidade com o modelo. Se ele for muito complexo, não seremos capazes disso.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.
- Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.
- Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado.
- Isto significa que ele provavelmente vai ser difícil de analisar matematicamente.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.
- Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado.
- Isto significa que ele provavelmente vai ser difícil de analisar matematicamente.
- Por isto, pode ser razoável considerar modelos que reproduzem apenas algumas das características dos dados subjacentes.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.
- Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado.
- Isto significa que ele provavelmente vai ser difícil de analisar matematicamente.
- Por isto, pode ser razoável considerar modelos que reproduzem apenas algumas das características dos dados subjacentes.
- Queremos reproduzir no modelo as principais características em que estamos mais interessados no momento.

As propriedades costumam ser conflitantes

- Muitas vezes, não é possível ter todas as três propriedades simultaneamente.
- Por exemplo, um modelo para gerar dados que sejam bem realistas talvez tenha que se tornar muito complicado.
- Isto significa que ele provavelmente vai ser difícil de analisar matematicamente.
- Por isto, pode ser razoável considerar modelos que reproduzem apenas algumas das características dos dados subjacentes.
- Queremos reproduzir no modelo as principais características em que estamos mais interessados no momento.
- O processo de modelagem é geralmente difícil, exige experiência, e muitas vezes é uma ciência E uma arte.

Modelos para quê?

- Por que estamos interessados em elaborar modelos matemáticos para os nossos dados observados?

Modelos para quê?

- Por que estamos interessados em elaborar modelos matemáticos para os nossos dados observados?
- Um bom modelo dá certo significado aos nossos dados e ajuda a entender de forma aproximada o mecanismo por meio do qual os dados são criados.

Modelos para quê?

- Por que estamos interessados em elaborar modelos matemáticos para os nossos dados observados?
- Um bom modelo dá certo significado aos nossos dados e ajuda a entender de forma aproximada o mecanismo por meio do qual os dados são criados.
- Muitas vezes, o modelo é apenas uma CARICATURA da situação real.

Modelos para quê?

- Por que estamos interessados em elaborar modelos matemáticos para os nossos dados observados?
- Um bom modelo dá certo significado aos nossos dados e ajuda a entender de forma aproximada o mecanismo por meio do qual os dados são criados.
- Muitas vezes, o modelo é apenas uma CARICATURA da situação real.
- Caricatura é um desenho de um personagem da vida real que enfatiza e exagera algumas das características físicas ou comportamentais da pessoa de uma forma humorística.

Modelos para quê?

- Por que estamos interessados em elaborar modelos matemáticos para os nossos dados observados?
- Um bom modelo dá certo significado aos nossos dados e ajuda a entender de forma aproximada o mecanismo por meio do qual os dados são criados.
- Muitas vezes, o modelo é apenas uma CARICATURA da situação real.
- Caricatura é um desenho de um personagem da vida real que enfatiza e exagera algumas das características físicas ou comportamentais da pessoa de uma forma humorística.

Modelo para rede complexa

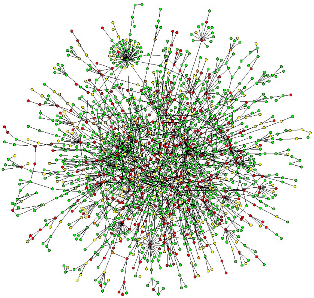
- Redes complexas possuem a maioria dos seus vértices com poucas arestas.

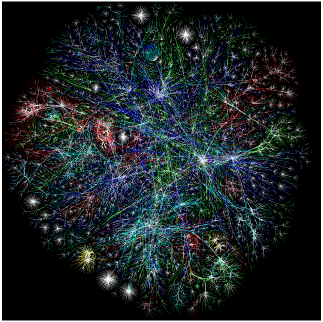
Modelo para rede complexa

- Redes complexas possuem a maioria dos seus vértices com poucas arestas.
- Entretanto, alguns poucos vértices possuem muitas arestas (são os hubs da rede).

Modelo para rede complexa

- Redes complexas possuem a maioria dos seus vértices com poucas arestas.
- Entretanto, alguns poucos vértices possuem muitas arestas (são os hubs da rede).





Modelo para rede complexa

- Seja $\mathbb{P}(k)$ a probabilidade de um vértice escolhido ao acaso possuir k arestas.

Modelo para rede complexa

- Seja $\mathbb{P}(k)$ a probabilidade de um vértice escolhido ao acaso possuir k arestas.
- Quase sempre, encontramos em redes complexas que $\mathbb{P}(k) \approx c/k^\gamma$ onde c e γ são constantes.

Modelo para rede complexa

- Seja $\mathbb{P}(k)$ a probabilidade de um vértice escolhido ao acaso possuir k arestas.
- Quase sempre, encontramos em redes complexas que $\mathbb{P}(k) \approx c/k^\gamma$ onde c e γ são constantes.
- Isto é chamado uma distribuição de probabilidade na forma power-law (potência inversa de k).

Modelo para rede complexa

- Seja $\mathbb{P}(k)$ a probabilidade de um vértice escolhido ao acaso possuir k arestas.
- Quase sempre, encontramos em redes complexas que $\mathbb{P}(k) \approx c/k^\gamma$ onde c e γ são constantes.
- Isto é chamado uma distribuição de probabilidade na forma power-law (potência inversa de k).
- Como isto pode acontecer na prática?

Modelo para rede complexa

- Seja $\mathbb{P}(k)$ a probabilidade de um vértice escolhido ao acaso possuir k arestas.
- Quase sempre, encontramos em redes complexas que $\mathbb{P}(k) \approx c/k^\gamma$ onde c e γ são constantes.
- Isto é chamado uma distribuição de probabilidade na forma power-law (potência inversa de k).
- Como isto pode acontecer na prática?

Modelo de Polya-Erdös

- Suponha que cada par de vértices joga uma moeda para o alto com probabilidade θ de sair cara.

Modelo de Polya-Erdös

- Suponha que cada par de vértices joga uma moeda para o alto com probabilidade θ de sair cara.
- Se der cara, um link é estabelecido entre eles.

Modelo de Polya-Erdös

- Suponha que cada par de vértices joga uma moeda para o alto com probabilidade θ de sair cara.
- Se der cara, um link é estabelecido entre eles.
- Se der coroa, eles não se ligam.

Modelo de Polya-Erdös

- Suponha que cada par de vértices joga uma moeda para o alto com probabilidade θ de sair cara.
- Se der cara, um link é estabelecido entre eles.
- Se der coroa, eles não se ligam.
- Como veremos mais tarde, o número de links de um nó num grafo com n vértices segue uma distribuição Binomial $\text{Bin}(n-1, \theta)$
- Por mero acaso, alguns vértices terão um número de links maior que outros.

Modelo de Polya-Erdös

- Suponha que cada par de vértices joga uma moeda para o alto com probabilidade θ de sair cara.
- Se der cara, um link é estabelecido entre eles.
- Se der coroa, eles não se ligam.
- Como veremos mais tarde, o número de links de um nó num grafo com n vértices segue uma distribuição Binomial $\text{Bin}(n-1, \theta)$
- Por mero acaso, alguns vértices terão um número de links maior que outros.

Modelo de Polya-Erdös

- Este modelo de Polya-Erdös não é capaz de gerar a característica power-law vista em grafos reais de redes complexas.

Modelo de Polya-Erdös

- Este modelo de Polya-Erdös não é capaz de gerar a característica power-law vista em grafos reais de redes complexas.
- O número de links de um vértice tem pouca variação em torno da média.
- Nunca aparecem os hubs dominantes que vemos nos casos reais.

Modelo de Polya-Erdös

- Este modelo de Polya-Erdös não é capaz de gerar a característica power-law vista em grafos reais de redes complexas.
- O número de links de um vértice tem pouca variação em torno da média.
- Nunca aparecem os hubs dominantes que vemos nos casos reais.
- Este não é um bom modelo para as redes complexas da realidade.

Modelo de preferential attachment

- O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa.

Modelo de preferential attachment

- O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa.
- Comece com poucos vértices ligados ao acaso entre si pelo modelo anterior.

Modelo de preferential attachment

- O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa.
- Comece com poucos vértices ligados ao acaso entre si pelo modelo anterior.
- Introduza novos vértices sequencialmente.

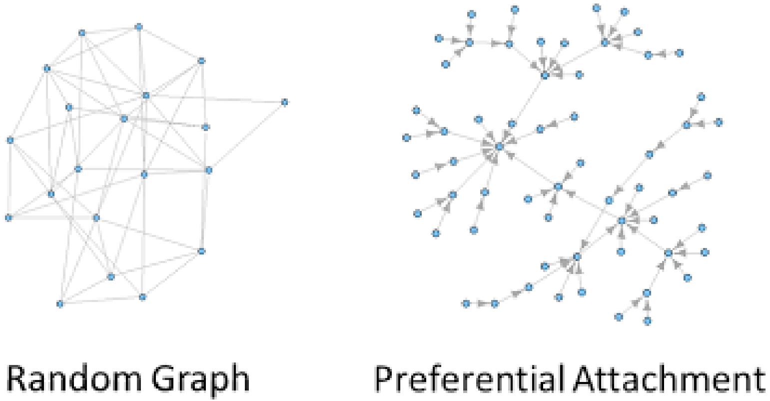
Modelo de preferential attachment

- O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa.
- Comece com poucos vértices ligados ao acaso entre si pelo modelo anterior.
- Introduza novos vértices sequencialmente.
- Um novo vértice conecta-se a um nó já existente com uma probabilidade proporcional ao número de arestas que o nó antigo já possui.

Modelo de preferential attachment

- O modelo de rede social *preferential-attachment* de Barabási-Albert é uma alternativa.
- Comece com poucos vértices ligados ao acaso entre si pelo modelo anterior.
- Introduza novos vértices sequencialmente.
- Um novo vértice conecta-se a um nó já existente com uma probabilidade proporcional ao número de arestas que o nó antigo já possui.

Os dois modelos: exemplo de realizações



Modelo de preferential attachment

- Este não é um modelo perfeito para as redes complexas reais.

Modelo de preferential attachment

- Este não é um modelo perfeito para as redes complexas reais.
- Mas ele induz uma distribuição nos graus dos vértices de redes complexas que possui uma forma de power-law, com cauda pesada.

Modelo de preferential attachment

- Este não é um modelo perfeito para as redes complexas reais.
- Mas ele induz uma distribuição nos graus dos vértices de redes complexas que possui uma forma de power-law, com cauda pesada.
- Temos em mãos então um mecanismo hipótetico que produz um aspecto muito visível e característico das redes complexas.

Modelo de preferential attachment

- Este não é um modelo perfeito para as redes complexas reais.
- Mas ele induz uma distribuição nos graus dos vértices de redes complexas que possui uma forma de power-law, com cauda pesada.
- Temos em mãos então um mecanismo hipótetico que produz um aspecto muito visível e característico das redes complexas.
- Temos uma caricatura do processo gerador REAL das redes complexas.

Modelo de preferential attachment

- Este não é um modelo perfeito para as redes complexas reais.
- Mas ele induz uma distribuição nos graus dos vértices de redes complexas que possui uma forma de power-law, com cauda pesada.
- Temos em mãos então um mecanismo hipótetico que produz um aspecto muito visível e característico das redes complexas.
- Temos uma caricatura do processo gerador REAL das redes complexas.

Modelos para quê?

- Outro uso de um bom modelo é fazer predições.

Modelos para quê?

- Outro uso de um bom modelo é fazer predições.
- Um modelo de classificação de risco de crédito serve para isto.

Modelos para quê?

- Outro uso de um bom modelo é fazer predições.
- Um modelo de classificação de risco de crédito serve para isto.
- Com base em várias features (características) de um usuário, conseguimos prever se ele vai pagar ou não na data combinada um eventual empréstimo.

Modelos para quê?

- Outro uso de um bom modelo é fazer predições.
- Um modelo de classificação de risco de crédito serve para isto.
- Com base em várias features (características) de um usuário, conseguimos prever se ele vai pagar ou não na data combinada um eventual empréstimo.
- Isto é feito com dados históricos: temos uma enorme coleção de indivíduos que tomaram empréstimo e qual foi o resultado ($Y = 1$, pagou; $Y = 0$, não pagou).

Modelos para quê?

- Outro uso de um bom modelo é fazer predições.
- Um modelo de classificação de risco de crédito serve para isto.
- Com base em várias features (características) de um usuário, conseguimos prever se ele vai pagar ou não na data combinada um eventual empréstimo.
- Isto é feito com dados históricos: temos uma enorme coleção de indivíduos que tomaram empréstimo e qual foi o resultado ($Y = 1$, pagou; $Y = 0$, não pagou).
- Para cada indivíduo, temos também suas características coletadas como um vetor x.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.
- Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.
- Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo.
- Um cliente com as características x chega e pede um empréstimo.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.
- Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo.
- Um cliente com as características x chega e pede um empréstimo.
- Calcule $\mathbb{P}(Y = 1|\mathrm{x})$ usando o modelo.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.
- Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo.
- Um cliente com as características x chega e pede um empréstimo.
- Calcule $\mathbb{P}(Y = 1|\mathrm{x})$ usando o modelo.
- Se a probabilidade é baixa, não conceda o empréstimo.

Modelos para quê?

- Algumas das características: sexo, idade, tempo como correntista, saldo médio, etc.
- Com estes dados estatísticos, encontramos um modelo para $\mathbb{P}(Y = 1|\mathrm{x})$.
- Isto é, um modelo para a probab de pagar dado que possui as características x.
- Este modelo é usado para predizer o comportamento de futuros tomadores de empréstimo.
- Um cliente com as características x chega e pede um empréstimo.
- Calcule $\mathbb{P}(Y = 1|\mathrm{x})$ usando o modelo.
- Se a probabilidade é baixa, não conceda o empréstimo.

Modelos para quê?

- Tomar decisões...

Modelos para quê?

- Tomar decisões...
- Conceder o empréstimo?

Modelos para quê?

- Tomar decisões...
- Conceder o empréstimo?
- Oferecer desconto a cliente se é grande a chance dele comprar um item muito caro.

Modelos para quê?

- Tomar decisões...
- Conceder o empréstimo?
- Oferecer desconto a cliente se é grande a chance dele comprar um item muito caro.
- Cortar a conexão a uma rede se a chance de que certas atividades na rede sejam ação de hackers.

Modelos para quê?

- Tomar decisões...
- Conceder o empréstimo?
- Oferecer desconto a cliente se é grande a chance dele comprar um item muito caro.
- Cortar a conexão a uma rede se a chance de que certas atividades na rede sejam ação de hackers.
- Construir uma nova estação metereológica numa localização (x, y) se esta posição minimiza a incerteza de previsões para a região como um todo a partir da rede existente mais a nova estação.

Modelos para quê?

- Tomar decisões...
- Conceder o empréstimo?
- Oferecer desconto a cliente se é grande a chance dele comprar um item muito caro.
- Cortar a conexão a uma rede se a chance de que certas atividades na rede sejam ação de hackers.
- Construir uma nova estação metereológica numa localização (x, y) se esta posição minimiza a incerteza de previsões para a região como um todo a partir da rede existente mais a nova estação.

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.
- Veremos muitas aplicações e exemplos reais mas a ênfase está nos fundamentos.

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.
- Veremos muitas aplicações e exemplos reais mas a ênfase está nos fundamentos.
- Nível de matemática requerido: básico (a esta altura, você já viu a lista 01).

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.
- Veremos muitas aplicações e exemplos reais mas a ênfase está nos fundamentos.
- Nível de matemática requerido: básico (a esta altura, você já viu a lista 01).
- Cálculo de várias variáveis: derivada parcial, gradiente, integral múltipla, máximização de $f(x, y)$.

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.
- Veremos muitas aplicações e exemplos reais mas a ênfase está nos fundamentos.
- Nível de matemática requerido: básico (a esta altura, você já viu a lista 01).
- Cálculo de várias variáveis: derivada parcial, gradiente, integral múltipla, máximização de $f(x, y)$.
- Precisamos mais dos conceitos do que da manipulação algébrica exaustiva.

Objetivos da disciplina

- Estudar os fundamentos dos modelos estatísticos úteis para análise de dados.
- Veremos muitas aplicações e exemplos reais mas a ênfase está nos fundamentos.
- Nível de matemática requerido: básico (a esta altura, você já viu a lista 01).
- Cálculo de várias variáveis: derivada parcial, gradiente, integral múltipla, máximização de $f(x, y)$.
- Precisamos mais dos conceitos do que da manipulação algébrica exaustiva.
- Algebra de matrizes: importante, quanto mais você souber, melhor para você, inclusive a manipulação.
- Espero que você já tenha sido exposto a um curso de probabilidade anteriormente: vamos revisar muito rapidamente.

Livro Texto

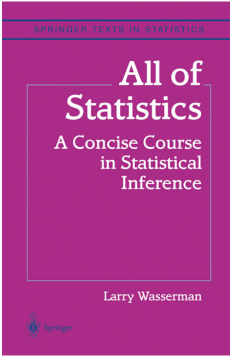
- Vamos seguir dois livros.

Livro Texto

- Vamos seguir dois livros.
- A primeira parte da disciplina (3 semanas) cobre probabilidade.

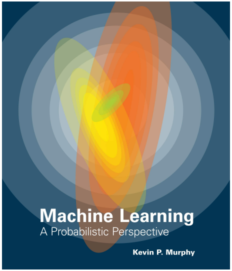
Livro Texto

- Vamos seguir dois livros.
- A primeira parte da disciplina (3 semanas) cobre probabilidade.
- Vou usar os 5 primeiros capítulos de *All of Statistics*, de Larry Wasserman, do Depto de Machine Learning de Carnegie Mellon.



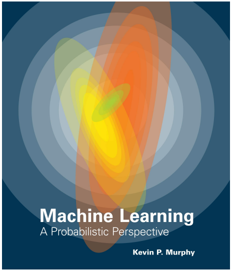
...

- A segunda parte da disciplina vai se basear no livro *Machine Learning, a Probabilistic Approach*, de Kevin Murphy.
- Atualmente, no Google `http://research.google.com/pubs/KevinMurphy.html`



...

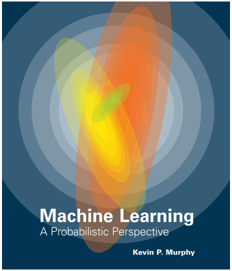
- A segunda parte da disciplina vai se basear no livro *Machine Learning, a Probabilistic Approach*, de Kevin Murphy.
- Atualmente, no Google http://research.google.com/pubs/KevinMurphy.html



- Vamos cobrir os capítulos integralmente os capítulos: 2, 4, 6, 7, 8, 9.

...

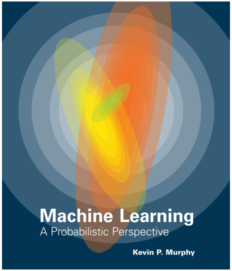
- A segunda parte da disciplina vai se basear no livro *Machine Learning, a Probabilistic Approach*, de Kevin Murphy.
- Atualmente, no Google `http://research.google.com/pubs/KevinMurphy.html`



- Vamos cobrir os capítulos integralmente os capítulos: 2, 4, 6, 7, 8, 9.
- Cobriremos parcialmente os capítulos 3 (3.2 e 3.5) , 12 (12.1, 12.2, 12.3), 13 (13.3 e 13.4), 14 (14.2 e 14.3)

...

- A segunda parte da disciplina vai se basear no livro *Machine Learning, a Probabilistic Approach*, de Kevin Murphy.
- Atualmente, no Google
  `http://research.google.com/pubs/KevinMurphy.html`



- Vamos cobrir os capítulos integralmente os capítulos: 2, 4, 6, 7, 8, 9.
- Cobriremos parcialmente os capítulos 3 (3.2 e 3.5) , 12 (12.1, 12.2, 12.3), 13 (13.3 e 13.4), 14 (14.2 e 14.3)

Avaliação

- Listas de exercícios semanais: 40 pontos ao todo.

Avaliação

- Listas de exercícios semanais: 40 pontos ao todo.
- Teremos 3 provas de 20 pontos cada.

Avaliação

- Listas de exercícios semanais: 40 pontos ao todo.
- Teremos 3 provas de 20 pontos cada.
- Prova serÃ¡ SEM consulta.
- Site inicial da disciplina (preciso remontar): `http://homepages.dcc.ufmg.br/~assuncao/EstatCC/`
- Site oficial da disciplina: moodle.

Fundamentos Estatísticos para Ciência dos dados