Fundamentos Estatísticos de Ciência dos Dados A

FECD A

Renato Assunção - DCC/UFMG e ESRI Inc.

Logística

- Nome da disciplina:
  a. Fundamentos Estatísticos para Ciência dos Dados A
- Horário:
  a. Terças e Quintas, de 9:25 às 11:05
- Local:
  a. Pela plataforma TEAMS
- Link para a equipe na plataforma TEAMS

- Monitor: não há monitor este semestre.
- Horário da monitoria: não há monitor

Objetivos

Permitir ao aluno:

dominar princípios, técnicas e metodologias

associadas ao raciocínio probabilístico

e à análise de dados estatísticos.

Explosão no uso e desenvolvimento de modelos probabilísticos e estatísticos bastante sofisticados em computação.

Uso em:

robótica probabilística, processamento de linguagem natural, recuperação de informação, redes de sensores e análise de redes sociais.

Disciplinas de:

aprendizado de máquina, mineração de dados e modelos gráficos probabilísticos

É impossível desenvolver pesquisa de qualidade nestas últimas três áreas sem um sólido conhecimento de probabilidade e estatística.

Objetivos

Disciplina dirigida aos alunos de ciência da computação e similares.

Visa a oferecer os fundamentos de probabilidade, processos estocásticos e inferência para análise de dados.

O objetivo principal é que os alunos possam acompanhar e desenvolver novas idéias nos campos de aplicação de computação que têm usado métodos estatísticos com intensidade.

Ao final do curso, os alunos devem ser capazes de ler os capítulos mais avançados dos livros de aprendizado de máquina,

Novidade a partir de 2021

Eu nunca consigo cobrir o programa de FECD → FECD A e FECD B

FECD A

1. Revisão de Probabilidade: Regra de Bayes.
2. Principais distribuições de probabilidade;
3. Distribuições conjuntas e condicionais.
4. Momentos e desigualdades.
5. Convergência de variáveis aleatórias;
6. Boostrap, Jackknife e teste de Kolmogorov.
7. Revisão de álgebra de matrizes
8. Normal multivariada
9. Métrica para medir distância entre distribuições de probabilidade: entropia, Kullback-Leibler e Kolmogorov
10. Modelos probabilísticos com a normal multivariada:
    a. Classificação e regra ótima de Bayes
    b. PCA
    c. Análise fatorial (fatores latentes)
11. Seleção de modelos:
    a. entropia, critério de informação de Akaike (AIC) e minimum description length (MDL)
12. Modelos de mistura: análise de clusters

FECD B

1. Regressão linear múltipla
2. Regressão com dados não-gaussianos: logística e Poisson
3. Modelos com dados dependentes: espaciais e temporais.
4. Princípios de inferência estatística: vício, variância, consistência, eficiência.
5. Métodos de estimação: distância mínima, mínimos quadrados, mínimo qui-quadrado, máxima verossimilhança.
6. Exemplos com máxima verossimilhança.
7. Propriedades de otimalidade do estimador de máxima verossimilhança.
8. Algoritmo EM.
9. Intervalos de Confiança
10. Testes de Hipótese e p-valores.
11. Famílias exponenciais de distribuições e GLM
12. Modelos de fatores latentes: filtragem colaborativa.
13. Modelos de mistura: análise de clusters com modelos probabilísticos
14. Seleção de modelos de novo.

Fundamentos x prática de análise de dados

fom Lynne Schneider no Twitter: the data doesn't have the answer if you don't know the question.

Vamos focar nos *fundamentos* da análise de dados. Uma analogia seria: este curso está mais para AEDS que para Programação.

A prática da análise de dados é diferente da teoria (os fundamentos).

> Ao analisar dados reais é preciso compreender o problema aplicado para decidir quais são os aspectos relevantes e que devem ser modelados probabilisticamente.
>
> Outros aspectos, considerados irrelevantes, são todos despejados numa cesta genérica do modelo (o termo de "erro").
>
> Esta escolha de quais fatores são os relevantes é muitas vezes subjetiva e vai sendo aprimorada na prática da análise de dados.
>
> É preciso também decidir quais dados devem ser coletados e como eles devem ser processados.
>
> É preciso explorar e visualizar os dados, transformar os dados em medidas ou escalas mais convenientes para o problema.
>
> Selecionar os modelos e métodos computacionais a serem aplicados
>
> Comunicar os resultados das análises para gerentes ou o público em geral.

Novidade em 2021

- Flipped class: agora vai?

- A partir da 2a semana, **pelo menos** uma das aulas semanais será de discussão "prática"

- Os alunos devem assistir aos vídeos com o material expositivo

- No horário de aula, vamos discutir dúvidas, perguntas mais gerais, conexões com a prática e discussão de projetos individuais

- Haverá TPs? ** Não sei ainda ** Minha tendência é **não** dar TPs

- Talvez discutir trabalhos individuais voluntários…

Avaliação

40% - Listas de exercícios semanais (aprox 15 listas)

Entrega via moodle, às terças-feiras até 23h59.

Todos as listas contam igualmente e totalizam 40% da nota final.

As três notas mais baixas das listas serão descartadas.

Conseqüentemente, não será aceita nenhuma lista atrasada por qualquer motivo.

60% da nota: provas (duas ou três)

Uma aula extra no meio do semestre

Como ir fazer pós-graduação nos EUA ou Canadá:

sem sair de BH

sem pagar a universidade

ganhando uma bolsa para cobrir as despesas

É mais fácil que entrar na UFMG…

Uma palestra de duas horas sobre isto no meio do semestre

Se não puder esperar, clique aqui...

O professor

- Bacharel (UFMG) e Mestre (IMPA) em matemática
- PhD (Univ of Washington, Seattle) em Estatística
- Pesquisador do CNPq no comitê de matemática
- Ex- Professor Titular do DCC - UFMG

- Aposentado mas professor voluntário ….
- Atualmente, trabalho para a ESRI Inc.