Gabarito da Terceira Prova
Fundamentos Estatísticos para Ciência dos Dados

08/06/2016

1. *UC Berkeley, CS 194-10, Introduction to Machine Learning Fall 2011, Prof Stuart Russell* Um modelo de regressão linear da resposta y_i tem um único atributo x_i que tem sempre valores positivos (isto é, $x_i > 0$ em qualquer observação i). Todas as hipóteses usuais aplicam-se (dados gaussianos, independência, etc) EXCETO que a variância seja constante para todas as n observações. Verifica-se que a variância do ruído ou erro ϵ_i é proporcional a x_i.

   - Qual das seguintes densidades para o vetor $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ descreve corretamente a verossimilhança dos parâmetros $\beta_0, \beta_1, \sigma^2$:

     –
     $$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \prod_i x_i} \exp\left(- \sum_i \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2 x_i^2 \sigma^2} \right)$$

     –
     $$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left(- \sum_i \frac{x_i \, (y_i - (\beta_0 + \beta_1 x_i))^2}{2 \sigma^2} \right)$$

     –
     $$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \prod_i \sqrt{x_i}} \exp\left(- \sum_i \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2 x_i \sigma^2} \right)$$

   - Identifique na Figura 1, com notação de matriz qual (ou quais) dos plots poderia ser gerado por uma instância do modelo identificado no item anterior.

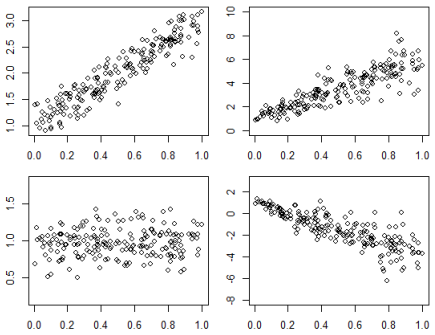

Figura 1: Quatro conjuntos de dados para regressão com variância não constante.

   - Obtenha a derivada com respeito a β_1 da **LOG**-verossimilhança escolhida no primeiro item. Obtenha uma expressão para o MLE de β_1 supondo que os demais parâmetros são conhecidos.

Solução: Dado x_i, temos $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ onde $\epsilon_i \sim N(0, \sigma_i^2) = N(0, \sigma^2 x_i)$ pois a variância do ruído ou erro ϵ_i é proporcional a x_i e σ^2 é agora simplesmente a constante de proporcionalidade. Desssa forma, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_i^2)$ com densidade dada por

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2 x_i}} \exp\left(- \frac{1}{\sigma^2 x_i} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \right)$$

Como $Y_1, \ldots, Y_n$ são v.a.'s independentes, a sua densidade conjunta é o produto das densidades. Portanto, a verossmilhança é

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2 x_i}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2 x_i} \right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_i \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2x_i\sigma^2} \right) \end{aligned}$$

Portanto, a terceira opção é a correta.

A resposta são os plots nas posições $(1,2)$ e $(2,2)$. Nos dois, a v.a. $(Y|x)$ tem seu valor esperado $\mathbb{E}(Y|x)$ variando linearmente com x (isto é, $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$). Em $(1,1)$ temos $\beta_1 > 0$ e em $(2,2)$ temos $\beta_1 < 0$. Nestes dois plots, a variabilidade de $(Y|x)$ cresce com x. Vemos que $\mathbb{V}(Y|x)$ aumenta com x nos dois casos. No plot na posição $(1,1)$, $\mathbb{E}(Y|x)$ aumenta com x mas $\mathbb{V}(Y|x)$ é a mesma para todo x. No caso do plot $(2,1)$, $\mathbb{E}(Y|x)$ não varia com x (isto é, $\beta_1 = 0$) e $\mathbb{V}(Y|x)$ também não varia com x.

A log-verossmilhança é dada por

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_i \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2x_i\sigma^2}$$

e portanto

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_1} &= -\sum_i \frac{1}{2x_i\sigma^2} 2\left(y_i - (\beta_0 + \beta_1 x_i)\right)\ (-x_i) \\ &= \frac{1}{\sigma^2}\left(\sum_i y_i - (\beta_0 n + \beta_1 \sum_i x_i) \right) \\ &= \frac{n}{\sigma^2}\left(\bar{y} - (\beta_0 + \beta_1 \bar{x})\right) \end{aligned}$$

Note que para esta derivada ser igual a zero, estamos pedindo que β_0 e β_1 sejam tais que a média $\bar{y}$ iguale o valor da reta quando tomarmos x igual à média $\bar{x}$ (isto é, que $\bar{y} = \beta_0 + \beta_1 \bar{x}$).

---

2. Um modelo de regressão linear da resposta y_i tem um único atributo x_i com valores sempre positivos. Sabe-se que a relação linear entre y e x é tal que o intercepto $\beta_0 = 0$. Assim, o modelo de regressão linear é da forma $y_i = \beta x_i + \epsilon_i$ onde $\epsilon_i \sim N(0, \sigma^2)$. A verossimilhança desse modelo é dada por

$$L(\beta, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_i \frac{(y_i - \beta x_i)^2}{2\sigma^2} \right)$$

   - Obtenha o MLE β_{mle}.
   - Um modelo de regressão regularizado modifica a log-verossmilhança para a função

$$J(\beta, \sigma^2) = \log(L(\beta, \sigma^2)) - \frac{\lambda}{\sigma^2}\beta^2$$

     onde $\lambda > 0$ é uma constante conhecida. Obtenha o valor β_{reg} que maximiza a log-verossimilhança regularizada $J(\beta, \sigma^2)$.
   - É possível adiantar a desigualdade entre os estimadores. Isto é, podemos dizer de teremos sempre $|\beta_{reg}| > |\beta_{mle}|$?
   - A variância do erro de estimação está associada com o valor de $|\partial^2 J/\partial\beta^2|$. Obtenha este valor.

Solução: Derivando diretamente a verossimilhança e igualando a zero, temos:

$$\begin{aligned}\frac{\partial L(\beta,\sigma^2)}{\partial\beta} &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}\right)\frac{\partial}{\partial\beta}\left(-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}\right)\\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}\right)\left(-\frac{1}{2\sigma^2}\sum_i 2(y_i-\beta x_i)(-x_i)\right)=0\end{aligned}$$

Como

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}\right)>0$$

pois $e^x>0$ para todo x, a única forma dessa expressão ser zero é se

$$-\frac{1}{2\sigma^2}\sum_i 2(y_i-\beta x_i)(-x_i)=0$$

o que implica em

$$\sum_i y_i x_i-\beta\sum_i x_i^2=0$$

ou seja,

$$\beta_{mle}=\sum_i y_i x_i/\sum_i x_i^2\,.$$

Observe que nós *quase nunca* derivamos diretamente a verossimilhança como fizemos acima. Em geral, tomamos o log e derivamos. Neste caso,

$$\begin{aligned}\frac{\partial\log(L(\beta,\sigma^2))}{\partial\beta} &= \frac{\partial}{\partial\beta}\left(-\frac{n}{2}\log(2\pi\sigma^2)-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}\right)\\ &= -\frac{1}{2\sigma^2}\sum_i 2(y_i-\beta x_i)(-x_i)\\ &= -\frac{1}{\sigma^2}\left(\sum_i y_i x_i-\beta\sum_i x_i^2\right)\end{aligned}$$

Igualando a zero, obtemos o mesmo MLE de antes.

A log-verossimilhança regularizada é igual a

$$J(\beta,\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)-\sum_i\frac{(y_i-\beta x_i)^2}{2\sigma^2}-\frac{\lambda}{\sigma^2}\beta^2$$

Derivando em relação a β e igualando a zero, obtemos:

$$\frac{\partial J}{\partial\beta}=\frac{1}{\sigma^2}\left(\sum_i x_i y_i-\beta\sum_i x_i^2-2\lambda\beta\right)=0$$

o que implica em

$$\beta_{reg}=\frac{\sum_i x_i y_i}{\sum_i x_i^2+2\lambda}$$

Como $\lambda>0$, temos $|\beta_{reg}|<|\beta_{mle}|$.

Por simples derivação adicional encontramos

$$\frac{\partial^2 J}{\partial\beta^2}=-\frac{\sum_i x_i^2+2\lambda}{\sigma^2}$$

---

3. O peso (em quilos) de um brasileiro do sexo masculino entre 20 e 24 anos escolhido ao acaso segue uma distribuição com valor esperado 75. O desvio-padrão desse peso é denotado por σ.

   - O valor de σ é um dos números na seguinte lista: $\{1, 2, 5, 10, 15, 20, 25\}$. Diga qual deles é o mais razoável na sua opinião (justificando, claro).
   - O elevador do DCC no ICEx diz que o limite máximo de segurança para seu uso é de 975 quilos ou 13 pessoas (note que $75 \times 13 = 975$). Usando o valor de σ escolhido anteriormente, use o TCL para calcular aproximadamente a probabilidade de que 10 pessoas escolhidas ao acaso entrem no elevador e o peso ultrapasse 975 quilos.
   - Se n pessoas escolhidas ao acaso entrarem no elevador ao mesmo tempo, existe certa probabilidade de que a soma de seus pesos ultrapasse 975 quilos. Esta probabilidade aumenta com n. Qual o menor n tal que esta probabilidade seja 0.10? OBS: Use que $\mathbb{P}(N(0,1) > 1.28) = 0.10$.

   Solução: Distribuições de medidas antropométricas como peso e altura seguem aproximadamente uma distribuição gaussiana. Aproximadamente 95% dos indivíduos deverão estar no intervalo $(75 - 2\sigma, 75 + 2\sigma)$. Considerando que o peso de brasileiros adultos deve estar na faixa de 100 a 50 quilos, devemos ter um valor de σ em torno de $(100 - 50)/4 = 12.5$. Assim, a resposta $\sigma = 10$ ou $\sigma = 15$ são as mais razoáveis.

   Sejam $X_1, \ldots, X_{10}$ os pesos aleatórios de 10 adultos do sexo masculino escolhidos ao acaso. Então, com $\overline{X} = (X_1 + \ldots + X_{10})/10$, temos

$$\begin{aligned}
\mathbb{P}(X_1 + \ldots + X_{10} > 975) &= \mathbb{P}\left(\frac{X_1 + \ldots + X_{10}}{10} > \frac{975}{10}\right) \\
&= \mathbb{P}\left(\frac{\overline{X} - 75}{\sigma/\sqrt{10}} > \frac{97.5 - 75}{\sigma/\sqrt{10}}\right) \\
&\approx \mathbb{P}(N(0,1) > \frac{71.15}{\sigma}) \\
&= \mathbb{P}(N(0,1) > 4.74) \quad \text{usando } \sigma = 15 \\
&= 1.07 \times 10^{-6}\,,
\end{aligned}$$

   uma chance muito pequena.

   Com um n genérico, fazendo $\overline{X}_n = (X_1 + \ldots + X_n)/n$, queremos calcular

$$\begin{aligned}
\mathbb{P}(X_1 + \ldots + X_n > 975) &= \mathbb{P}\left(\frac{X_1 + \ldots + X_n}{n} > \frac{975}{n}\right) \\
&= \mathbb{P}\left(\frac{\overline{X}_n - 75}{\sigma/\sqrt{n}} > \frac{975/n - 75}{\sigma/\sqrt{n}}\right) \\
&\approx \mathbb{P}\left(N(0,1) > \frac{975/n - 75}{15/\sqrt{n}}\right) \quad \text{usando } \sigma = 15\,.
\end{aligned}$$

   No caso de uma $N(0,1)$, temos $\mathbb{P}(N(0,1) > -1.28) = 0.10$. Portanto, queremos um n tal que $(975/n - 75)/(15/\sqrt{n}) \approx -1.28$. Isto é, n tal que

$$\frac{975}{n} - 75 \approx -\frac{19.2}{\sqrt{n}}$$

   ou $975 - 75n + 19.2\sqrt{n} \approx 0$. Substituindo $x = \sqrt{n}$ e a proximação por uma igualdade, devemos resolver uma equação do segundo grau, $975 - 75x^2 + 19.2x = 0$, cujas soluções são $x = 3.479673$ e $x = -3.735983$. Como $x = \sqrt{n} > 0$, a segunda solução pode ser descartada. Usando a primeira solução, temos $n \approx 12.10$, ou $n = 12$.

---

4. Sejam $Y_1, Y_2, \ldots, Y_n$ v.a's contínuas i.i.d. com a seguinte densidade de probabilidade:

$$f(x|\theta) = \theta x^{-(\theta+1)}$$

onde $\theta > 1$ e $x \geq 1$. Ache o MLE de θ.

Solução: A log-verossimilhança é

$$\ell(\theta) = \log\left(\theta^n \left(\prod_i x_i\right)^{-(\theta+1)}\right) = n\log(\theta) - (\theta+1)\sum_i \log(x_i)$$

Derivando e igualando a zero:

$$\frac{\partial \ell(\theta)}{\partial \beta} = \frac{n}{\theta} - \sum_i \log(x_i) = 0$$

o que implica em $\theta_{mle} = n/\sum_i \log(x_i)$, o inverso da a média dos logs dos valores.

---

5. *O efeito de centrar os atributos.* O objetivo deste exercício é mostrar que, ao centrar os atributos, temos coeficientes relacionados de forma simples aos coeficientes obtidos com coeficientes não-centrados. Seja $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$ o vetor que minimiza

$$\sum_i \left(y_i - (\beta_0^* + \beta_1^*(x_{i1} - \bar{x}_1) + \beta_2^*(x_{i2} - \bar{x}_2))\right)^2$$

Isto é, a matriz de desenho tem suas colunas com média zero (exceto a primeira coluna). Seja $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ o coeficiente que minimiza a distância entre $\mathbf{Y}$ e as combinações lineares das colunas *não-centradas*:

$$\sum_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\right)^2$$

Mostre que as soluções dos dois problemas estão relacionadas da seguinte forma:

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\beta}_0^* - \hat{\beta}_1^*\bar{x}_1 - \hat{\beta}_2^*\bar{x}_2 \\
\hat{\beta}_1 &= \hat{\beta}_1^* \\
\hat{\beta}_2 &= \hat{\beta}_2^*
\end{aligned}$$

Solução: Seja

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^* &= \min_{\boldsymbol{\beta}^*}\left\{\sum_i \left(y_i - (\beta_0^* + \beta_1^*(x_{i1} - \bar{x}_1) + \beta_2^*(x_{i2} - \bar{x}_2))\right)^2\right\} \\
&= \min_{\boldsymbol{\beta}^*}\left\{\sum_i \left(y_i - ((\beta_0^* - \beta_1^*\bar{x}_1 - \beta_2^*\bar{x}_2) + +\beta_1^* x_{i1} + \beta_2^* x_{i2})\right)^2\right\} \\
&= \min_{\boldsymbol{\beta}^*}\left\{\sum_i \left(y_i - (\beta_0 + \beta_1^* x_{i1} + \beta_2^* x_{i2})\right)^2\right\}
\end{aligned}$$

onde fizemos $\beta_0 = \beta_0^* - \beta_1^*\bar{x}_1 - \beta_2^*\bar{x}_2$. Como todo vetor da forma $(\beta_0, \beta_1^*, \beta_2^*)$ tem um único correspondente da forma $(\beta_0^*, \beta_1^*, \beta_2^*)$ com $\beta_0 = \beta_0^* - \beta_1^*\bar{x}_1 - \beta_2^*\bar{x}_2$. Assim, a solução do problema centrado obtém um vetor ótimo $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$ que minimiza a última expressão e portanto resolve também o problema ótimo para as variáveis não-centradas bastando tomar

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\beta}_0^* - \hat{\beta}_1^*\bar{x}_1 - \hat{\beta}_2^*\bar{x}_2 \\
\hat{\beta}_1 &= \hat{\beta}_1^* \\
\hat{\beta}_2 &= \hat{\beta}_2^*
\end{aligned}$$