Terceira Prova - FECD - Gabarito

Ano da pandemia 2020

1. **4 PONTOS** As v.a.'s binárias $y_1, y_2, \ldots, y_n$ são independentes com $y_i \sim$ Bernoulli(p_i) onde

$$p_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip})}} = \frac{1}{1+e^{-\mathbf{x}_i^t \boldsymbol{\theta}}}$$

onde o vetor-COLUNA $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^t$ é a i-ésima LINHA da matriz $\mathbf{X}$ de dimensão $n \times (p+1)$ e $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor-coluna de parâmetros desconhecidos. Mostre EM DETALHES que a equação de verossimilhança é dada por

$$\mathbf{0} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{X}^t(\mathbf{y}-\mathbf{p})$$

OBS: Eu saltei alguns detalhes nos slides. Por exemplo, não mostrei como os p_i's aparecem na equação acima. Você deve mostrar estes detalhes também, não basta copiar os slides.

Solução: Os slides possuem toda a derivação necessária faltando poucos detalhes. Seja

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}.$$

a matriz $n \times (p+1)$ com os preditores (ou features) como colunas (lembrando que a primeira coluna é formada por 1's). A log-verossimilhança é

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{n} P(Y_i=1)^{y_i}\ P(Y_i=0)^{1-y_i}\right) \\ &= \sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] = \sum_{i=1}^{n}\left[y_i\ \mathbf{x}_i^t\boldsymbol{\theta} - \log(1+e^{\mathbf{x}_i^t\boldsymbol{\theta}})\right] \end{aligned}$$

Para maximizar $\ell(\boldsymbol{\theta})$, tomamos derivadas em relação a cada coordenada de $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p)$ e igualamos a zero. O único detalhe relevante que deve ser acrescentado é que

$$p_i = p_i(\boldsymbol{\theta})) = \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\theta}}} = \frac{e^{\mathbf{x}_i^t\boldsymbol{\theta}}}{1+e^{\mathbf{x}_i^t\boldsymbol{\theta}}}$$

Assim,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j} &= \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \frac{e^{\mathbf{x}_i^t\boldsymbol{\theta}}}{1+e^{\mathbf{x}_i^t\boldsymbol{\theta}}} \\ &= \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} p_i(\boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} x_{ij}(y_i - p_i(\boldsymbol{\theta})) \end{aligned}$$

para todo $j = 0, 1, \ldots, p$. Devemos ter esta derivada igual a zero, o que implica em ter $\sum_{i=1}^{n} x_{ij}(y_i - p_i(\boldsymbol{\theta})) = 0$ ou $\sum_i x_{ij} y_i = \sum_i x_{ij} p_i(\boldsymbol{\theta})$. Os dois lados desta igualdade representam uma soma ponderada dos n valores do j-ésimo preditor ou feature (valores da coluna j da matriz $\mathbf{X}$). O lado esquerdo dá um peso binário, 0 ou 1, a cada x_{ij}. Assim, ele é a soma dos valores do predito j para aqueles itens que tiveram $y_i = 1$. O lado direito usa a probabilidade $p_i(\boldsymbol{\theta})$ que depende da escolha dos coeficientes no vetor $\boldsymbol{\theta}$. Assim, encontrar o MLE significa escolher $\widehat{\boldsymbol{\theta}}$ de forma a equilibrar (ou igualar) estas duas somas ponderadas. Na verdade, devemos ter este equilíbrio para todos as colunas de $\mathbf{X}$ usando a mesma escolha para $\widehat{\boldsymbol{\theta}}$. São $p+1$ equações formando um sistema de equações não-lineares em $\boldsymbol{\theta}$. Então

$$\begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_0} \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n 1 \cdot (y_i - p_i(\boldsymbol{\theta})) \\ \sum_{i=1}^n x_{i1} \cdot (y_i - p_i(\boldsymbol{\theta})) \\ \vdots \\ \sum_{i=1}^n x_{ip} \cdot (y_i - p_i(\boldsymbol{\theta})) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & & & \vdots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} y_1 - p_1(\boldsymbol{\theta}) \\ y_2 - p_2(\boldsymbol{\theta}) \\ \vdots \\ y_n - p_n(\boldsymbol{\theta}) \end{bmatrix} = \mathbf{X}'(\mathbf{y} - \mathbf{p}(\boldsymbol{\theta})) = \mathbf{0}$$

onde $\mathbf{p}(\boldsymbol{\theta})$ é o vetor coluna $(p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), \dots, p_n(\boldsymbol{\theta}))'$. O MLE é a solução $\hat{\boldsymbol{\theta}}$ do sistema de equações não-lineares acima, que pode ser escrito em forma matricial como $\mathbf{X}^t(\mathbf{y} - \mathbf{p}(\boldsymbol{\theta})) = \mathbf{0}$ ou, de forma equivalente, $\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\mathbf{p}(\boldsymbol{\theta})$.

---

2. **4 PONTOS** Ainda com relação ao problema acima, faça a derivação detalhada mostrando que

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

e esclarecendo o que é a matriz $\mathbf{W}$.

Solução: Vamos considerar uma das derivadas parciais de segunda ordem:

$$\begin{aligned} \frac{\partial}{\partial \beta_k} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j} &= \frac{\partial}{\partial \beta_k} \sum_{i=1}^n x_{ij}(y_i - p_i(\boldsymbol{\theta})) \\ &= -\sum_{i=1}^n x_{ij} \frac{\partial p_i(\boldsymbol{\theta})}{\partial \beta_k} \end{aligned}$$

Temos

$$\begin{aligned} \frac{\partial p_i(\boldsymbol{\theta})}{\partial \beta_k} &= \frac{-1}{(1+e^{-\mathbf{x}_i'\boldsymbol{\theta}})^2}\, e^{-\mathbf{x}_i'\boldsymbol{\theta}}\, (-x_{ik}) \\ &= x_{ik} \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\theta}}} \frac{e^{-\mathbf{x}_i'\boldsymbol{\theta}}}{1+e^{-\mathbf{x}_i'\boldsymbol{\theta}}} \\ &= x_{ik}\; p_i(\boldsymbol{\theta})\; (1 - p_i(\boldsymbol{\theta})) \end{aligned}$$

Dessa forma,

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_k \partial \beta_j} = -\sum_{i=1}^n x_{ij}\; x_{ik}\; p_i(\boldsymbol{\theta})\; (1 - p_i(\boldsymbol{\theta}))$$

Esta expressão é o elemento (j,k) da matriz $-\mathbf{X}^t\mathbf{W}\mathbf{X}$ onde $\mathbf{W}$ é uma matriz diagonal $n \times n$ com i-ésimo elemento diagonal igual a $p_i(\boldsymbol{\theta})\; (1 - p_i(\boldsymbol{\theta}))$.

---

3. **4 PONTOS** Encontre o MLE do parâmetro σ^2 se $Y_1, \dots, Y_n$ são i.i.d. com distribuição gaussiana $N(\mu_0, \sigma^2)$ onde μ_0 *é um valor conhecido*. Por exemplo, assuma que $\mu_0 = 7$ e forneça o MLE de σ^2.

Solução: A verossimilhança de σ^2 é

$$L(\sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_0)^2}{2\sigma^2}\right)$$

e portanto a log-verossmilhança é

$$\ell(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu_0)^2$$

Precisamos agora derivar com respeito a σ^2. A única possível dificuldade aqui é que alguns alunos costumam derivar com respeito a σ ou ficam confusos com a presença do expoente 2. Uma maneira muito simples de resolver isto é substituir σ^2 na log-verossmilhança por, por exemplo, v e derivar com respeito a v. No final, voltamos a substituir v por σ^2. Assim,

$$\ell(v) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(v) - \frac{1}{2v}\sum_{i=1}^n (y_i - \mu_0)^2$$

e portanto

$$\frac{\partial \ell(v)}{\partial v} = -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(y_i-\mu_0)^2$$

Igulando esta derivada a zero e isolando v encontramos $\hat{v} = \sum_{i=1} n(y_i-\mu_0)^2/n$. Ou seja, quando o valor esperado μ_0 é conhecido, o MLE de σ^2 é igual a

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i-\mu_0)^2}{n}$$

---

4. **4 PONTOS** Seja Y uma v.a. com $\mu = \mathbb{E}(Y)$. Prove que $m = \mu$ é o valor que minimiza $\mathbb{E}(Y-m)^2$. Isto é, que $\mu = \arg\min_m \mathbb{E}(Y-m)^2$.

   Solução: Vamos mostrar duas soluções. Para a primeira, vou assumir inicialmente que a v.a. Y seja contínua com densidade $f(y)$. Comece derivando $g(m) = \mathbb{E}(Y-m)^2$ com respeito a m e igualando a zero para encontrar os seus pontos críticos:

$$\begin{aligned}
\frac{\partial g(m)}{\partial m} &= \frac{\partial}{\partial m}\mathbb{E}(Y-m)^2 = \frac{\partial}{\partial m}\int (y-m)^2 f(y)dy \\
&= \int f(y)\frac{\partial (y-m)^2}{\partial m} f(y)dy \\
&= \int f(y)\frac{\partial (y-m)^2}{\partial m} f(y)dy \\
&= \int -2(y-m)f(y)dy = -2\int yf(y)dy + 2m\int f(y)dy \\
&= -2\mu + 2m \times 1 = 0
\end{aligned}$$

   o que implica em $m = \mu$ é o único ponto crítico da função $g(m)$. Ele é um ponto de mínimo pois

$$\frac{\partial^2 g(m)}{\partial m^2} = \int (-2)(-1)f(y)dy = 2\int f(y)dy = 2 > 0$$

   O caso de v.a.'s discretas é idêntico, substituindo as integrais por somas.

   Matemáticos e probabilistas teriam objeção à solução acima pois existem v.a.'s que são misturas de v.a.'s contínuas e discretas. Neste caso, a próxima solução é melhor pois é geral e serve para qulaquer v.a. Use o velho truque de somar e subtrair μ no parÃªnteses de $g(m)$:

$$\begin{aligned}
g(m) &= \mathbb{E}(Y-m)^2 \\
&= \mathbb{E}(Y-\mu+\mu-m)^2 \\
&= \mathbb{E}\left[(Y-\mu)^2 + 2(Y-\mu)(\mu-m) + (\mu-m)^2\right] \\
&= \mathbb{E}\left[(Y-\mu)^2\right] + 2\mathbb{E}\left[(Y-\mu)(\mu-m)\right] + \mathbb{E}\left[(\mu-m)^2\right] \\
&= \mathbb{V}(Y) + 2(\mu-m)\mathbb{E}\left[(Y-\mu)\right] + (\mu-m)^2 \quad \text{pois } (\mu-m) \text{ é uma constante} \\
&= \mathbb{V}(Y) + 2(\mu-m)\times 0 + (\mu-m)^2 \quad \text{pois } \mu = \mathbb{E}(Y) \\
&= \mathbb{V}(Y) + (\mu-m)^2
\end{aligned}$$

   O primeiro termo é a variância de Y e ela não depende da constante m que você vai escolher. Assim, para minimizar $g(m)$ você deve considerar aqpenas o segundo termo. Ele é não-negativo e será minimizado se for igual a zero. Isto ocorre se tomarmos $m = \mu$.

---

5. **4 PONTOS** Seja

$$\mathbf{H}_1 = \frac{1}{n}\mathbf{1}\mathbf{1}' = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

   Verifique que, qualquer que seja o vetor resposta $\mathbf{Y}$ de dimensão $n \times 1$, ele pode ser decomposto como

$$\mathbf{Y} = \mathbf{H}_1\mathbf{Y} + (\mathbf{I} - \mathbf{H}_1)\,\mathbf{Y} = \bar{y}\mathbf{1} + (\mathbf{Y} - \bar{y}\mathbf{1})$$

e que os dois vetores do lado direito da equação são ortogonais entre si.

Solução: Como

$$\mathbf{H}_1\mathbf{Y} = \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Y} = \mathbf{1}\left(\frac{1}{n}\mathbf{1}'\mathbf{Y}\right) = \mathbf{1}\bar{y}\,,$$

temos então, somando e subtraindo $\mathbf{H}_1\mathbf{Y}$,

$$\begin{aligned} \mathbf{Y} &= \mathbf{Y} + \mathbf{H}_1\mathbf{Y} - \mathbf{H}_1\mathbf{Y} = \mathbf{H}_1\mathbf{Y} + (\mathbf{I} - \mathbf{H}_1)\,\mathbf{Y} \\ &= \bar{y}\mathbf{1} + \mathbf{Y} - \bar{y}\mathbf{1} \end{aligned}$$

A parte mais relevante é mostrar que os componentes desta decomposição são ortogonais entre si. Temos

$$(\bar{y}\mathbf{1})' \cdot (\mathbf{Y} - \bar{y}\mathbf{1}) = \bar{y}\,(n\bar{y}) - (\bar{y})^2\,\mathbf{1}'\mathbf{1} = n\,(\bar{y})^2 - (\bar{y})^2\,n = 0$$

Outra opção trabalhando com as matrizes, é checar que $\mathbf{H}_1\mathbf{H}_1 = \mathbf{H}_1^2 = \mathbf{H}_1$ (isto é, é uma matriz idempotente). Além disso, $\mathbf{H}_1$ é simétrica ($\mathbf{H}_1 = \mathbf{H}_1'$). Em seguida, veja que

$$(\mathbf{H}_1\mathbf{Y})' \cdot (\mathbf{I} - \mathbf{H}_1)\,\mathbf{Y} = \mathbf{Y}'\mathbf{H}_1'\,(\mathbf{I} - \mathbf{H}_1)\,\mathbf{Y} = \mathbf{Y}'\,(\mathbf{H}_1 - \mathbf{H}_1\mathbf{H}_1)\,\mathbf{Y} = \mathbf{Y}'\,(\mathbf{H}_1 - \mathbf{H}_1)\,\mathbf{Y} = \mathbf{Y}'\,(\mathbf{0})\,\mathbf{Y} = 0$$