Seminário 1a: Holmes e Adams (2002)

Paulo C. Marques F.

Seminário ministrado no Insper

5 de Fevereiro de 2016

Solução bayesiana do problema de classificação

Solução bayesiana do problema de classificação

- Como na primeira aula, temos pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ em que as preditoras X_i são vetores aleatórios em $\mathbb{R}^d$ e as respostas Y_i assumem apenas os valores 0 e 1.

Solução bayesiana do problema de classificação

- Como na primeira aula, temos pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ em que as preditoras X_i são vetores aleatórios em $\mathbb{R}^d$ e as respostas Y_i assumem apenas os valores 0 e 1.
- Retomando o exemplo inicial, $X_i = (X_{i1}, X_{i2})$ são o comprimento e o peso do i-ésimo peixe e Y_i é uma das duas espécies possíveis (salmão (0) ou robalo (1)).

Solução bayesiana do problema de classificação

- Como na primeira aula, temos pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ em que as preditoras X_i são vetores aleatórios em $\mathbb{R}^d$ e as respostas Y_i assumem apenas os valores 0 e 1.
- Retomando o exemplo inicial, $X_i = (X_{i1}, X_{i2})$ são o comprimento e o peso do i-ésimo peixe e Y_i é uma das duas espécies possíveis (salmão (0) ou robalo (1)).
- Por conveniência, usaremos as notações $Y = (Y_1, \ldots, Y_n)$ e $X = (X_1, \ldots, X_n)$.

Solução bayesiana do problema de classificação

- Como na primeira aula, temos pares $(X_1, Y_1), \ldots, (X_n, Y_n)$ em que as preditoras X_i são vetores aleatórios em $\mathbb{R}^d$ e as respostas Y_i assumem apenas os valores 0 e 1.
- Retomando o exemplo inicial, $X_i = (X_{i1}, X_{i2})$ são o comprimento e o peso do i-ésimo peixe e Y_i é uma das duas espécies possíveis (salmão (0) ou robalo (1)).
- Por conveniência, usaremos as notações $Y = (Y_1, \ldots, Y_n)$ e $X = (X_1, \ldots, X_n)$.
- Se observarmos uma nova preditora $X_{n+1} = x_{n+1}$, a solução bayesiana para o problema de classificação consiste na construção de um classificador $\varphi : \mathbb{R}^d \to \{0, 1\}$ baseado na probabilidade

$$\Pr\{\, Y_{n+1} = 1 \mid X_{n+1} = x_{n+1}, \{(X_i, Y_i) = (x_i, y_i)\}_{i=1}^{n} \}\,.$$

O modelo de Holmes e Adams (1)

O modelo de Holmes e Adams (1)

- De acordo com a métrica subjacente, denote por $N_k(x_i)$ os índices das k preditoras mais próximas de x_i.

O modelo de Holmes e Adams (1)

- De acordo com a métrica subjacente, denote por $N_k(x_i)$ os índices das k preditoras mais próximas de x_i.
- Introduzindo um variável aleatória $B \in \mathbb{R}_+$, que regula a "intensidade da interação entre os vizinhos", e uma variável aleatória $K \in \{1, \ldots, n\}$, que define o número de vizinhos mais próximos interagentes, Holmes e Adams (2002) propuseram o seguinte modelo.

O modelo de Holmes e Adams (1)

- De acordo com a métrica subjacente, denote por $N_k(x_i)$ os índices das k preditoras mais próximas de x_i.
- Introduzindo um variável aleatória $B \in \mathbb{R}_+$, que regula a "intensidade da interação entre os vizinhos", e uma variável aleatória $K \in \{1, \ldots, n\}$, que define o número de vizinhos mais próximos interagentes, Holmes e Adams (2002) propuseram o seguinte modelo.
- Suponha que $Y_1, \ldots, Y_n$ são condicionalmente independentes, dados X, B e K, com (função massa de) probabilidade condicional

$$f_{Y_i|X,B,K}(y_i \mid x, \beta, k) = \frac{\exp\left(\beta \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k\right)}{\sum_{t=0,1} \exp\left(\beta \sum_{j \in N_k(x_i)} I_{\{t\}}(y_j) \Big/ k\right)} \qquad (*)$$

para $i = 1, \ldots, n$.

O modelo de Holmes e Adams (2)

O modelo de Holmes e Adams (2)

- A função indicadora $I_A(x) = 1$, se $x \in A$, e $I_A(x) = 0$, se $x \notin A$.

O modelo de Holmes e Adams (2)

- A função indicadora $I_A(x) = 1$, se $x \in A$, e $I_A(x) = 0$, se $x \notin A$.
- Devido à independência condicional, deveríamos ter a seguinte expressão para a probabilidade conjunta:

$$f_{Y|X,B,K}(y \mid x, \beta, k) = \prod_{i=1}^{n} \frac{\exp\left(\beta \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k\right)}{\sum_{t=0,1} \exp\left(\beta \sum_{j \in N_k(x_i)} I_{\{t\}}(y_j) \Big/ k\right)}.$$

Contradição

Contradição

- O problema é que em $(*)$ a expressão da probabilidade condicional não depende apenas de x, β e k.

Contradição

- O problema é que em $(*)$ a expressão da probabilidade condicional não depende apenas de x, β e k.
- A probabilidade condicional $(*)$ também depende dos valores das preditoras dos k vizinhos mais próximos.

Contradição

- O problema é que em $(*)$ a expressão da probabilidade condicional não depende apenas de x, β e k.
- A probabilidade condicional $(*)$ também depende dos valores das preditoras dos k vizinhos mais próximos.
- Esta especificação incorreta leva diretamente a uma contradição: a probabilidade conjunta não está normalizada.

Contradição

- O problema é que em $(*)$ a expressão da probabilidade condicional não depende apenas de x, β e k.
- A probabilidade condicional $(*)$ também depende dos valores das preditoras dos k vizinhos mais próximos.
- Esta especificação incorreta leva diretamente a uma contradição: a probabilidade conjunta não está normalizada.
- Por exemplo, quando $n = 2$, temos que

$$\sum_{(y_1,y_2)\in\{0,1\}^2} f_{Y_1,Y_2|X,B,K}(y_1, y_2 \mid x, \beta, k) = \frac{2(1 + e^{2\beta/k})}{(1 + e^{\beta/k})^2} \neq 1.$$

Contradição

- O problema é que em $(*)$ a expressão da probabilidade condicional não depende apenas de x, β e k.
- A probabilidade condicional $(*)$ também depende dos valores das preditoras dos k vizinhos mais próximos.
- Esta especificação incorreta leva diretamente a uma contradição: a probabilidade conjunta não está normalizada.
- Por exemplo, quando $n = 2$, temos que

$$\sum_{(y_1,y_2)\in\{0,1\}^2} f_{Y_1,Y_2|X,B,K}(y_1, y_2 \mid x, \beta, k) = \frac{2(1 + e^{2\beta/k})}{(1 + e^{\beta/k})^2} \neq 1.$$

- Esta contradição foi descoberta por Cucala et al. (2009).

Condicionais completas (1)

Condicionais completas (1)

- Uma alternativa seria tentar especificar o modelo conjunto através das probabilidades condicionais completas

$$f_{Y_i \mid Y_{-i}, X, B, K}(y_i \mid y_{-i}, x, \beta, k).$$

Condicionais completas (1)

- Uma alternativa seria tentar especificar o modelo conjunto através das probabilidades condidionais completas

$$f_{Y_i \mid Y_{-i}, X, B, K}(y_i \mid y_{-i}, x, \beta, k).$$

- Esse caminho é, em geral, ardiloso.

Condicionais completas (1)

- Uma alternativa seria tentar especificar o modelo conjunto através das probabilidades condicionais completas

$$f_{Y_i \mid Y_{-i}, X, B, K}(y_i \mid y_{-i}, x, \beta, k).$$

- Esse caminho é, em geral, ardiloso.
- Suponha que $U \mid V = v \sim \mathrm{N}(v, 1)$ e $V \mid U = u \sim \mathrm{N}(u, 1)$.

Condicionais completas (1)

- Uma alternativa seria tentar especificar o modelo conjunto através das probabilidades condicionais completas

$$f_{Y_i \mid Y_{-i}, X, B, K}(y_i \mid y_{-i}, x, \beta, k).$$

- Esse caminho é, em geral, ardiloso.
- Suponha que $U \mid V = v \sim \mathrm{N}(v, 1)$ e $V \mid U = u \sim \mathrm{N}(u, 1)$.
- Existe uma densidade conjunta $f_{U,V}$ que possui estas condicionais completas?

Condicionais completas (1)

- Uma alternativa seria tentar especificar o modelo conjunto através das probabilidades condicionais completas

$$f_{Y_i \mid Y_{-i}, X, B, K}(y_i \mid y_{-i}, x, \beta, k).$$

- Esse caminho é, em geral, ardiloso.
- Suponha que $U \mid V = v \sim \mathrm{N}(v, 1)$ e $V \mid U = u \sim \mathrm{N}(u, 1)$.
- Existe uma densidade conjunta $f_{U,V}$ que possui estas condicionais completas?
- Supondo que sim e calculando formalmente com as expressões usuais, temos que

$$f_{U,V}(u, v) = f_{U \mid V}(u \mid v) f_V(v) = f_{V \mid U}(v \mid u) f_U(u);$$

$$\frac{f_{V \mid U}(v \mid u)}{f_{U \mid V}(u \mid v)} = \frac{f_V(v)}{f_U(u)}.$$

Condicionais completas (2)

Condicionais completas (2)

- Integrando em relação a y, obtemos

$$\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)} \, dv = \frac{1}{f_U(u)}.$$

Condicionais completas (2)

- Integrando em relação a y, obtemos

$$\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv = \frac{1}{f_U(u)}.$$

- Portanto, recuperaríamos a densidade conjunta fazendo

$$f_{U,V}(u, v) = \frac{f_{V|U}(v \mid u)}{\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv}.$$

Condicionais completas (2)

- Integrando em relação a y, obtemos

$$\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv = \frac{1}{f_U(u)}.$$

- Portanto, recuperaríamos a densidade conjunta fazendo

$$f_{U,V}(u,v) = \frac{f_{V|U}(v \mid u)}{\displaystyle\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv}.$$

- No entanto, em nosso exemplo a integral do denominador diverge.

Condicionais completas (2)

- Integrando em relação a y, obtemos

$$\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv = \frac{1}{f_U(u)}.$$

- Portanto, recuperaríamos a densidade conjunta fazendo

$$f_{U,V}(u,v) = \frac{f_{V|U}(v \mid u)}{\displaystyle\int_{-\infty}^{\infty} \frac{f_{V|U}(v \mid u)}{f_{U|V}(u \mid v)}\, dv}.$$

- No entanto, em nosso exemplo a integral do denominador diverge.
- Dado um conjunto de condicionais completas, precisamos de certas condições de compatibilidade (veja, por exemplo, Arnold e Press (1989)) para garantir a existência da conjunta correspondente.

Hammersley-Clifford

Hammersley-Clifford

- Há outro caminho, trilhado nos anos 70 pelos pesquisadores de sistemas em rede e processos espaciais (Besag (1974)).

Hammersley-Clifford

- Há outro caminho, trilhado nos anos 70 pelos pesquisadores de sistemas em rede e processos espaciais (Besag (1974)).
- Em nosso contexto, se postularmos que as probabilidades condicionais completas tem uma estrutura "local" (só dependem dos vizinhos mais próximos), sob certas condições, o teorema de Hammersley-Clifford determina a forma funcional da distribuição conjunta.

Hammersley-Clifford

- Há outro caminho, trilhado nos anos 70 pelos pesquisadores de sistemas em rede e processos espaciais (Besag (1974)).
- Em nosso contexto, se postularmos que as probabilidades condicionais completas tem uma estrutura "local" (só dependem dos vizinhos mais próximos), sob certas condições, o teorema de Hammersley-Clifford determina a forma funcional da distribuição conjunta.
- Este é exatamente o caminho seguido por Cucala et al. (2009).

Distribuição de Boltzmann

Distribuição de Boltzmann

- Cucala et al. adotam para o modelo conjunto uma distribuição tipo Boltzmann (também chamada de distribuição de Gibbs)

$$f_{Y|X,B,K}(y \mid x, \beta, k) = \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k\right) \Big/ Z(\beta, k).$$

Distribuição de Boltzmann

- Cucala et al. adotam para o modelo conjunto uma distribuição tipo Boltzmann (também chamada de distribuição de Gibbs)

$$f_{Y|X,B,K}(y \mid x, \beta, k) = \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right) \Big/ Z(\beta, k).$$

- A constante de normalização é dada por

$$Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right).$$

Distribuição de Boltzmann

- Cucala et al. adotam para o modelo conjunto uma distribuição tipo Boltzmann (também chamada de distribuição de Gibbs)

$$f_{Y|X,B,K}(y \mid x, \beta, k) = \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right) \Big/ Z(\beta, k).$$

- A constante de normalização é dada por

$$Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right).$$

- Em geral, a soma nesta constante de normalização é computacionalmente intratável, pois temos 2^n termos.

Distribuição de Boltzmann

- Cucala et al. adotam para o modelo conjunto uma distribuição tipo Boltzmann (também chamada de distribuição de Gibbs)

$$f_{Y|X,B,K}(y \mid x, \beta, k) = \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right) \Big/ Z(\beta, k).$$

- A constante de normalização é dada por

$$Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp\left(\beta \sum_{i=1}^{n} \sum_{j \in N_k(x_i)} I_{\{y_i\}}(y_j) \Big/ k \right).$$

- Em geral, a soma nesta constante de normalização é computacionalmente intratável, pois temos 2^n termos.
- Este modelo apresenta uma transição de frase: existe um β_{max} crítico a partir do qual que a probabilidade de todos os Y_i's serem iguais a 0 (ou iguais a 1) se aproxima rapidamente de um.

Path sampling (1)

Path sampling (1)

- Método para aproximar numericamente $Z(\beta, k)$, também conhecido como integração termodinâmica (Ogata (1989)).

Path sampling (1)

- Método para aproximar numericamente $Z(\beta, k)$, também conhecido como integração termodinâmica (Ogata (1989)).
- Defina $A(y)$ de modo que $Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp(\beta A(y))$.

Path sampling (1)

- Método para aproximar numericamente $Z(\beta, k)$, também conhecido como integração termodinâmica (Ogata (1989)).
- Defina $A(y)$ de modo que $Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp(\beta A(y))$.
- Segue que:

$$\frac{dZ(\beta, k)}{d\beta} = \sum_{y \in \{0,1\}^n} A(y) \exp(\beta A(y));$$

Path sampling (1)

- Método para aproximar numericamente $Z(\beta, k)$, também conhecido como integração termodinâmica (Ogata (1989)).
- Defina $A(y)$ de modo que $Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp(\beta A(y))$.
- Segue que:

$$\frac{dZ(\beta, k)}{d\beta} = \sum_{y \in \{0,1\}^n} A(y) \exp(\beta A(y));$$

$$\frac{1}{Z(\beta, k)} \frac{dZ(\beta, k)}{d\beta} = \sum_{y \in \{0,1\}^n} A(y) \frac{\exp(\beta A(y))}{Z(\beta, k)};$$

Path sampling (1)

- Método para aproximar numericamente $Z(\beta, k)$, também conhecido como integração termodinâmica (Ogata (1989)).
- Defina $A(y)$ de modo que $Z(\beta, k) = \sum_{y \in \{0,1\}^n} \exp(\beta A(y))$.
- Segue que:

$$\frac{dZ(\beta, k)}{d\beta} = \sum_{y \in \{0,1\}^n} A(y) \exp(\beta A(y));$$

$$\frac{1}{Z(\beta, k)} \frac{dZ(\beta, k)}{d\beta} = \sum_{y \in \{0,1\}^n} A(y) \, \frac{\exp(\beta A(y))}{Z(\beta, k)};$$

$$\frac{d \log Z(\beta, k)}{d\beta} = \mathrm{E}_\beta[A(y)].$$

Path sampling (2)

Path sampling (2)

- Integrando em β, temos

$$\log\left(\frac{Z(\beta_1,k)}{Z(\beta_0,k)}\right) = \int_{\beta_0}^{\beta_1} \mathrm{E}_\beta[A(y)]\,d\beta.$$

Path sampling (2)

- Integrando em β, temos

$$\log\left(\frac{Z(\beta_1, k)}{Z(\beta_0, k)}\right) = \int_{\beta_0}^{\beta_1} \mathrm{E}_\beta[A(y)]\, d\beta.$$

- Uma vez que $Z(0, k) = 2^n$, obtemos a aproximação

$$Z(\beta, k) = 2^n \exp\left(\int_0^\beta \mathrm{E}_{\beta'}[A(y)]\, d\beta'\right).$$

Path sampling (2)

- Integrando em β, temos

$$\log\left(\frac{Z(\beta_1,k)}{Z(\beta_0,k)}\right) = \int_{\beta_0}^{\beta_1} \mathrm{E}_\beta[A(y)]\,d\beta.$$

- Uma vez que $Z(0,k) = 2^n$, obtemos a aproximação

$$Z(\beta,k) = 2^n \exp\left(\int_0^{\beta} \mathrm{E}_{\beta'}[A(y)]\,d\beta'\right).$$

- Para aproximar a esperança acima via Monte Carlo, é possível simular a conjunta via Gibbs sampler, pois temos as condicionais completas

$$f_{Y_i\,|\,Y_{-i},X,B,K}(y_i \mid y_{-i},x,\beta,k) =$$

$$\propto \exp\left(\beta\left(\sum_{j\in N_k(x_i)} I_{\{y_i\}}(y_j) + \sum_{\{\ell\neq i:\, i\in N_k(x_\ell)\}} I_{\{y_\ell\}}(y_i)\right)\Big/ k\right).$$