Mineração de Dados

Aula 6: Finalização de Regressão e Classificação

Rafael Izbicki

Como fazer um IC para o risco estimado?

Vamos assumir que $(\tilde{\mathbf{X}}_1, \tilde{Y}_1), \ldots, (\tilde{\mathbf{X}}_s, \tilde{Y}_s)$ são elementos de um conjunto de teste, não utilizado nem para o treinamento, nem para a validação do modelo.

Um estimador não viesado para o risco de g (que foi estimada com base nos conjuntos de treinamento e validação) é dado por

$$\widehat{R}(g) = \frac{1}{s} \sum_{k=1}^{s} \underbrace{\left(\tilde{Y}_k - g(\tilde{\mathbf{X}}_k) \right)^2}_{W_k}.$$

Como trata-se de uma média de variáveis i.i.d., sabemos pelo Teorema do Limite Central que

$$\widehat{R}(g) \approx \mathit{Normal}\left(R(g), \frac{1}{s} \mathbb{V}\left[W_1 \right] \right)$$

$$\widehat{R}(g) \approx \mathit{Normal}\left(R(g), \frac{1}{s}\mathbb{V}\left[W_1\right]\right)$$

Como $W_1, \ldots, W_s$ são i.i.d.'s, podemos estimar $\mathbb{V}[W_1]$ com

$$\widehat{S^2} = \frac{1}{s}\sum_{k=1}^{s}\left(W_k - \bar{W}\right)^2,$$

onde $\bar{W} = \frac{1}{s}\sum_{k=1}^{s} W_k$.

Assim, um IC aproximado para $R(g)$ (confiança 95%) é dado por

$$\widehat{R}(g) \pm 2\sqrt{\frac{1}{s}\widehat{S^2}}$$

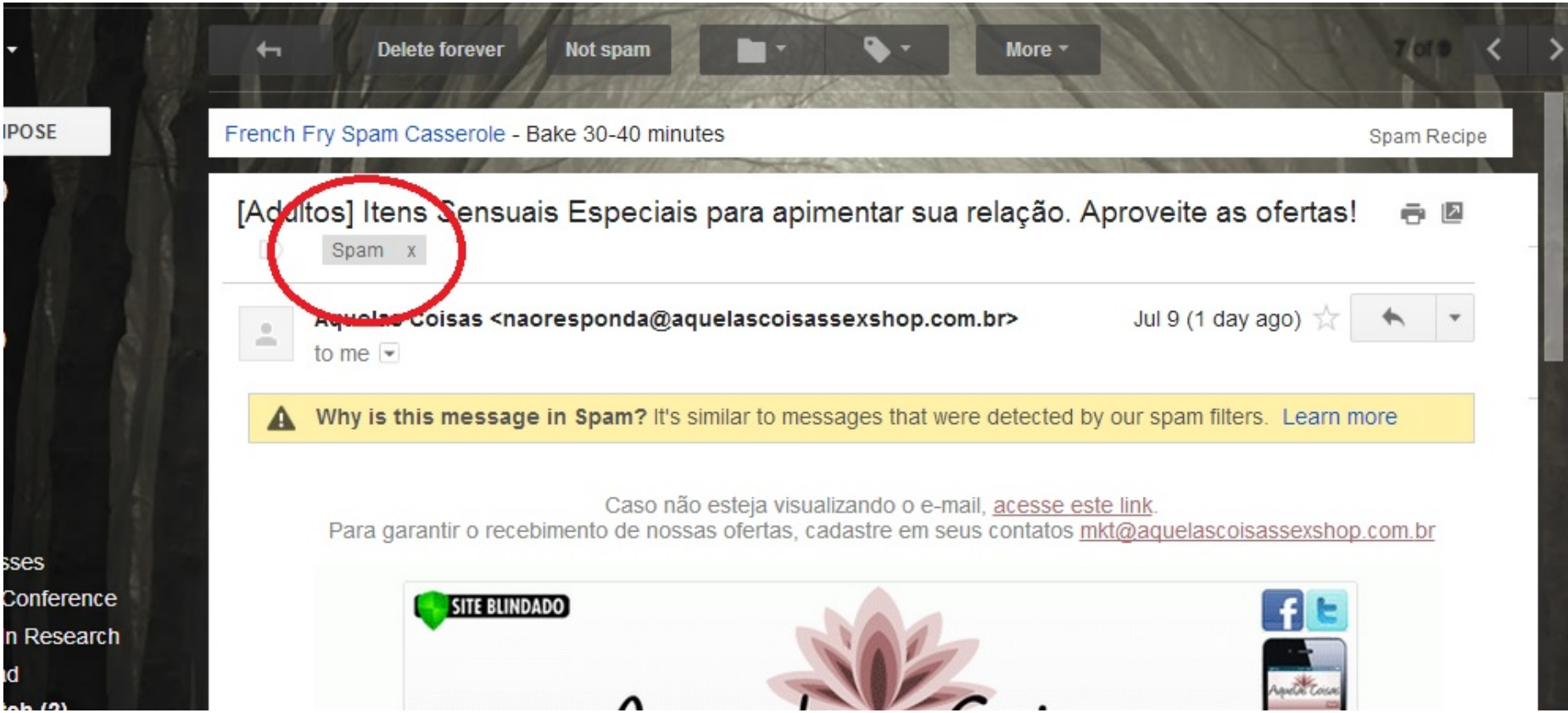
Até agora, vimos como prever uma variável quantitativa Y dado o valor de um vetor $\mathbf{x}$.

Para tanto, estudamos métodos para encontrar uma função $g(\mathbf{x})$ tal que o risco $R(g) = \mathbb{E}[(Y - g(\mathbf{x}))^2]$ fosse baixo.

Em particular, vimos que g que minimiza $R(g)$ é dada pela função de regressão $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$, e assim os métodos vistos basearam-se em estimadores para tal função.

Em muitos problemas, a variável Y é uma variável qualitativa, e não variável quantitativa.

Exemplo: Detecção de Spams

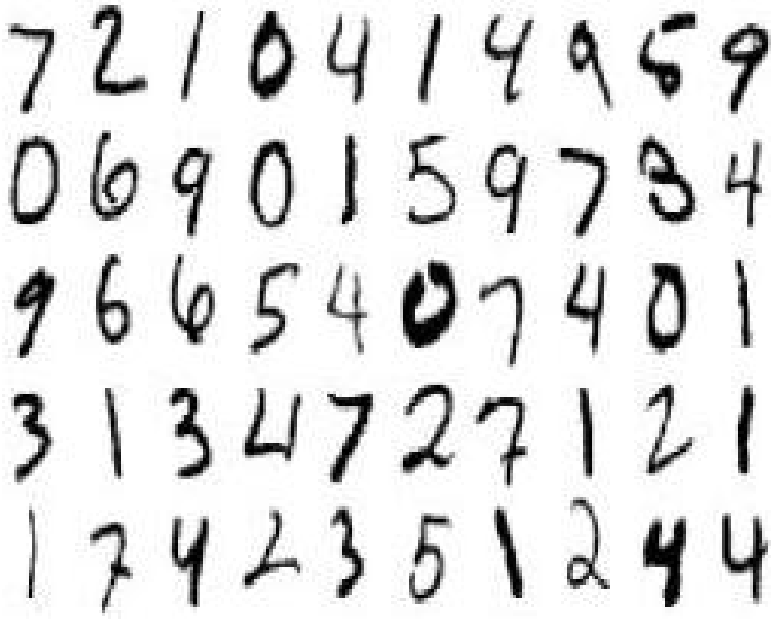


$\mathbf{X}_i \longrightarrow$ email

$Y_i \in \{\text{spam, não spam}\}$

Objetivo: prever Y_i com base em $\mathbf{X}_i$

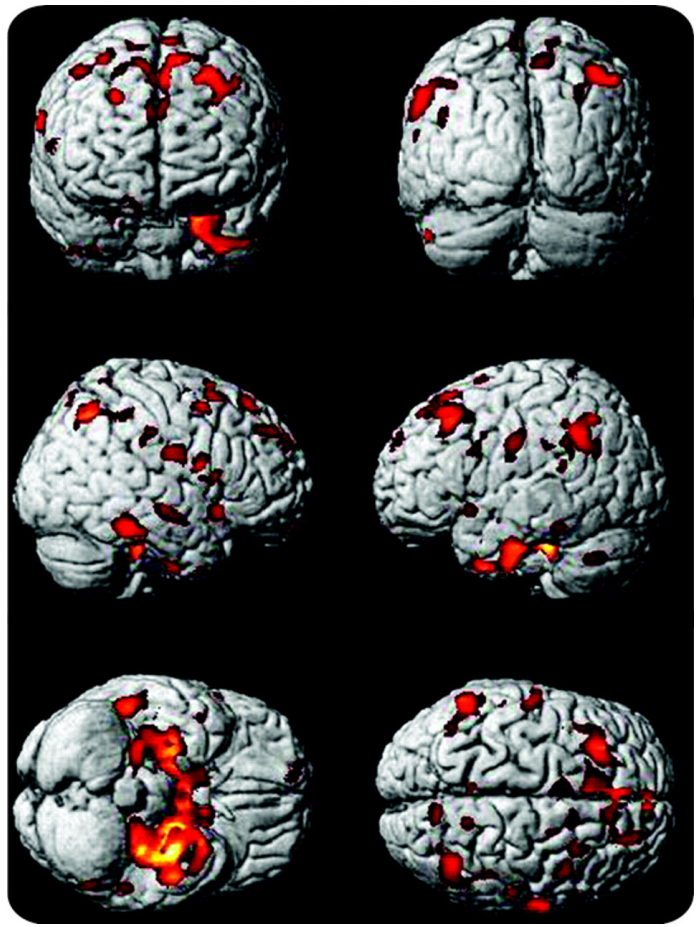
Exemplo: Reconhecimento de Dígitos



$\mathbf{X}_i \longrightarrow$ imagem de um dígito

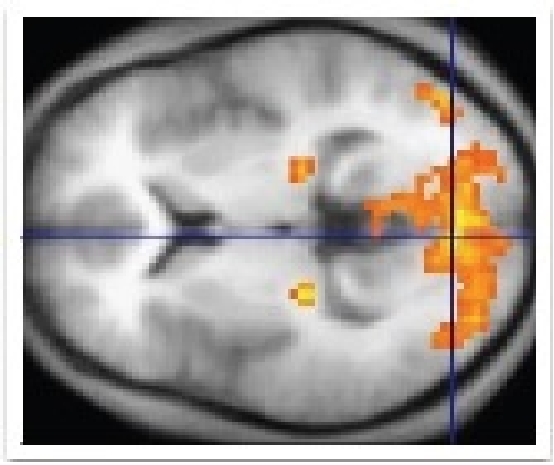
$Y_i \in \{0, 1, \ldots, 9\}$

Exemplo: Predição de Alzheimer



$\mathbf{X}_i \longrightarrow$ imagem da ressonância magnética
$Y_i \in \{\text{com Alzheimer, sem Alzheimer}\}$

Exemplo: Leitura de Pensamentos



⇓⇓⇓

$\mathbf{X}_i \longrightarrow$ imagem da ressonância magnética

$Y_i \in \{\text{Professor chato, Férias, Não to entendendo, É sexta!, } \ldots\}$

Quando Y é uma variável qualitativa, chamamos um problema de predição de um problema de classificação.

O que muda em um problema de classificação?

$\Rightarrow$ **O risco** $R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$ **não faz mais sentido**

Ao invés dele, é comum se usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X})),$$

ou seja, os risco de g é a probabilidade de erro em uma nova observação $(\mathbf{X}, Y)$.

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}\left(Y \neq g(\mathbf{X})\right).$$

Vimos que em regressão a função que minimiza $\mathbb{E}[(Y - g(\mathbf{X}))^2]$ é dada pela função de regressão $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Existe um análogo para classificação?

Sim!

Para simplificar, digamos que Y assume só dois valores, digamos, c_1 e c_2 (i.e., temos um problema *binário*, e.g, spam/não spam).

Então

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X})) =$$
$$\int_{\mathbf{x}} \mathbb{P}(Y \neq g(\mathbf{X})|\mathbf{x}) f(\mathbf{x}) d\mathbf{x} =$$
$$\int_{\mathbf{x}} [\mathbb{I}(g(\mathbf{x}) = c_2)\mathbb{P}(Y = c_1|\mathbf{x}) + \mathbb{I}(g(\mathbf{x}) = c_1)\mathbb{P}(Y = c_2|\mathbf{x})]\, f(\mathbf{x}) d\mathbf{x}$$

Para um dado $\mathbf{x}$, devemos escolher $g(\mathbf{x}) = c_1$ quando

$$\mathbb{P}(Y = c_1|\mathbf{x}) \geq \mathbb{P}(Y = c_2|\mathbf{x}),$$

caso contrário devemos escolher $g(\mathbf{x}) = c_2$.

Para um dado x, devemos escolher $g(\mathbf{x}) = c_1$ quando

$$\mathbb{P}(Y = c_1|\mathbf{x}) \geq \mathbb{P}(Y = c_2|x),$$

caso contrário devemos escolher $g(\mathbf{x}) = c_2$.

Em outras palavras, a melhor g é dada por

$$g(\mathbf{x}) = \arg \max_{d \in \{c_1, c_2\}} \mathbb{P}(Y = d|\mathbf{x})$$

Tal classificador é conhecido como classificador de Bayes (não confundir com teorema de Bayes)

Para o caso binário, ele pode ser reescrito como

$$g(\mathbf{x}) = c_1 \iff \mathbb{P}(Y = c_1|\mathbf{x}) \geq \frac{1}{2}.$$

Daqui para frente, assumiremos que Y assume valores em um conjunto $\mathcal{C}$ (e.g, $\mathcal{C} = \{\text{Spam, não spam}\}$)

No caso com várias categorias, a melhor g é dada por

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(Y = c|\mathbf{x})$$

Resultado análogo ao que diz que em um problema de predição quantitativo, a regressão é a melhor função de predição.

Obs: é comum na literatura denotar os elementos de $\mathcal{C}$ por 0, 1, 2, etc. Em particular, para o caso binário, é comum usar as classes 0 e 1. Note que tal escolha é válida, mas arbitrária (i.e., não deve-se entender que há uma ordenação entre esses elementos).

Este resultado sugere um abordagem simples para resolver um problema de predição:

(1) Estimamos $\mathbb{P}(Y = c|\mathbf{x})$, para cada categoria $c \in \mathcal{C}$.

(2) Tomamos então

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Abordagem conhecida como plug-in classifier.

Sob esta abordagem, criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Como fazer isso?

Regressão Logística

Vamos assumir por enquanto que Y é binário.

A regressão logística assume que (denotando $c_1 = 1, c_2 = 0$)

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}$$

Note que, como no caso de regressão, não estamos assumindo que está relação é válida (que o modelo é razoável para descrever os dados), mas sim apenas que ela nos leva a um bom classificador.

Ao contrário do estimador de mínimos quadrados de uma regressão, é necessário usar algoritmos numéricos para maximizar a verossimilhança induzida pela regressão logística e, assim, chegar nas estimativas para os coeficientes β.

Breve Revisão sobre Regressão Logística

Vamos assumir por enquanto que $Y \in \{0, 1\}$.

A regressão logística assume que a relação entre Y e o vetor $\mathbf{x}$ é dada por

$$\mathbb{P}(Y = 1|\mathbf{x}, \beta) = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}$$

O que significam os coeficientes β?

Assim, a função de verossimilhança dada uma amostra i.i.d. $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, condicional nas covariáveis, é dada por

$$L(y; (\mathbf{x}, \beta)) = \prod_{k=1}^{n} (\mathbb{P}(Y_k = 1|\mathbf{x}_k, \beta))^{y_k} (1 - \mathbb{P}(Y_k = 1|\mathbf{x}_k, \beta))^{1-y_k}$$

$$\prod_{k=1}^{n} \left(\frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}} \right)^{y_k} \left(\frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i}} \right)^{1-y_k}$$

Para encontrarmos as estimativas do coeficientes β, maximizamos $L(y; (\mathbf{x}, \beta))$.

Temos que fazer isso numericamente. No R:

```
glm.fit = glm(formula,data = dados, family = binomial)
```

Uma alternativa: Regressão Linear

Novamente denotando $c_1 = 1, c_2 = 0$, temos

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}].$$

Assim, podemos usar uma regressão linear para estimar $\mathbb{P}(Y = 1|\mathbf{x})$,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Nossas estimativas de $\mathbb{P}(Y = 1|\mathbf{x})$ podem ser menores que 0, e maiores que 1.

Ainda assim podemos usar estimador para definir, por exemplo,

$$g(\mathbf{x}) = \mathbb{I}(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}) \geq 1/2) = \mathbb{I}(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p \geq 1/2).$$

Podemos inclusive estimar os coeficientes usando o *lasso* e outras abordagens vistas em aula.

Uma outra abordagem

Outra abordagem para estimar $\mathbb{P}(Y = c|\mathbf{x})$ consiste em usar o Teorema de Bayes (aqui estamos assumindo que x é *contínuo*):

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Assim, pode-se obter uma estimativa de $\mathbb{P}(Y = c|\mathbf{x})$ estimando-se as probabilidades $\mathbb{P}(Y = s)$ e as densidades $f(\mathbf{x}|Y = s)$.

$\mathbb{P}(Y = s)$ pode ser facilmente estimada usando-se as proporções amostrais de cada classe.

Para estimar $f(\mathbf{x}|Y = s)$, precisamos assumir algum modelo para as covariáveis.

Naive Bayes

Uma suposição comum sobre $f(\mathbf{x}|Y = s)$ é que ela pode ser fatorada como

$$f(\mathbf{x}|Y = s) = f((x_1, \ldots, x_d)|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

i.e., supomos que as componentes de $\mathbf{x}$ são independentes condicionalmente à classe Y.

Apesar de tal suposição não ser razoável em muitos problemas (Naive=Ingênuo), ela é muito conveniente, e leva a bons classificadores.

Porque essa suposição ajuda na estimação das probabilidades condicionais?

Naive Bayes

$$f(\mathbf{x}|Y=s) = f((x_1,\ldots,x_d)|Y=s) = \prod_{j=1}^{d} f(x_j|Y=s).$$

Podemos estimar cada $f(x_j|Y=s)$ assumindo, por exemplo, que

$$X_j|Y=s \sim N(\mu_{j,s}, \sigma^2_{j,s}),\ j=1,\ldots,p$$

Em outras palavras, assumimos que cada componente do vetor $\mathbf{x}$ tem distribuição normal, com parâmetros que dependem da classe e da componente em questão.

Os parâmetros deste modelo podem ser facilmente estimados usando-se EMV:

$$\widehat{\mu_{j,s}} = \frac{1}{|\mathcal{C}_s|}\sum_{k\in\mathcal{C}_s} X_{j,k} \qquad \widehat{\sigma^2_{j,s}} = \frac{1}{|\mathcal{C}_s|}\sum_{k\in\mathcal{C}_s} (X_{j,k}-\widehat{\mu}_{j,s})^2$$

em que $\mathcal{C}_s = \{j : Y_j = s\}$ é o conjunto de todas observações de treinamento da classe s.

Assim,

$$\widehat{f}(\mathbf{x}|Y=c) = \prod_{k=1}^{d} \widehat{f}(x_k|Y=c) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi\widehat{\sigma^2}_{k,s}}} e^{-\left(\frac{(x_k-\widehat{\mu_{k,s}})^2}{2\widehat{\sigma^2_{k,s}}}\right)}$$

Claro, podemos supor outras distribuições que não a distribuição normal.

Em particular, se $\mathbf{X}$ tem componentes discretas, o Teorema de Bayes diz que

$$\mathbb{P}(Y = c|x) = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} \mathbb{P}(\mathbf{X} = \mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Neste caso, é comum assumir que

$$X_j|Y = c \sim Multinomial(1, \theta_{j,c}),$$

onde $\theta_{j,c} \in \Re^q$ é um vetor com q dimensões, o número de categorias que X_j assume.

Computacionalmente: para dados discretos, há a função `naiveBayes` no pacote `e1071` do R.

Para dados contínuos, implementar sua própria rotina.

Dica: calcular produtos como $\prod_{k=1}^{d} \widehat{f}(x_j|Y=c)$ é difícil numericamente, pois eles pequenos e o produto acaba sendo aproximado por 0.

Uma alternativa é trabalhar com logaritmos:

$$\log \widehat{f}(\mathbf{x}|Y=c) = \log\left(\prod_{k=1}^{d} \widehat{f}(x_j|Y=c)\right) = \sum_{k=1}^{d} \log\left(\widehat{f}(x_j|Y=c)\right)$$

Na hora de comparar densidades usamos somente esses logaritmos. Isto é, usamos que o classificador plugin pode ser escrito como

$$\begin{aligned} g(\mathbf{x}) &= \arg\max_{c\in\mathcal{C}} \widehat{\mathbb{P}}(Y=c|\mathbf{x}) = \arg\max_{c\in\mathcal{C}} \widehat{f}(\mathbf{x}|Y=c)\widehat{\mathbb{P}}(Y=c) \\ &= \arg\max_{c\in\mathcal{C}} \left(\prod_{k=1}^{d} \widehat{f}(x_j|Y=c)\right)\widehat{\mathbb{P}}(Y=c) \\ &= \arg\max_{c\in\mathcal{C}} \sum_{k=1}^{d} \log\left(\widehat{f}(x_j|Y=c)\right) + \log\widehat{\mathbb{P}}(Y=c) \end{aligned}$$

Seleção de Modelos

Como estimamos o novo risco

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))?$$

Podemos novamente separar nossos dados em treinamento e validação, e usar

$$\widehat{R}(g) := \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(Y_k' \neq g(\mathbf{X}_k')),$$

onde $(\mathbf{X}_1', Y_1'), \ldots, (\mathbf{X}_m', Y_m')$ é o conjunto de validação.

Usando estimativas do risco podemos selecionar um entre vários classificadores.

Podemos também criar um IC para o risco.

Exemplo: Detecção de SPAMs

$n = 4.601$ observações de emails. Alguns são SPAMs, outros não.

Medimos a frequência relativa de algumas palavras:

`internet, free, credit, money, data, technology, direct, ...`

Resultados:

Regressão logística: Risco Estimado=0.086

Regressão linear: Risco Estimado=0.111

Naive Bayes: Risco Estimado=0.188

Revisão

Em um problema de classificação, é comum se usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}\left(Y \neq g(\mathbf{X})\right),$$

para medir o risco de um classificador g.

A função g que minimza $R(g)$ é dada por

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(Y = c|\mathbf{x})$$

Tal classificador é conhecido como classificador de Bayes (não confundir com teorema de Bayes)

Como no caso de regressão, não conhecemos tal g, mas podemos estimá-la usando

$$\widehat{g}(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{\mathbb{P}}(Y = c|\mathbf{x})$$

Como no caso de regressão, não conhecemos tal g, mas podemos estimá-la usando

$$\widehat{g}(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{\mathbb{P}}(Y = c|\mathbf{x})$$

A estimativa $\widehat{\mathbb{P}}(Y = c|\mathbf{x})$ pode ser obtida usando-se, por exemplo, regressão logística ou regressão linear.

Como no problema de regressão, não assumimos que esta relação (logística ou linear) é válida, mas apenas que ela criar bons classificadores.

Vimos também uma outra abordagem.

Esta consiste em usar o Teorema de Bayes:

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Estimamos então tanto $f(\mathbf{x}|Y = c)$ quanto $\mathbb{P}(Y = c)$.

A probabilidade $\mathbb{P}(Y = c)$ pode ser facilmente estimada utilizando-se as proporções amostrais de cada categoria de Y.

Já para estimar $f(\mathbf{x}|Y = c)$, é comum fazermos alguma suposição sobre esta distribuição.

Aqui vimos o Naive Bayes, que assume independência condicional das covariáveis:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_p|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

Finalmente, vimos que podemos selecionar modelos utilizando-se validação cruzada.

Para tanto, podemos utilizar o seguinte estimador do risco:

$$\widehat{R}(g) := \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(Y_k' \neq g(\mathbf{X}_k')),$$

Próxima Aula: