Aula 1: k-Nearest Neighbors

Paulo C. Marques F.

Aula ministrada no Insper

29 de Janeiro de 2016

O problema geral de classificação

O problema geral de classificação

- Imagine uma fábrica na qual temos uma esteira pela qual descem peixes de duas espécies: salmões e robalos.

O problema geral de classificação

- Imagine uma fábrica na qual temos uma esteira pela qual descem peixes de duas espécies: salmões e robalos.
- Nosso objetivo é construir uma máquina que, tomando o peso e o maior comprimento dos peixes, classifique cada um deles como salmão ou robalo.

O problema geral de classificação

- Imagine uma fábrica na qual temos uma esteira pela qual descem peixes de duas espécies: salmões e robalos.
- Nosso objetivo é construir uma máquina que, tomando o peso e o maior comprimento dos peixes, classifique cada um deles como salmão ou robalo.
- Formalmente, temos um vetor de variáveis *preditoras* $x \in \mathbb{R}^d$, uma variável *resposta* $y \in C = \{0, \ldots, c\}$, e queremos construir um *classificador* $\varphi : \mathbb{R}^d \to C$.

O problema geral de classificação

- Imagine uma fábrica na qual temos uma esteira pela qual descem peixes de duas espécies: salmões e robalos.
- Nosso objetivo é construir uma máquina que, tomando o peso e o maior comprimento dos peixes, classifique cada um deles como salmão ou robalo.
- Formalmente, temos um vetor de variáveis *preditoras* $x \in \mathbb{R}^d$, uma variável *resposta* $y \in C = \{0, \ldots, c\}$, e queremos construir um *classificador* $\varphi : \mathbb{R}^d \to C$.
- Sem perda de generalidade, vamos supor que temos apenas duas classes: $C = \{0, 1\}$ (salmão e robalo).

O problema geral de classificação

- Imagine uma fábrica na qual temos uma esteira pela qual descem peixes de duas espécies: salmões e robalos.
- Nosso objetivo é construir uma máquina que, tomando o peso e o maior comprimento dos peixes, classifique cada um deles como salmão ou robalo.
- Formalmente, temos um vetor de variáveis *preditoras* $x \in \mathbb{R}^d$, uma variável *resposta* $y \in C = \{0, \ldots, c\}$, e queremos construir um *classificador* $\varphi : \mathbb{R}^d \to C$.
- Sem perda de generalidade, vamos supor que temos apenas duas classes: $C = \{0, 1\}$ (salmão e robalo).
- O caso em que existe um classificador φ que nunca erra é de pouco interesse prático/científico. Existem salmões e robalos que tem exatamente o mesmo peso e comprimento.

Aprendizagem supervisionada

Aprendizagem supervisionada

- Introduzimos incertezas em nossa descrição através de um vetor aleatório $(X, Y) \in \mathbb{R}^d \times C$ com função de distribuição conjunta $F_{X,Y}$.

Aprendizagem supervisionada

- Introduzimos incertezas em nossa descrição através de um vetor aleatório $(X, Y) \in \mathbb{R}^d \times C$ com função de distribuição conjunta $F_{X,Y}$.
- Metáfora: um oráculo gera um x a partir da função de distribuição marginal F_X e depois disso gera um y a partir da função de distribuição condicional $F_{X|Y}(\cdot \mid x)$ (conhecida como distribuição do supervisor).

Aprendizagem supervisionada

- Introduzimos incertezas em nossa descrição através de um vetor aleatório $(X, Y) \in \mathbb{R}^d \times C$ com função de distribuição conjunta $F_{X,Y}$.
- Metáfora: um oráculo gera um x a partir da função de distribuição marginal F_X e depois disso gera um y a partir da função de distribuição condicional $F_{X|Y}(\,\cdot \mid x)$ (conhecida como distribuição do supervisor).
- A função de distribuição conjunta fica determinada formalmente por

$$F_{X,Y}(x, y) = \int_{(-\infty, x]} F_{Y|X}(y \mid t)\, dF_X(t),$$

na qual usamos a notação $(-\infty, x] := (-\infty, x_1] \times \cdots \times (-\infty, x_d]$.

Aprendizagem supervisionada

- Introduzimos incertezas em nossa descrição através de um vetor aleatório $(X, Y) \in \mathbb{R}^d \times C$ com função de distribuição conjunta $F_{X,Y}$.
- Metáfora: um oráculo gera um x a partir da função de distribuição marginal F_X e depois disso gera um y a partir da função de distribuição condicional $F_{X|Y}(\,\cdot \mid x)$ (conhecida como distribuição do supervisor).
- A função de distribuição conjunta fica determinada formalmente por

$$F_{X,Y}(x, y) = \int_{(-\infty, x]} F_{Y|X}(y \mid t)\, dF_X(t),$$

  na qual usamos a notação $(-\infty, x] := (-\infty, x_1] \times \cdots \times (-\infty, x_d]$.
- Estamos fazendo inferência: não conhecemos $F_{X,Y}$.

Aprendizagem supervisionada

- Introduzimos incertezas em nossa descrição através de um vetor aleatório $(X, Y) \in \mathbb{R}^d \times C$ com função de distribuição conjunta $F_{X,Y}$.
- Metáfora: um oráculo gera um x a partir da função de distribuição marginal F_X e depois disso gera um y a partir da função de distribuição condicional $F_{X|Y}(\,\cdot \mid x)$ (conhecida como distribuição do supervisor).
- A função de distribuição conjunta fica determinada formalmente por

$$F_{X,Y}(x, y) = \int_{(-\infty, x]} F_{Y|X}(y \mid t)\, dF_X(t),$$

  na qual usamos a notação $(-\infty, x] := (-\infty, x_1] \times \cdots \times (-\infty, x_d]$.
- Estamos fazendo inferência: não conhecemos $F_{X,Y}$.
- Nosso contexto é não paramétrico: a menos do suporte, não impomos quaisquer restrições a $F_{X,Y}$.

Erro de classificação

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.
- Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.
- Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

- O classificador de Bayes tem um papel formal: em um problema real não conhecemos $F_{X,Y}$. Portanto, também não conhecemos $\eta(x)$ e não conseguimos construir φ^*.

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.
- Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

- O classificador de Bayes tem um papel formal: em um problema real não conhecemos $F_{X,Y}$. Portanto, também não conhecemos $\eta(x)$ e não conseguimos construir φ^*.
- Apesar do nome, não estamos fazendo inferência bayesiana.

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.
- Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

- O classificador de Bayes tem um papel formal: em um problema real não conhecemos $F_{X,Y}$. Portanto, também não conhecemos $\eta(x)$ e não conseguimos construir φ^*.
- Apesar do nome, não estamos fazendo inferência bayesiana.
- Quando temos dados simulados a partir de uma distribuição conhecida, podemos construir o classificador de Bayes φ^*.

Erro de classificação

- Definimos o *erro de classificação* de um classificador φ pela probabilidade do classificador errar: $L[\varphi] = \Pr\{\varphi(X) \neq Y\}$.
- Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

- O classificador de Bayes tem um papel formal: em um problema real não conhecemos $F_{X,Y}$. Portanto, também não conhecemos $\eta(x)$ e não conseguimos construir φ^*.
- Apesar do nome, não estamos fazendo inferência bayesiana.
- Quando temos dados simulados a partir de uma distribuição conhecida, podemos construir o classificador de Bayes φ^*.
- O classificador de Bayes é ótimo: para qualquer classificador φ, temos que $L[\varphi^*] \leq L[\varphi]$.

O classificador de Bayes é ótimo (1)

Demonstração

Para qualquer classificador φ e todo $x \in \mathbb{R}^d$, note que $\Pr\{\varphi(X) = Y \mid X = x\} = \Pr\{\varphi(x) = Y \mid X = x\}$ é igual a $\Pr\{Y = 0 \mid X = x\} = 1 - \eta(x)$, quando $\varphi(x) = 0$, e é igual a $\Pr\{Y = 1 \mid X = x\} = \eta(x)$, quando $\varphi(x) = 1$. Assim,

$$\begin{aligned}\Pr\{\varphi(X) \neq Y \mid X = x\} &= 1 - \Pr\{\varphi(X) = Y \mid X = x\} \\ &= 1 - \left(I_{\{\varphi(x)=0\}}(1 - \eta(x)) + I_{\{\varphi(x)=1\}}\eta(x)\right) \\ &= \eta(x) - (2\eta(x) - 1)I_{\{\varphi(x)=1\}},\end{aligned}$$

uma vez que $I_{\{\varphi(x)=0\}} = 1 - I_{\{\varphi(x)=1\}}$. Portanto,

$$\Pr\{\varphi(X) \neq Y \mid X = x\} - \Pr\{\varphi^*(X) \neq Y \mid X = x\}$$

$$= (2\eta(x) - 1)\left(I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}}\right).$$

O classificador de Bayes é ótimo (2)

Demonstração (continuação)

Temos dois casos: se $\varphi^*(x) = 0$, então, pela definição do classificador de Bayes, temos que $2\eta(x) - 1 \leq 0$ e $I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}} \leq 0$. Quando $\varphi^*(x) = 1$, temos que $2\eta(x) - 1 \geq 0$ e $I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}} \geq 0$. Assim, em ambos os casos, temos que

$$\Pr\{\varphi(X) \neq Y \mid X = x\} - \Pr\{\varphi^*(X) \neq Y \mid X = x\} \geq 0. \qquad (*)$$

Pela definição de probabilidade condicional, para todo classificador φ, temos que

$$\Pr\{\varphi(X) \neq Y\} = \int_{(-\infty, x]} \Pr\{\varphi(X) \neq Y \mid X = x\}\, dF_X(x).$$

Obtemos o resultado desejado integrando $(*)$ com respeito a $dF_X(x)$.

Vapnik e Chervonenkis (1)

Vapnik e Chervonenkis (1)

- Já que em geral o classificador de Bayes é inacessível, como escolher um "bom" classificador dentro de uma classe de classificadores $\mathscr{C} = \{\varphi_1, \ldots, \varphi_m\}$?

Vapnik e Chervonenkis (1)

- Já que em geral o classificador de Bayes é inacessível, como escolher um "bom" classificador dentro de uma classe de classificadores $\mathscr{C} = \{\varphi_1, \ldots, \varphi_m\}$?
- Dada uma amostra de pares

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

independentes e identicamente distribuídos com função de distribuição $F_{X,Y}$.

Vapnik e Chervonenkis (1)

- Já que em geral o classificador de Bayes é inacessível, como escolher um "bom" classificador dentro de uma classe de classificadores $\mathscr{C} = \{\varphi_1, \ldots, \varphi_m\}$?
- Dada uma amostra de pares

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

independentes e identicamente distribuídos com função de distribuição $F_{X,Y}$.

- Defina o *erro empírico de classificação* de um classificador $\varphi \in \mathscr{C}$ por

$$\hat{L}_n[\varphi] = \frac{1}{n} \sum_{i=1}^{n} I_{\{\varphi(X_i) \neq Y_i\}}.$$

Vapnik e Chervonenkis (2)

Vapnik e Chervonenkis (2)

- Vapnik e Chernonenkis preconizam que devemos escolher o classificador $\hat{\varphi}$ que minimiza o erro empírico:

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{C}} \hat{L}_n[\varphi].$$

Vapnik e Chervonenkis (2)

- Vapnik e Chernonenkis preconizam que devemos escolher o classificador $\hat{\varphi}$ que minimiza o erro empírico:

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{C}} \hat{L}_n[\varphi].$$

- Note-se que, pela lei forte dos grandes números, $\hat{L}_n[\varphi]$ é um estimador fortemente consistente de $L[\varphi]$, ou seja, $\hat{L}_n[\varphi] \to L[\varphi]$ com probabilidade 1, quando $n \to \infty$, para toda $F_{X,Y}$.

Vapnik e Chervonenkis (2)

- Vapnik e Chernonenkis preconizam que devemos escolher o classificador $\hat{\varphi}$ que minimiza o erro empírico:

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{C}} \hat{L}_n[\varphi].$$

- Note-se que, pela lei forte dos grandes números, $\hat{L}_n[\varphi]$ é um estimador fortemente consistente de $L[\varphi]$, ou seja, $\hat{L}_n[\varphi] \to L[\varphi]$ com probabilidade 1, quando $n \to \infty$, para toda $F_{X,Y}$.
- Muito importante: isto não ocorreria, em geral, se o classificador φ fosse uma função de toda a amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Vapnik e Chervonenkis (2)

- Vapnik e Chernonenkis preconizam que devemos escolher o classificador $\hat{\varphi}$ que minimiza o erro empírico:

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{C}} \hat{L}_n[\varphi].$$

- Note-se que, pela lei forte dos grandes números, $\hat{L}_n[\varphi]$ é um estimador fortemente consistente de $L[\varphi]$, ou seja, $\hat{L}_n[\varphi] \to L[\varphi]$ com probabilidade 1, quando $n \to \infty$, para toda $F_{X,Y}$.
- Muito importante: isto não ocorreria, em geral, se o classificador φ fosse uma função de toda a amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Vale lembrar: φ é simplesmente uma função de $\mathbb{R}^d$ em C. O classificador φ não é um objeto aleatório.

Vapnik e Chervonenkis (2)

- Vapnik e Chernonenkis preconizam que devemos escolher o classificador $\hat{\varphi}$ que minimiza o erro empírico:

$$\hat{\varphi} = \arg\min_{\varphi \in \mathscr{C}} \hat{L}_n[\varphi].$$

- Note-se que, pela lei forte dos grandes números, $\hat{L}_n[\varphi]$ é um estimador fortemente consistente de $L[\varphi]$, ou seja, $\hat{L}_n[\varphi] \to L[\varphi]$ com probabilidade 1, quando $n \to \infty$, para toda $F_{X,Y}$.
- Muito importante: isto não ocorreria, em geral, se o classificador φ fosse uma função de toda a amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Vale lembrar: φ é simplesmente uma função de $\mathbb{R}^d$ em C. O classificador φ não é um objeto aleatório.
- Menos importante: o estimador é não viciado: $\mathrm{E}[\hat{L}_n[\varphi]] = L[\varphi]$.

Vapnik e Chervonenkis (3)

Vapnik e Chervonenkis (3)

- Desigualdade de Hoeffding: sejam $U_1, \ldots U_n$ variáveis aleatórias independentes tais que $\Pr\{a_i \leq U_i \leq b_i]\} = 1$. Definindo $\bar{U}_n = (U_1 + \cdots + U_n)/n$, temos que

$$\Pr\left\{\left|\bar{U}_n - \mathrm{E}[\bar{U}_n]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Vapnik e Chervonenkis (3)

- Desigualdade de Hoeffding: sejam $U_1, \ldots U_n$ variáveis aleatórias independentes tais que $\Pr\{a_i \leq U_i \leq b_i]\} = 1$. Definindo $\bar{U}_n = (U_1 + \cdots + U_n)/n$, temos que

$$\Pr\left\{\left|\bar{U}_n - \mathrm{E}[\bar{U}_n]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

- Usando a desigualdade de Hoeffding, temos que

$$\Pr\left\{\left|\hat{L}_n[\hat{\varphi}] - L[\hat{\varphi}]\right| \geq \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

Vapnik e Chervonenkis (3)

- Desigualdade de Hoeffding: sejam $U_1, \dots U_n$ variáveis aleatórias independentes tais que $\Pr\{a_i \leq U_i \leq b_i]\} = 1$. Definindo $\bar{U}_n = (U_1 + \cdots + U_n)/n$, temos que

$$\Pr\left\{\left|\bar{U}_n - \mathrm{E}[\bar{U}_n]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

- Usando a desigualdade de Hoeffding, temos que

$$\Pr\left\{\left|\hat{L}_n[\hat{\varphi}] - L[\hat{\varphi}]\right| \geq \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

- Portanto, para algum $\alpha = (0, 1]$, fazendo $2e^{-2n\epsilon^2} = \alpha$, temos que

$$\hat{L}_n[\hat{\varphi}] \pm \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

é um intervalo de confiança para $L[\hat{\varphi}]$ com nível de confiança não menor do que $(1 - \alpha)\%$.

k-Nearest Neighbors (1)

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).
- Diversas distâncias podem ser utilizadas.

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).
- Diversas distâncias podem ser utilizadas.
- Euclidiana: $d(x, z) = \sqrt{(x - z)^\top (x - z)}$.

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).
- Diversas distâncias podem ser utilizadas.
- Euclidiana: $d(x, z) = \sqrt{(x - z)^\top (x - z)}$.
- Mahalanobis: $d(x, z) = \sqrt{(x - z)^\top S^{-1} (x - z)}$, em que S é a matriz de covariâncias amostral.

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).
- Diversas distâncias podem ser utilizadas.
- Euclidiana: $d(x, z) = \sqrt{(x - z)^\top (x - z)}$.
- Mahalanobis: $d(x, z) = \sqrt{(x - z)^\top S^{-1} (x - z)}$, em que S é a matriz de covariâncias amostral.
- E muitas outras. Especialmente quando algumas das preditoras são categóricas.

k-Nearest Neighbors (1)

- Para uma certa distância definida em $\mathbb{R}^d$, dados n pares $(x_1, y_1), \ldots, (x_n, y_n)$ e um novo $x \in \mathbb{R}^d$, o classificador k-NN determina os k pontos em $\{x_1, \ldots, x_n\}$ mais próximos de x e classifica x como pertencente à classe mais frequente entre os y_i's destes k vizinhos mais próximos (voto da maioria).
- Diversas distâncias podem ser utilizadas.
- Euclidiana: $d(x, z) = \sqrt{(x - z)^\top (x - z)}$.
- Mahalanobis: $d(x, z) = \sqrt{(x - z)^\top S^{-1} (x - z)}$, em que S é a matriz de covariâncias amostral.
- E muitas outras. Especialmente quando algumas das preditoras são categóricas.
- Quando temos muitas preditoras, o classificador k-NN sofre a "maldição da dimensionalidade", pois, grosso modo, em um espaço euclidiano de dimensão muito alta todos os x_i's estariam aproximadamente à mesma distância da origem.

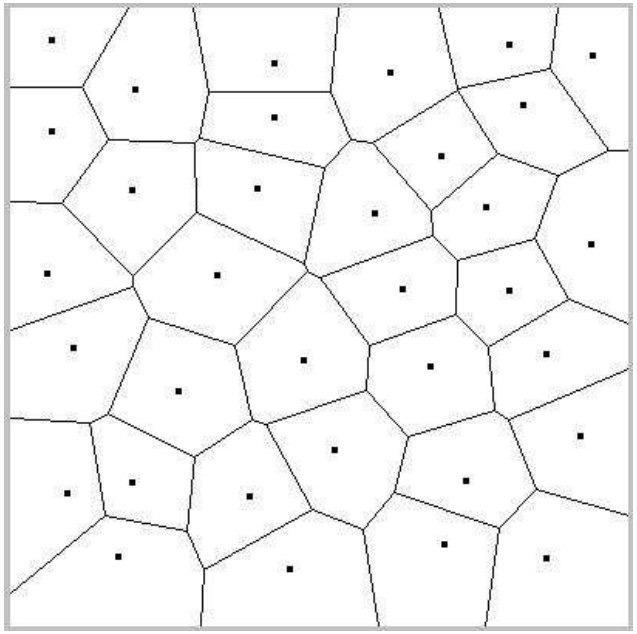
k-Nearest Neighbors (2)

k-Nearest Neighbors (2)

- No caso $k = 1$, os pontos $x_1, \ldots, x_n$ definem células de classificação que formam uma estrutura geométrica conhecida como tesselação (mosaico) de Voronoi.

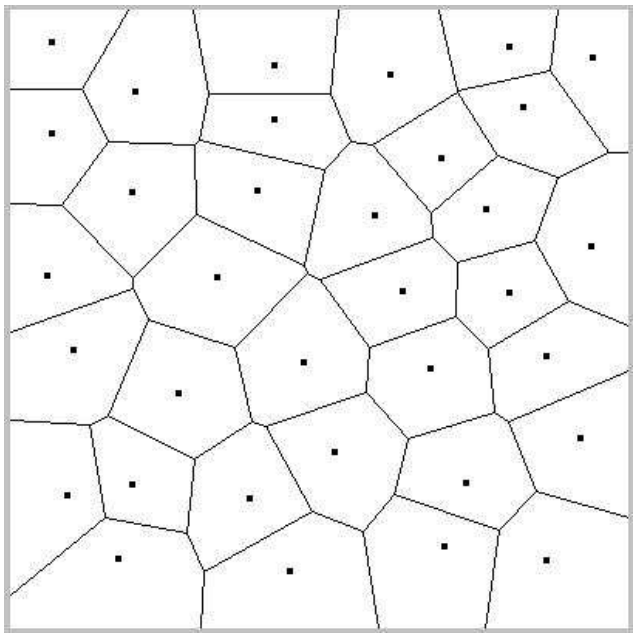
k-Nearest Neighbors (2)

- No caso $k = 1$, os pontos $x_1, \ldots, x_n$ definem células de classificação que formam uma estrutura geométrica conhecida como tesselação (mosaico) de Voronoi.



k-Nearest Neighbors (2)

- No caso $k = 1$, os pontos $x_1, \ldots, x_n$ definem células de classificação que formam uma estrutura geométrica conhecida como tesselação (mosaico) de Voronoi.



- Cover e Hart provaram que, assintoticamente, o erro de classificação da regra 1-NN nunca é maior do que o dobro do erro de Bayes, de maneira universal, ou seja, para qualquer $F_{X,Y}$.

Como escolher k? (1)

Como escolher k? (1)

- A escolha de k é crítica. As regiões de classificação podem ser substancialmente diferentes para k's distintos.

Como escolher k? (1)

- A escolha de k é crítica. As regiões de classificação podem ser substancialmente diferentes para k's distintos.
- Se, erroneamente, tentassemos minimizar o "erro empírico" do classificador contruído com toda a amostra, escolheríamos sempre $k = 1$, pois a regra 1-NN, aparentemente, teria "erro empírico" igual a zero. Conforme discutido em slides anteriores, esta interpretação é incorreta.

Como escolher k? (1)

- A escolha de k é crítica. As regiões de classificação podem ser substancialmente diferentes para k's distintos.
- Se, erroneamente, tentassemos minimizar o "erro empírico" do classificador contruído com toda a amostra, escolheríamos sempre $k = 1$, pois a regra 1-NN, aparentemente, teria "erro empírico" igual a zero. Conforme discutido em slides anteriores, esta interpretação é incorreta.
- O procedimento clássico em Statistical Learning é dividir a amostra em m *dados de treinamento* e $n - m$ *dados de teste*:

$$\underbrace{(X_1, Y_1), \ldots, (X_m, Y_m)}_{\text{dados de treinamento}}, \underbrace{(X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)}_{\text{dados de teste}}.$$

Como escolher k? (2)

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.
- Há critérios assintóticos (tipo Stone) para a divisão da amostra em dados de treinamento e dados de teste.

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.
- Há critérios assintóticos (tipo Stone) para a divisão da amostra em dados de treinamento e dados de teste.
- Não há critérios universais para n finito.

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.
- Há critérios assintóticos (tipo Stone) para a divisão da amostra em dados de treinamento e dados de teste.
- Não há critérios universais para n finito.
- O que se vê entre os praticamentes são critérios de divisão do tipo 70-30.

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.
- Há critérios assintóticos (tipo Stone) para a divisão da amostra em dados de treinamento e dados de teste.
- Não há critérios universais para n finito.
- O que se vê entre os praticamentes são critérios de divisão do tipo 70-30.
- Matematicamente, o classificador obtido depende de como a amostra foi dividida.

Como escolher k? (2)

- Usa-se os dados de treinamento para construir classificadores com $k = 1, \ldots, m$ e calcula-se o erro empírico de cada classificador usando *apenas* os $n - m$ dados de teste. O menor erro empírico determina o valor de k.
- Há critérios assintóticos (tipo Stone) para a divisão da amostra em dados de treinamento e dados de teste.
- Não há critérios universais para n finito.
- O que se vê entre os praticamentes são critérios de divisão do tipo 70-30.
- Matematicamente, o classificador obtido depende de como a amostra foi dividida.
- Uma variante da k-NN é utilizada para regressão não paramétrica: ao invés do voto da maioria, toma-se a média das respostas dos k vizinhos mais próximos.

Obrigado pela presença!