Aula 17 - Outros Métodos Supervisionados

João Florindo

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned} \operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y). \end{aligned}$$

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned}\operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y).\end{aligned}$$

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned}\operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y).\end{aligned}$$

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned}\operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y).\end{aligned}$$

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned}\operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y).\end{aligned}$$

- Vimos até agora modelos para $p(y|x;\theta)$. EX.: Na regressão logística, $p(y=1|x;\theta) = g(\theta^T X)$.
- Essa é a abordagem **discriminativa**.
- EX.: Separar cachorros de elefantes.
- Mas temos também a **generativa**, que modela $p(x|y)$ (e a probabilidade *a priori* da classe $p(y)$).
- A probabilidade *posteriori* é obtida por Bayes:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Para fazer predições, nem precisamos do denominador pois

$$\begin{aligned}\operatorname*{argmax}_y p(y|x) &= \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x|y)p(y).\end{aligned}$$

Análise Discriminante Gaussiana (GDA)

- Assume que $p(x|y)$ segue uma distribuição normal (Gaussiana) multivariada.
- Em n dimensões, com o **vetor média** $\mu \in \mathbb{R}^n$ e a **matriz de covariância** $\Sigma \in \mathbb{R}^{n \times n}$, temos

$$p(x|y) \sim \mathcal{N}(\mu, \Sigma),$$

com densidade

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

- A **covariância** de uma variável aleatória vetorial Z é

$$Cov(Z) = E[(Z - E[Z])(Z - E[Z])^T].$$

Análise Discriminante Gaussiana (GDA)

- Assume que $p(x|y)$ segue uma distribuição normal (Gaussiana) multivariada.
- Em n dimensões, com o **vetor média** $\mu \in \mathbb{R}^n$ e a **matriz de covariância** $\Sigma \in \mathbb{R}^{n \times n}$, temos

$$p(x|y) \sim \mathcal{N}(\mu, \Sigma),$$

com densidade

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

- A **covariância** de uma variável aleatória vetorial Z é

$$Cov(Z) = E[(Z - E[Z])(Z - E[Z])^T].$$

Análise Discriminante Gaussiana (GDA)

- Assume que $p(x|y)$ segue uma distribuição normal (Gaussiana) multivariada.
- Em n dimensões, com o **vetor média** $\mu \in \mathbb{R}^n$ e a **matriz de covariância** $\Sigma \in \mathbb{R}^{n \times n}$, temos

$$p(x|y) \sim \mathcal{N}(\mu, \Sigma),$$

  com densidade

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right).$$

- A **covariância** de uma variável aleatória vetorial Z é

$$Cov(Z) = E[(Z - E[Z])(Z - E[Z])^T].$$

Análise Discriminante Gaussiana (GDA)

- Problema de classificação em que x é uma variável contínua.
- Temos então

$$y \sim \text{Bernoulli}(\phi)$$
$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$$
$$x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

- Escrevendo as distribuições:

$$y = \phi^y (1 - \phi)^{1-y}$$
$$p(x|y = 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1} (x - \mu_0)\right)$$
$$p(x|y = 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1} (x - \mu_1).\right)$$

- Note que temos vetores μ_0 e μ_1 mas uma única Σ.

Análise Discriminante Gaussiana (GDA)

- Problema de classificação em que x é uma variável contínua.
- Temos então

$$y \sim \text{Bernoulli}(\phi)$$
$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$$
$$x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

- Escrevendo as distribuições:

$$y = \phi^y(1-\phi)^{1-y}$$
$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$
$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1).\right)$$

- Note que temos vetores μ_0 e μ_1 mas uma única Σ.

Análise Discriminante Gaussiana (GDA)

- Problema de classificação em que x é uma variável contínua.
- Temos então

$$
\begin{aligned}
y &\sim \text{Bernoulli}(\phi) \\
x|y = 0 &\sim \mathcal{N}(\mu_0, \Sigma) \\
x|y = 1 &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}
$$

- Escrevendo as distribuições:

$$
\begin{aligned}
y &= \phi^y (1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1).\right)
\end{aligned}
$$

- Note que temos vetores μ_0 e μ_1 mas uma única Σ.

Análise Discriminante Gaussiana (GDA)

- Problema de classificação em que x é uma variável contínua.
- Temos então

$$
\begin{aligned}
y &\sim \text{Bernoulli}(\phi) \\
x|y = 0 &\sim \mathcal{N}(\mu_0, \Sigma) \\
x|y = 1 &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}
$$

- Escrevendo as distribuições:

$$
\begin{aligned}
y &= \phi^y (1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1).\right)
\end{aligned}
$$

- Note que temos vetores μ_0 e μ_1 mas uma única Σ.

Análise Discriminante Gaussiana (GDA)

- Parâmetros ϕ, Σ, μ_0 e μ_1 são aprendidos por máxima verossimilhança:

$$\phi = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}$$

$$\mu_0 = \frac{\sum_{i=1}^{m}(\mathbb{1}_{y^{(i)}=0})x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=0}}$$

$$\mu_1 = \frac{\sum_{i=1}^{m}(\mathbb{1}_{y^{(i)}=1})x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

Análise Discriminante Gaussiana (GDA)

- GDA, como apresentada aqui com Σ compartilhada é também conhecida por LDA (Linear Discriminant Analysis).
- Quando temos uma Σ para cada classe, o método se chama QDA (Quadratic Discriminant Analysis).
- GDA possui conexões com regressão logística.
- Se $p(x|y)$ segue uma Gaussiana multivariada (com Σ compartilhada), então $p(y|x)$ é dada por uma função logística $g(\theta^T x)$ para algum θ.
- O contrário não é verdadeiro.

Análise Discriminante Gaussiana (GDA)

- GDA, como apresentada aqui com Σ compartilhada é também conhecida por LDA (Linear Discriminant Analysis).
- Quando temos uma Σ para cada classe, o método se chama QDA (Quadratic Discriminant Analysis).
- GDA possui conexões com regressão logística.
- Se $p(x|y)$ segue uma Gaussiana multivariada (com Σ compartilhada), então $p(y|x)$ é dada por uma função logística $g(\theta^T x)$ para algum θ.
- O contrário não é verdadeiro.

Análise Discriminante Gaussiana (GDA)

- GDA, como apresentada aqui com Σ compartilhada é também conhecida por LDA (Linear Discriminant Analysis).
- Quando temos uma Σ para cada classe, o método se chama QDA (Quadratic Discriminant Analysis).
- GDA possui conexões com regressão logística.
- Se $p(x|y)$ segue uma Gaussiana multivariada (com Σ compartilhada), então $p(y|x)$ é dada por uma função logística $g(\theta^T x)$ para algum θ.
- O contrário não é verdadeiro.

Análise Discriminante Gaussiana (GDA)

- GDA, como apresentada aqui com Σ compartilhada é também conhecida por LDA (Linear Discriminant Analysis).
- Quando temos uma Σ para cada classe, o método se chama QDA (Quadratic Discriminant Analysis).
- GDA possui conexões com regressão logística.
- Se $p(x|y)$ segue uma Gaussiana multivariada (com Σ compartilhada), então $p(y|x)$ é dada por uma função logística $g(\theta^T x)$ para algum θ.
- O contrário não é verdadeiro.

Análise Discriminante Gaussiana (GDA)

- GDA, como apresentada aqui com Σ compartilhada é também conhecida por LDA (Linear Discriminant Analysis).
- Quando temos uma Σ para cada classe, o método se chama QDA (Quadratic Discriminant Analysis).
- GDA possui conexões com regressão logística.
- Se $p(x|y)$ segue uma Gaussiana multivariada (com Σ compartilhada), então $p(y|x)$ é dada por uma função logística $g(\theta^T x)$ para algum θ.
- O contrário não é verdadeiro.

Análise Discriminante Gaussiana (GDA)

- GDA assume condições mais fortes (normalidade multivariada de $p(x|y)$) do que a regressão logística.
- Quando essa condição é satisfeita, GDA tende a funcionar melhor que a regressão logística, mesmo com poucos dados de treino.

Análise Discriminante Gaussiana (GDA)

- GDA assume condições mais fortes (normalidade multivariada de $p(x|y)$) do que a regressão logística.
- Quando essa condição é satisfeita, GDA tende a funcionar melhor que a regressão logística, mesmo com poucos dados de treino.

Outline

- Assume que, dado y, os x_i's são condicionalmente independentes entre si (condição muito forte).
- Parâmetros (supondo $x_j \in \{0, 1\}$):

$$\phi_{j|y=1} = p(x_j = 1|y = 1)$$
$$\phi_{j|y=0} = p(x_j = 1|y = 0)$$
$$\phi_y = p(y = 1).$$

- Aprendidos por verossimilhança:

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=1}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}$$
$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=0}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=0}}$$
$$\phi_y = \frac{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}{m},$$

em que $\wedge$ denota "e".

- Assume que, dado y, os x_i's são condicionalmente independentes entre si (condição muito forte).
- Parâmetros (supondo $x_j \in \{0, 1\}$):

$$\phi_{j|y=1} = p(x_j = 1|y = 1)$$
$$\phi_{j|y=0} = p(x_j = 1|y = 0)$$
$$\phi_y = p(y = 1).$$

- Aprendidos por verossimilhança:

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=1}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}$$
$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=0}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=0}}$$
$$\phi_y = \frac{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}{m},$$

em que $\wedge$ denota "e".

- Assume que, dado y, os x_i's são condicionalmente independentes entre si (condição muito forte).
- Parâmetros (supondo $x_j \in \{0, 1\}$):

$$\begin{aligned} \phi_{j|y=1} &= p(x_j = 1|y = 1) \\ \phi_{j|y=0} &= p(x_j = 1|y = 0) \\ \phi_y &= p(y = 1). \end{aligned}$$

- Aprendidos por verossimilhança:

$$\begin{aligned} \phi_{j|y=1} &= \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=1}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}} \\ \phi_{j|y=0} &= \frac{\sum_{i=1}^{m} \mathbb{1}_{x_j=1 \wedge y^{(i)}=0}}{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=0}} \\ \phi_y &= \frac{\sum_{i=1}^{m} \mathbb{1}_{y^{(i)}=1}}{m}, \end{aligned}$$

em que $\wedge$ denota "e".

- Predição:

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$

$$\frac{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{n} p(x_j|y=0)\right) p(y=0)}.$$

- Se x_j não for binário, basta trocar Bernoulli por uma multinomial.

- É comum que se use Naive Bayes mesmo com atributos contínuos, discretizando-os.

- Quando os atributos originais são contínuos, mas não seguem uma normal multivariada, discretizá-los e aplicar Naive Bayes costuma funcionar melhor que GDA.

- Predição:

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$
$$\frac{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{n} p(x_j|y=0)\right) p(y=0)}.$$

- Se x_j não for binário, basta trocar Bernoulli por uma multinomial.

- É comum que se use Naive Bayes mesmo com atributos contínuos, discretizando-os.

- Quando os atributos originais são contínuos, mas não seguem uma normal multivariada, discretizá-los e aplicar Naive Bayes costuma funcionar melhor que GDA.

- Predição:

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$

$$\frac{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{n} p(x_j|y=0)\right) p(y=0)}.$$

- Se x_j não for binário, basta trocar Bernoulli por uma multinomial.

- É comum que se use Naive Bayes mesmo com atributos contínuos, discretizando-os.

- Quando os atributos originais são contínuos, mas não seguem uma normal multivariada, discretizá-los e aplicar Naive Bayes costuma funcionar melhor que GDA.

- Predição:

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$
$$\frac{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1) + \left(\prod_{j=1}^{n} p(x_j|y=0)\right) p(y=0)}.$$

- Se x_j não for binário, basta trocar Bernoulli por uma multinomial.

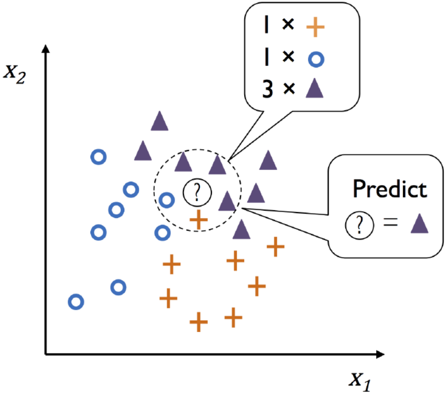
- É comum que se use Naive Bayes mesmo com atributos contínuos, discretizando-os.

- Quando os atributos originais são contínuos, mas não seguem uma normal multivariada, discretizá-los e aplicar Naive Bayes costuma funcionar melhor que GDA.

Outline

K Vizinhos Mais Próximos (KNN)

- Um dos métodos supervisionados mais simples.
- Dado x, prediz o rótulo mais frequente entre os K outros exemplos à menor distância de x.



K Vizinhos Mais Próximos (KNN)

- Distância padrão é a Euclideana (L^2):

$$d(x^{(a)}, x^{(b)}) = \sqrt{\sum_{j=1}^{n} \left(x_j^{(a)} - x_j^{(b)}\right)^2}.$$

- Pode ser usado para regressão: em vez de votação, calcula a média entre os k vizinhos.
- Altamente suscetível ao "mal da dimensionalidade".
- Processamento simples, mas alto consumo de memória.
- Considerado o modelo com a maior variância.

K Vizinhos Mais Próximos (KNN)

- Distância padrão é a Euclideana (L^2):

$$d(x^{(a)}, x^{(b)}) = \sqrt{\sum_{j=1}^{n} \left(x_j^{(a)} - x_j^{(b)}\right)^2}.$$

- Pode ser usado para regressão: em vez de votação, calcula a média entre os k vizinhos.
- Altamente suscetível ao "mal da dimensionalidade".
- Processamento simples, mas alto consumo de memória.
- Considerado o modelo com a maior variância.

K Vizinhos Mais Próximos (KNN)

- Distância padrão é a Euclideana (L^2):

$$d(x^{(a)}, x^{(b)}) = \sqrt{\sum_{j=1}^{n} \left(x_j^{(a)} - x_j^{(b)} \right)^2}.$$

- Pode ser usado para regressão: em vez de votação, calcula a média entre os k vizinhos.
- Altamente suscetível ao "mal da dimensionalidade".
- Processamento simples, mas alto consumo de memória.
- Considerado o modelo com a maior variância.

K Vizinhos Mais Próximos (KNN)

- Distância padrão é a Euclideana (L^2):

$$d(x^{(a)}, x^{(b)}) = \sqrt{\sum_{j=1}^{n} \left(x_j^{(a)} - x_j^{(b)}\right)^2}.$$

- Pode ser usado para regressão: em vez de votação, calcula a média entre os k vizinhos.
- Altamente suscetível ao "mal da dimensionalidade".
- Processamento simples, mas alto consumo de memória.
- Considerado o modelo com a maior variância.

K Vizinhos Mais Próximos (KNN)

- Distância padrão é a Euclideana (L^2):

$$d(x^{(a)}, x^{(b)}) = \sqrt{\sum_{j=1}^{n} \left(x_j^{(a)} - x_j^{(b)}\right)^2}.$$

- Pode ser usado para regressão: em vez de votação, calcula a média entre os k vizinhos.
- Altamente suscetível ao "mal da dimensionalidade".
- Processamento simples, mas alto consumo de memória.
- Considerado o modelo com a maior variância.