Aprendizado de Supervisionado

Fabrício Olivetti de França

Universidade Federal do ABC

Tópicos

1. Aprendendo com Exemplos
2. Classificação
3. Conceitos Gerais
4. Vizinho Mais Próximo

UFABC

Aprendendo com Exemplos

Aprendendo com exemplos

Uma das formas de aprendizado de máquina é definido como um conjunto de dados na forma $\{(\mathbf{x}, y)\}$, com $\mathbf{x} \in \mathbb{R}^d$ um vetor de atributos e $y \in \mathbb{R}$ uma variável alvo, queremos descobrir um mapa entre um certo $\mathbf{x}$ para seu valor y correspondente:

$$f : \mathbf{x} \to y.$$

Esse tipo de aprendizado é conhecido como Regressão.

Regressão

Exemplos de regressão:

- Predição de nota a ser atribuída para um filme.
- Estimativa do valor de um imóvel.
- Perspectiva de ganho em um investimento.
- Performance de um atleta em uma competição.

Classificação

Problema de Classificação

Similar ao problema de regressão temos um conjunto de dados na forma $\{(\mathbf{x}, y)\}$, com $\mathbf{x} \in \mathbb{R}^d$ um vetor de atributos e $y \in \mathbb{Y}$ uma variável alvo que define um conjunto finito de possíveis classificações, agora queremos descobrir uma função de probabilidade:

$$P(Y = y \mid \mathbf{x}).$$

UFABC

Exemplos de Problemas de Classificação

- Se um e-mail é spam ou não.
- Qual espécie é uma planta.
- Que tipo de doença um paciente tem.

Exemplos de Classificação

Supondo que temos a seguinte entrada de dados na tarefa de predição de se um e-mail é Spam ou não:

Símbolos	Links	Viagra	Contatos	Classe
3	1	4	0	Spam
5	3	0	0	Spam
1	1	0	2	Ham

UFABC

Exemplos de Classificação

Ao receber uma nova entrada sem classificação podemos deduzir a classe baseado no conhecimento prévio:

Símbolos	Links	Viagra	Contatos	Classe
3	1	4	0	Spam
5	3	0	0	Spam
1	1	0	2	Ham
2	1	1	1	???

Exemplos de Classificação

Como você classificaria?

Símbolos	Links	Viagra	Contatos	Classe
3	1	4	0	Spam
5	3	0	0	Spam
1	1	0	2	Ham
2	1	1	1	???

UFABC

Conceitos Gerais

Notações utilizadas

Durante o curso usaremos as seguintes notações:

- **Atributos (Features):** as variáveis mensuradas que serão utilizadas como entrada para o algoritmo.
- **Alvo (Target):** variável que queremos predizer, nosso objetivo.
- **Exemplo (Example):** uma amostra de um vetor de atributos.
- **Rótulo (Example):** uma amostra de um valor alvo, relacionada a um exemplo.

Notações utilizadas

	Atributos				Alvo
	Símbolos	**Links**	**Viagra**	**Contatos**	**Classe**
	3	1	4	0	Spam
Exemplo	5	3	0	0	Spam
	1	1	0	2	Ham
	2	1	1	1	???

UFABC

Notações utilizadas

	Atributos				Alvo
	Símbolos	**Links**	**Viagra**	**Contatos**	**Classe**
	3	1	4	0	Spam
Rótulo	5	3	0	0	Spam
	1	1	0	2	Ham
	2	1	1	1	???

UFABC

Vizinho Mais Próximo

Vizinhos mais próximos

A ideia do exemplo anterior é conhecida como *Vizinho mais próximo*.

Essa é uma técnica bem simples e que apresenta um desempenho razoável em muitas aplicações.

Vizinhos mais próximos

Digamos que queremos descobrir se um certo $\mathbf{x_i}$ pertence a classe $y_i = Spam$ ou $y_i = Ham$.

Se tivermos uma coleção de exemplos X já classificados, podemos dizer que $\mathbf{x_i}$ pertence a mesma classe do $\mathbf{x}' \in X$ mais similar a ele.

Vizinhos mais próximos

Para determinar qual é o exemplo mais próximo de nosso $\mathbf{x_i}$, precisamos definir medidas de distância (minimizar) ou similaridade (maximizar).

Medindo Similaridade

Uma tarefa importante é distinguir o quão similar são dois objetos:

- Se similares, são do mesmo tipo/classe
- Se similares, são do mesmo grupo
- A distância é proporcional ao aumento de um valor-objetivo (i.e., nota)

Medindo Distâncias

Por outro lado, ao medirmos a distância o inverso é válido.

Uma medida de distância $d : X \times X \to [0, \text{inf})$ é considerada uma métrica se:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \iff x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

UFABC

Distância de Minkowski

Métrica de distância entre dois valores dada uma ordem p:

$$D_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Essa distância é influenciada pela direção e intensidade do vetor.

UFABC

Distância de Minkowski

```
minkowski :: Int -> Vetor Double -> Vetor Double
          -> Double

def minkowski(p, x, y):
    difP      = np.power(np.abs(x-y), p)
    somatoria = np.sum(difAbs)
    return np.power(somatoria, 1.0/p)
```

UFABC

Distância Euclidiana

Distância de Minkowski com $p = 2$:

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{\frac{1}{2}}$$

UFABC

Distância de Manhattan

Distância de Minkowski com $p = 1$:

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

UFABC

Distância de Chebyshev

Distância de Minkowski com $p \to \inf$:

$$D(x, y) = \max_{i=1}^{n} |x_i - y_i|$$

UFABC

Distância de Minkowski com $p = 1, 2, \text{inf}$

```
manhattan  = lambda x, y: minkowski(1,x,y)
euclidiana = lambda x, y: minkowski(2,x,y)
chebyshev  = lambda x, y: np.max(np.abs(x-y))
```

UFABC

Similaridade dos Cossenos

Medida de similaridade (não é métrica) entre dois vetores que indica se eles apontam para a mesma direção:

$$S(x, y) = \frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2}$$

UFABC

Similaridade dos Cossenos

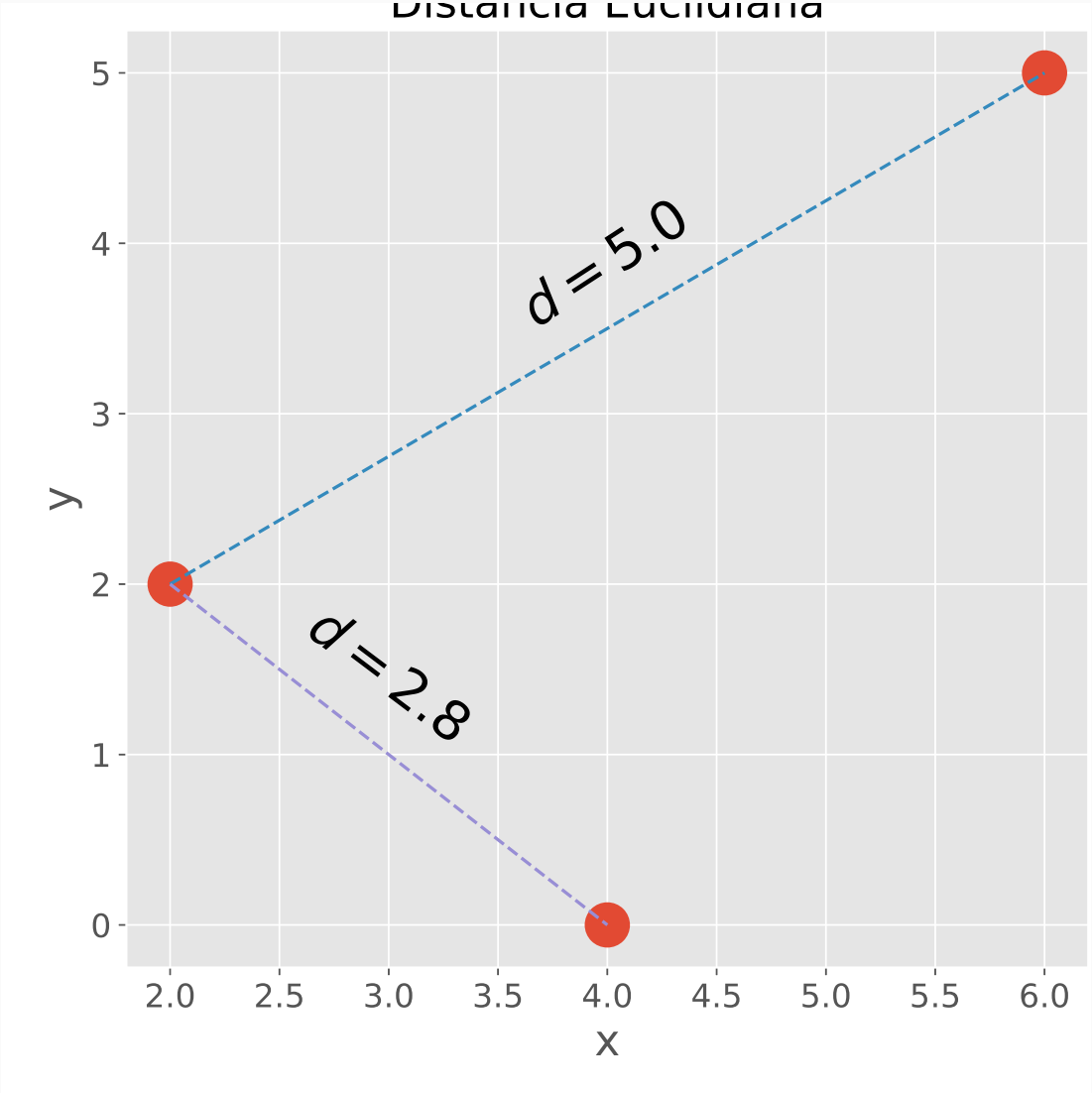
```
cosine :: Vector Double -> Vector Double
       -> Double

def cosine(x, y):
    return np.dot(x,y)/(norma(x) * norma(y))

def norma(x):
    return np.sqrt(np.dot(x,x))
```

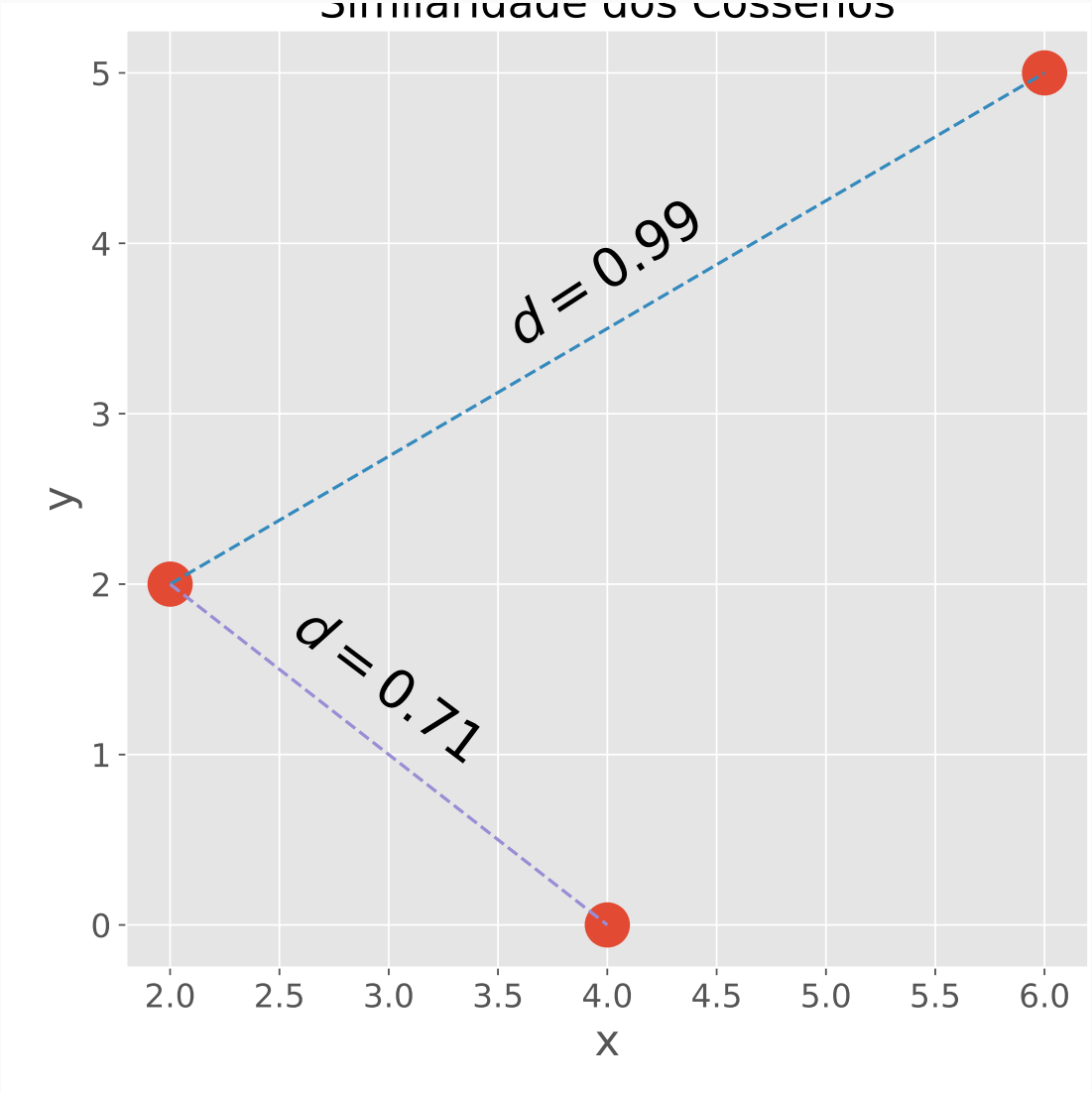
UFABC

Euclidiana vs Cosseno



UFABC

Euclidiana vs Cosseno



UFABC

Euclidiana vs Cosseno

Euclidiana

Quando a intensidade das variáveis possuem um papel principal:

- Uma pessoa com $40kg$ e $100cm$ não é parecida com uma pessoa com $80kg$ e $200cm$ se comparamos necessidade de consumo de energia.
- Um imóvel com $40m^2$ e 1 quarto não vale o mesmo que um de $120m^2$ e 3 quartos.

Cosseno

Quando nos interessamos pela proporcionalidade ou direção:

- Frequência de palavras: a proporção dos uso de palavras é mais importante que a intensidade.
- Uma pessoa com $40kg$ e $100cm$ é parecida com uma pessoa com $80kg$ e $200cm$ se estamos verificando o índice de massa corporal.

UFABC

Vizinhos mais próximos

O nosso algoritmo fica:

```
from scipy.spatial.distance import cdist

def vizinhoMaisProximo(xi, x, y):
    '''
    xi: exemplo a ser rotulado
    x : matriz de exemplos
    y : rótulos dos exemplos
    '''
    idx = np.argmin(cdist([xi], x)[0])
    return y[idx]
```

UFABC

Vizinhos mais próximos

Esse processo pode ser falho pois:

- O mais similar pode ter sido classificado incorretamente.
- A amostra está na fronteira entre c1 e c2 tendo chance do mais próximos ser da classe errada.

k-Vizinhos mais próximos

Para resolver, escolhemos os k pontos mais similares e calculamos a moda das classes desse ponto.

Dessa forma reduzimos a possibilidade de erros.

k-Vizinhos mais próximos

```
from scipy.spatial.distance import cdist

def kNN(xi, x, y):
    '''
    xi: exemplo a ser rotulado
    x : matriz de exemplos
    y : rótulos dos exemplos
    '''
    idx = np.argsort(cdist([xi], x)[0])[:k]
    return np.mode(y[idx])
```

UFABC

Próxima Aula

Na próxima aula aprenderemos sobre:

- Tipos de Variáveis
- Padronização e Normalização
- Extraindo variáveis de texto

UFABC

Atividade 01

Leia o material da Intel em: Introduction to Machine Learning and Toolkit e complete a atividade Introduction_to_Machine_Learning_and_Toolkit.ipynb.

Essa é uma atividade que demandará ao aluno pesquisar, estudar e aprender a ler documentações das bibliotecas:

- Pandas
- Scikit-Learn
- Numpy
- Matplotlib
- Seaborn