Universidade de São Paulo (USP)

Faculdade de Economia, Administração, Contabilidade e Atuária (FEA)

IA e ML Aplicados a Finanças

Prof. Leandro Maciel

AULA 8: Classificação e K-NN

Agenda

1. K-NN

- Regressão logística - **vantagens**:

  Eficiente, baixo custo computacional, interpretável;

  Não requer normalização de variáveis; saídas são probabilidades;

- **Desvantagens**:

  Hipóteses distribuição dados, baixo desempenho problemas não lineares;

  Exige não omissão variáveis, suscetível a **overfitting**.

1. K-NN

- São várias técnicas de **ML** para classificação;
- **Nearest Neighbor** ("vizinho mais próximo"):

  Considera o grau de similaridade aos elementos de uma classe.

- Classificar ave: considera *features* das demais aves de diferentes classes;
- Apesar da ideia simples, é uma técnica muito poderosa;
- Reconhecimento imagens, identificação padrões, etc...

1. K-NN

- Muito útil quando há inúmeras características em uma mesma classe;
- Relações de difícil entendimento, mas homogeneidade intra classe;
- Método simples, rápido e sem hipóteses acerca dos dados. Porém...
- Não é interpretável (*black box*), requer definição de parâmetros;
- Dados nominais e *missing values* requerem pré-processamento...

1. K-NN

- Método conhecido como **k-nearest neighbor algorithm** (k-NN);
- Classifica um item com base nos k (*user defined*) vizinhos mais próximos;
- Etapas na classificação com k-NN:
  1. Treinamento com um conjunto de dados rotulados (*supervised*);
  2. Localizar k vizinhos mais próximos a um objeto sem classificação;
  3. Objeto é associado ao mesmo grupo da maioria dos k vizinhos.

- **Exemplo** $\rightarrow$ suponha que queremos classificar um alimento (tomate);

Classes: proteínas, vegetais e frutas;

Features: doçura e crocância (ambas enumeradas de 1 a 10);

Suponha que tenhamos o seguinte conjunto treinamento:

Alimento	**Doçura**	**Crocância**	**Classe**
Maça	10	9	Fruta
Laranja	7	3	Fruta
Noz	3	6	Proteína
Cenoura	7	10	Vegetal
Uva	8	5	Fruta
Queijo	1	1	Proteína
Ervilha	3	7	Vegetal

1. K-NN

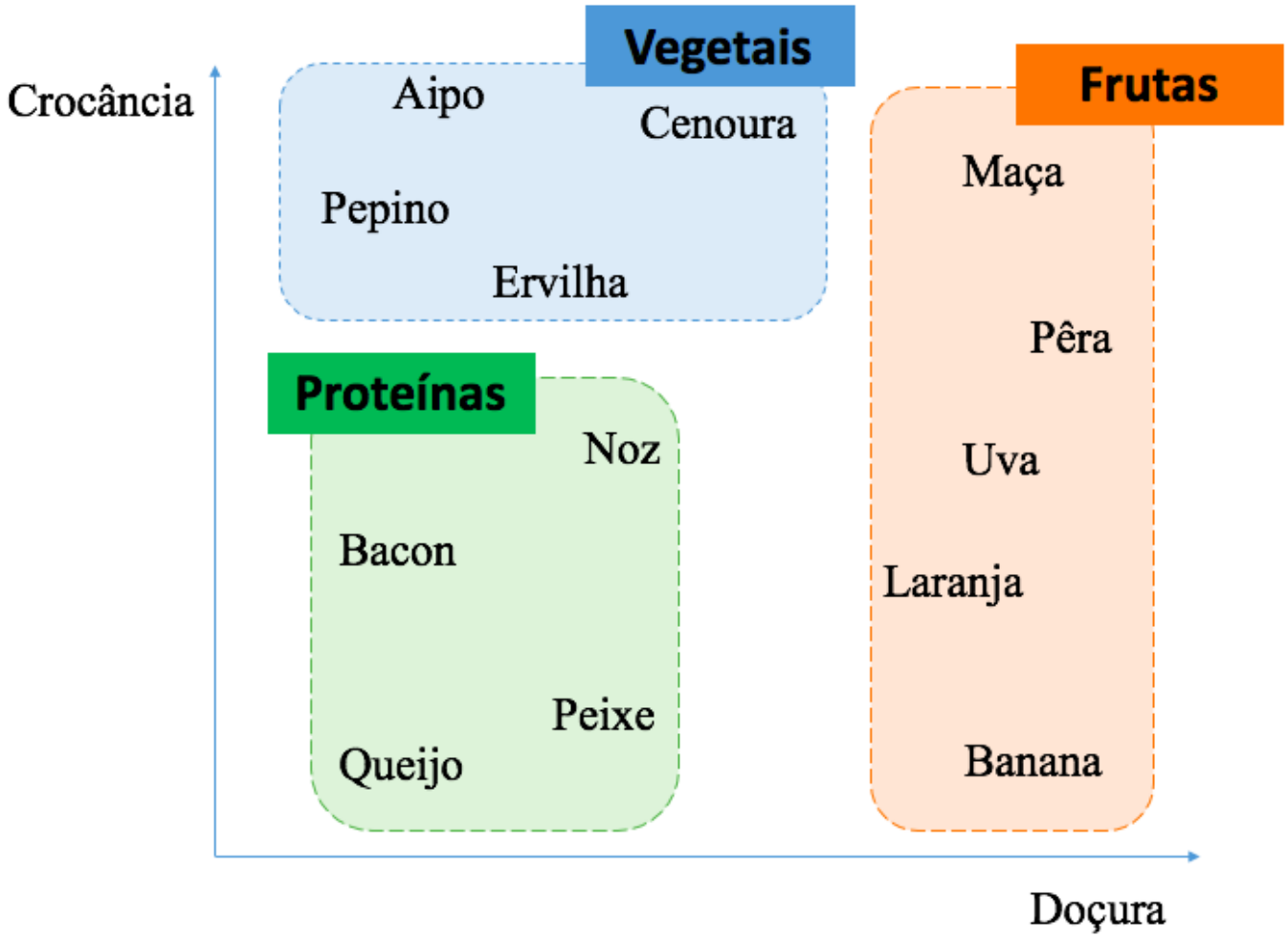
- Algoritmo k-NN trata as características como coordenadas;
- Coordenadas em um espaço multidimensional (*feature space*);
- Podemos ter n características (*features*) $\rightarrow \Re^n$:

$$(x_1, x_2, x_3, \ldots, x_n) \in \Re^n$$

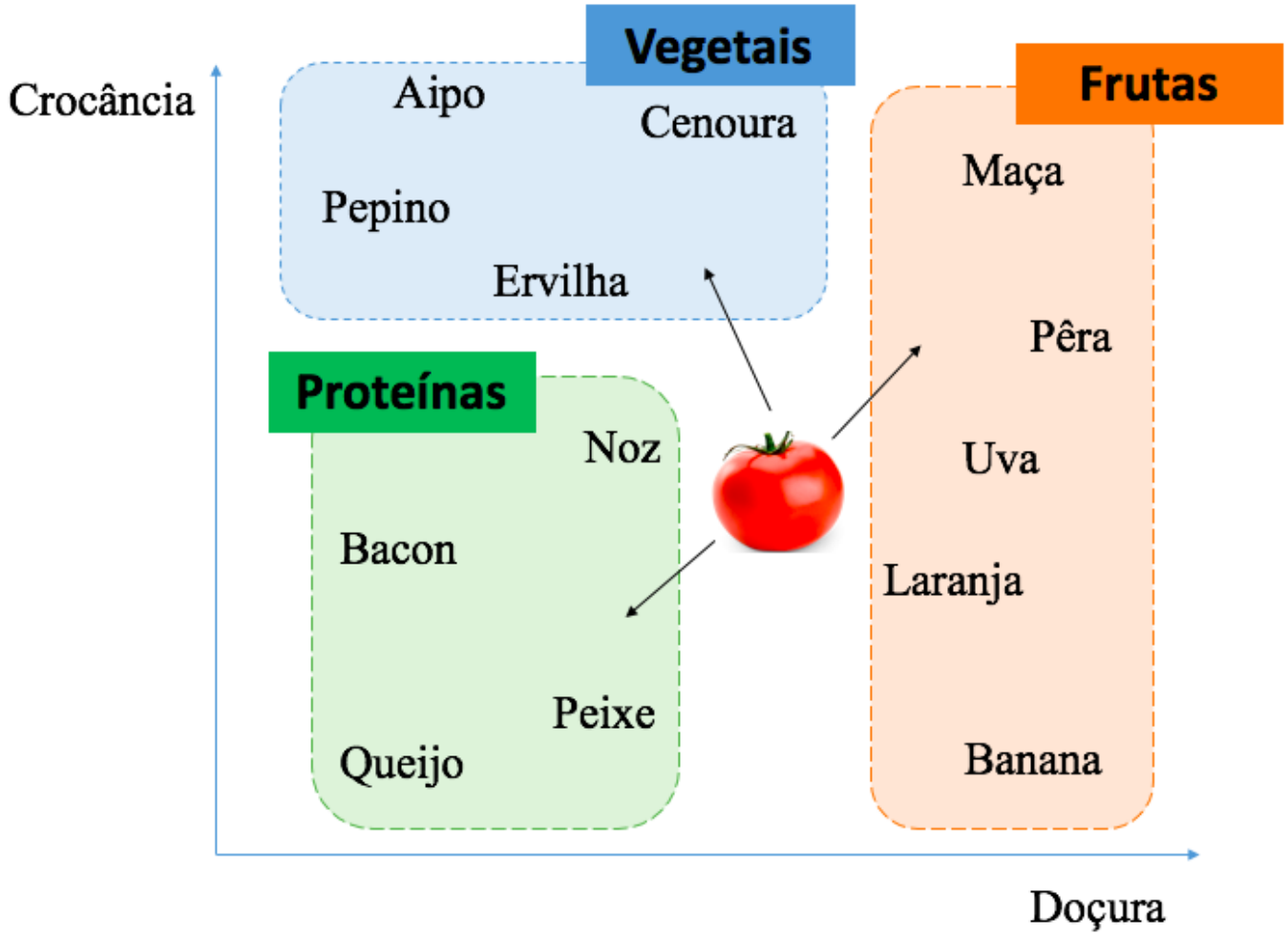
- Hipótese de que há um padrão dentre elementos de uma mesma classe;
- Mapeamento entrada-saída:

$$y = f(x_1, x_2, x_3, \ldots, x_n)$$

- Feature space da amostra treinamento:



- Como classificar um novo objeto (tomate - doçura 6 e crocância 4):



1. K-NN

- Precisamos localizar os k vizinhos mais próximos (similares);
- k-NN $\rightarrow$ mensura **similaridade** no espaço de características;
- Distância Euclidiana como medida de similaridade:

$$dist(a, b) = ||a - b|| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

- $1, 2, \ldots, n$ são as *features* dos objetos comparados;
- Objeto associado a classe da maioria dos k vizinhos mais próximos.

- Distância entre tomate e ervilha:

$$dist(\text{tomate}, \text{ervilha}) = \sqrt{(6-3)^2 + (4-7)^2} = 4,2$$

- Para outros vizinhos, por exemplo:

Alimento	**Doçura**	**Crocância**	**Classe**	**Distância**
Uva	8	5	Fruta	$\sqrt{(6-8)^2 + (4-5)^2} = 2,2$
Queijo	1	1	Proteína	$\sqrt{(6-1)^2 + (4-1)^2} = 5,8$
Noz	3	6	Proteína	$\sqrt{(6-3)^2 + (4-6)^2} = 3,6$
Laranja	7	3	Fruta	$\sqrt{(6-7)^2 + (4-3)^2} = 1,4$
Cenoura	7	10	Vegetal	$\sqrt{(6-7)^2 + (4-10)^2} = 6,1$
Maça	10	9	Fruta	$\sqrt{(6-10)^2 + (4-9)^2} = 6,4$

- Qual a classificação para $k = 1, 2, 3, 4$?

- **Como determinar o valor de k?**
- Impacta na capacidade de generalização do modelo;
- Maior $k \rightarrow$ menor impacto de ruídos (dados classificados erroneamente);
- Pode ignorar importantes padrões associados aos mais próximos vizinhos;
- Depende do número de dados, N, na amostra de treinamento;
- $k \approx \sqrt{N}$, mas pode ser escolhido por otimização (melhor resultado).

- Características (**features**) podem ser representadas em diferentes escalas;
- No exemplo, ambas medidas de 1 a 10;
- Diferenças de escala afeta a medida de distância;
- Para evitar esse problema, os dados devem ser **normalizados**.

- Normalização **min-max**:

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)}, \quad x_{norm} \in [0, 1]$$

- Normalização **z-score**:

$$x_{norm} = \frac{x - \mu_x}{\sigma_x}$$

- Características categóricas (dicotômicas), usamos variáveis binárias.

- Pseudo-código para algoritmo k-NN:

Algoritmo k-NN

1. entrada (X, Y, x), X treinamento, Y classes de X, x teste (tamanho m)
2. **for** $i = 1$ **to** m **do**
3. Calcule distância $d(X_i, x)$
4. **end for**
5. Selecione conjunto I com os índices das k menores distâncias $d(X_i, x)$
6. **return** classe majoritária para $\{Y_j,\ j \in I\}$

- Método implementado em R no pacote "**class**".

1. K-NN

- Exemplo de classificação de *default* de cartão crédito...
- Features $X \rightarrow$ características dos agentes;
- Classes $Y \rightarrow$ dafault (1) e não default (0);
- Variáveis devem estar normalizadas e especificar k;
- Comparação com regressão logística...

2. Árvores de Decisão

- Limitações K-NN:
  - sensibilidade *features* irrelevantes;
  - computacionalmente custoso ("*lazy learner*");
  - acurácia depende da qualidade dos dados;
  - performance prejudicada em altas dimensões (*features*);
  - sensível *outliers* e valores faltantes.
- **Modelos de Árvores de Decisão**.

2. Árvores de decisão

- Árvores de decisão:
  - problemas de regressão e classificação;
  - conjunto de regras de decisão;
  - extratificação resulta em regiões (**nós terminais ou folhas**);
  - saída é média dos indivíduos da região terminal (reg.);
  - saída é classe de maior freq. dos ind. da região terminal (clas.);
  - nós interpretáveis (relevância dos atributos);
  - nós mais internos $\rightarrow$ maior relevância.

Árvore de regressão para prever o salário em log de um jogador de beisebol

- Atributos: no número de anos que ele jogou nas grandes ligas (*years*) e número de rebatidas que ele fez no ano anterior (*hits*).
- Nó interno: ramo da esquerda corresponde a Anos < 4.5, e o ramo da direita corresponde a Anos >= 4.5.
- A árvore tem dois nós internos e três nós terminais, ou folhas.
- O número em cada folha é a média da resposta para as observações que caem ali. Classe mais frequente, se problema de classificação.

2. Árvores de decisão

- Treinamento (construção) das árvores de decisão:
  - crecer a árvore;
  - adicionar mecanismos de poda;
  - definir medida de erro (ajuste);
  - otimizar medida por validação cruzada.
- Principal problema: elevada variância;
- Alternativas: **bagging** e **random forests**;
- Pacotes no R: "party" (árvores de decisão) e "randomForest".

- Próxima aula...
  - "treinamento" $\rightarrow$ melhor modelo;
  - processo de otimização de parâmetros...

3. Bibliografia

HASTIE, Trevor, TIBSHIRANI, R., & FRIEDMAN, Jerome. **The Elements of Statistical Learning**. Data Mining, Inference, and Prediction. 2 Ed. Springer, 2008. Capítulo 13.

JAMES, Gareth, et al. **An Introduction to Statistical Learning** - With Applications in R. New York: Springer, 2013. Capítulos 4 e 8.

ADLER, Joseph. **R in a Nutshell** - A Desktop Quick Reference. 2 Ed. Sebastopol, CA: O'Reilly, 2012. Capítulo 21.

Prof. Leandro Maciel

leandromaciel@usp.br