FUNDAMENTOS DE MACHINE LEARNING: Classificação e Agrupamento (clustering)

Fevereiro de 2021

Ref.: Cap. 5 do livro texto

Machine Learning: uma visão panorâmica de métodos, algoritmos e modelos

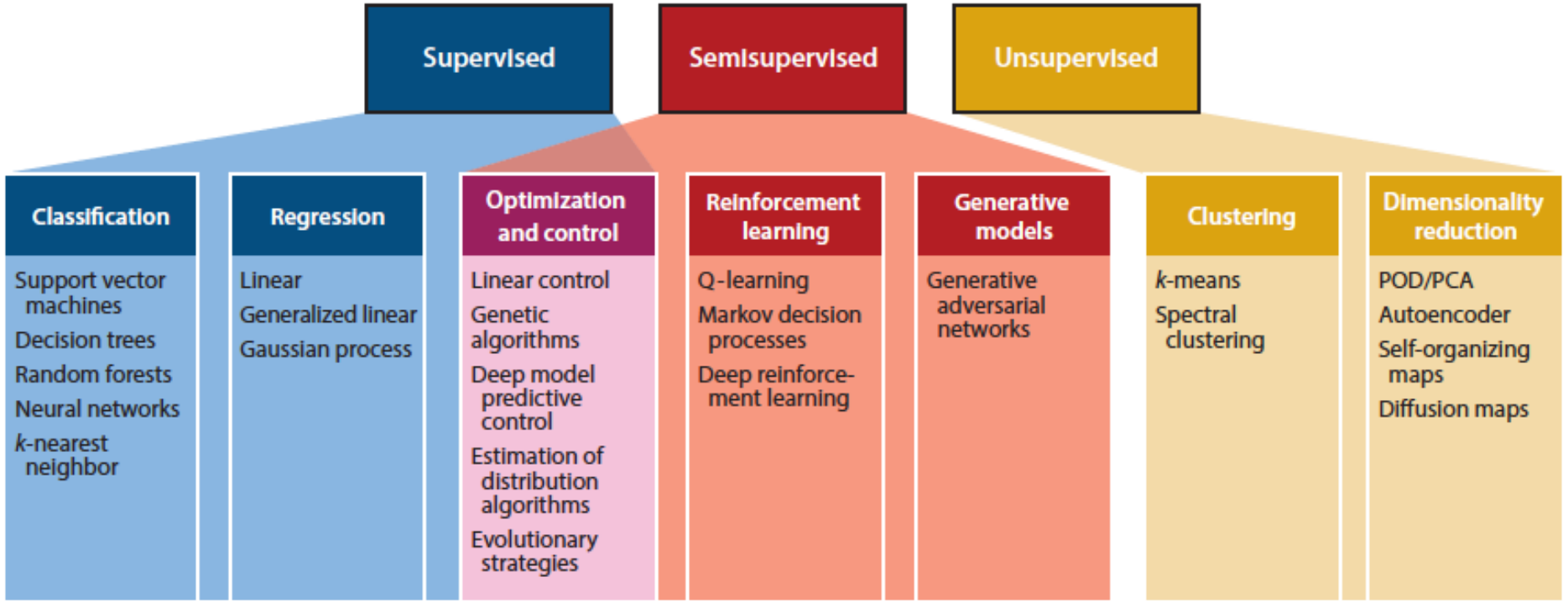


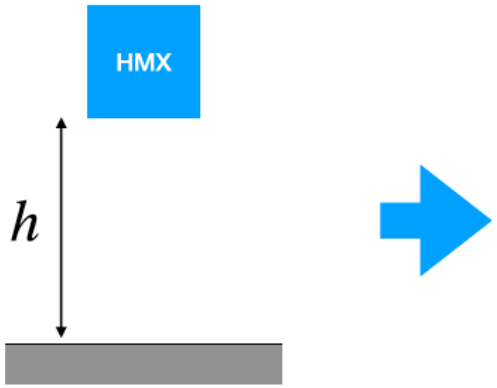
Figure 1

Machine learning algorithms may be categorized into supervised, unsupervised, and semisupervised, depending on the extent and type of information available for the learning process. Abbreviations: PCA, principal component analysis; POD, proper orthogonal decomposition.

Annual Review of Fluid Mechanics Machine Learning for Fluid Mechanics
Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos

Classificadores...

Example: Sensitivity of energetic materials



Does it explode or not when you drop it from height h?

Height (cm)	Results
40.5	E E E E E E E E E E
36.0	E N E E E E N E E E
32.0	E E N E E E N E N E
28.5	N E N N E N N N E N
25.5	N N N N N N E N N N
22.5	N N N N N N N N N N

Data from L. Smith, "Los Alamos National Laboratory explosives orientation course: Sensitivity and sensitivity tests to impact, friction, spark and shock," Los Alamos National Lab, NM (USA), Tech. Rep., 1987

PREDICTIVE SCIENCE LABORATORY

Classificação : aprendizado supervisionado e otimização

- Temos n valores (eventualmente x é um vetor d-dimensional) de entrada (variáveis de controle)

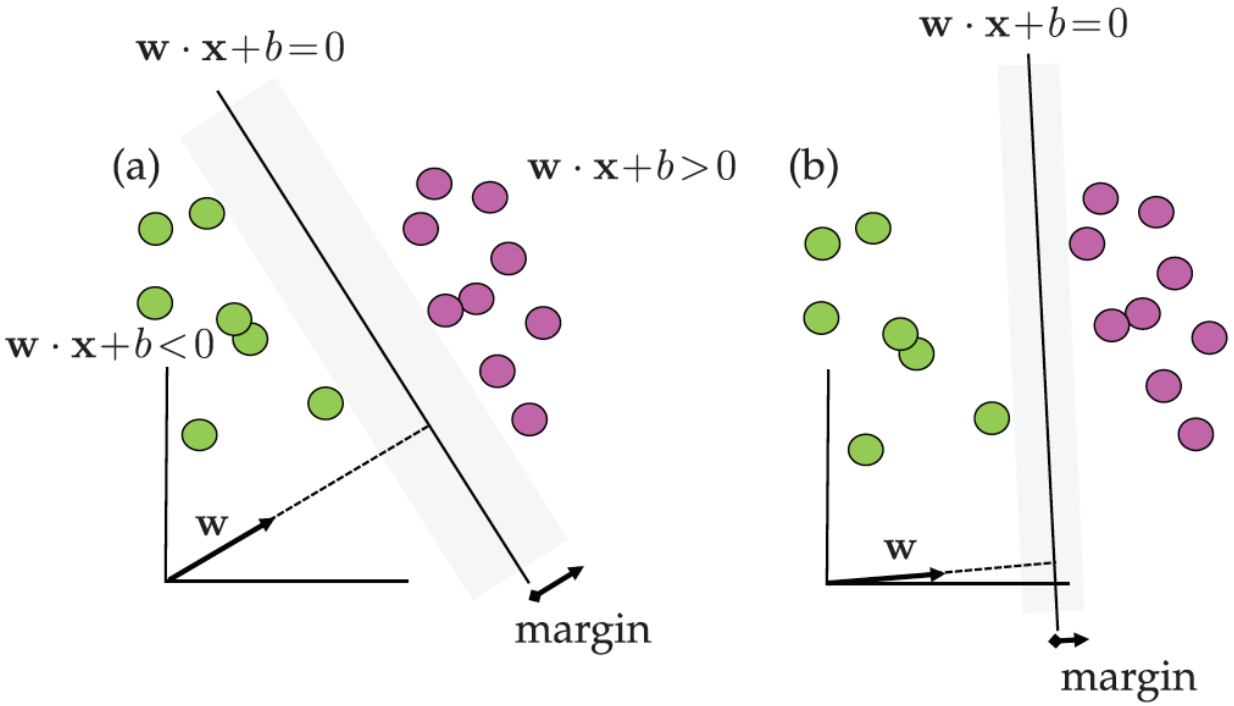
$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$

- e n valores de saída (eventualmente y é um vetor d'-dimensional)

$$y_{1:n} = (y_1, \ldots, y_n)$$

- Obs: no contexto de classificação, y não é uma variável assumindo valores contínuos (lembrem-se do exemplo dos explosivos)

Support Vector Machines (SVM)



$$\mathbf{y}_j(\mathbf{w}\cdot\mathbf{x}_j+b)=\text{sign}(\mathbf{w}\cdot\mathbf{x}_j+b)=\begin{cases}+1 & \text{magenta ball}\\ -1 & \text{green ball.}\end{cases}$$

Figure 5.22 The SVM classification scheme constructs a hyperplane $\mathbf{w}\cdot\mathbf{x}+b=0$ that optimally separates the labeled data. The area of the margin separating the labeled data is maximal in (a) and much less in (b). Determining the vector $\mathbf{w}$ and parameter b is the goal of the SVM optimization. Note that for data to the right of the hyperplane $\mathbf{w}\cdot\mathbf{x}+b>0$, while for data to the left $\mathbf{w}\cdot\mathbf{x}+b<0$. Thus the classification labels $\mathbf{y}_j\in\{\pm1\}$ for the data to the left or right of the hyperplane is given by $\mathbf{y}_j(\mathbf{w}\cdot\mathbf{x}_j+b)=\text{sign}(\mathbf{w}\cdot\mathbf{x}_j+b)$. So only the sign of $\mathbf{w}\cdot\mathbf{x}+b$ needs to be determined in order to label the data. The vectors touching the edge of the gray regions of are termed the *support vectors*.

SVM: como encontrar o hiperplano que separa os dados

Formular como um problema de otimização em que busca-se maior acerto (classificação correta dos dados, lembre-se trata-se de aprendizado supervisionado) e maximize a margem. Introduz-se como função de perda

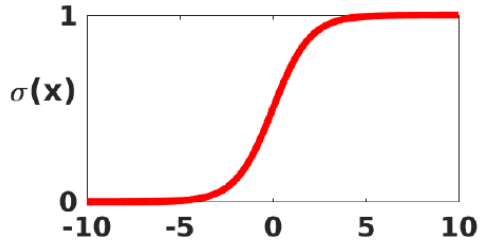
$$\ell(\mathbf{y}_j, \bar{\mathbf{y}}_j) = \ell(\mathbf{y}_j, \text{sign}(\mathbf{w} \cdot \mathbf{x}_j + b)) = \begin{cases} 0 & \text{if } \mathbf{y}_j = \text{sign}(\mathbf{w} \cdot \mathbf{x}_j + b) \\ +1 & \text{if } \mathbf{y}_j \neq \text{sign}(\mathbf{w} \cdot \mathbf{x}_j + b) \end{cases}$$

$$\ell(\mathbf{y}_j, \bar{\mathbf{y}}_j) = \begin{cases} 0 & \text{if data is correctly labeled} \\ +1 & \text{if data is incorrectly labeled} \end{cases}$$

$$\underset{\mathbf{w},b}{\text{argmin}} \sum_{j=1}^{m} \ell(\mathbf{y}_j, \bar{\mathbf{y}}_j) + \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad \min_j |\mathbf{x}_j \cdot \mathbf{w}| = 1.$$

Logistic Regression : classificação binária

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \text{sigm}\left(\sum_{j=1}^{m} w_j \phi_j(\mathbf{x})\right) = \text{sigm}$$



$$\text{sigm}(z) = \frac{1}{1 + e^{z}}$$

$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))$$

$$p(y = 0|\mathbf{x}, \mathbf{w}) = 1 - p(y = 1|\mathbf{x}, \mathbf{w}) = 1 - \text{sigm}\left(\mathbf{w}^T \phi(\mathbf{x})\right)$$

Regressão logística

$$p(y|\mathbf{x},\mathbf{w}) = \left[\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x})\right)\right]^y \left[1-\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x})\right)\right]^{1-y}$$

$$p(y_{1:n}|\mathbf{x}_{1:n},\mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i,\mathbf{w}) = \prod_{i=1}^{n}\left[\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x}_i)\right)\right]^{y_i}\left[1-\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x}_i)\right)\right]^{1-y_i}$$

$$p(\mathbf{w}|\mathbf{x}_{1:n},y_{1:n}) \propto p(y_{1:n}|\mathbf{x}_{1:n},\mathbf{w})p(\mathbf{w})$$

$$\log p(\mathbf{w}|\mathbf{x}_{1:n},y_{1:n}) =$$

$$\sum_{i=1}^{n}\left\{y_i\,\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x}_i)\right)+(1-y_i)\left[1-\text{sigm}\left(\mathbf{w}^T\phi(\mathbf{x}_i)\right)\right]\right\}+\log p(\mathbf{w})+\text{const.}$$

Frequentemente ... Usa-se uma estimativa pontual. w*

Tomando decisões a partir da regressão logística (dentro de uma visão probabilística)

$$p(y|\mathbf{x}, \mathbf{w} = \mathbf{w}^*)$$

$l(\hat{y}, y)$ Função custo expressando (nas circunstâncias da aplicação) o ônus da escolha

$$\min_{\hat{y}} \sum_{y=0,1} l(\hat{y}, y) p(y|\mathbf{x}, \mathbf{w} = \mathbf{w}^*)$$

Aprendizado não supervisionado

$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$

- Agrupamento (clustering) : promover separação entre os dados em diferentes classes de interesse
- Redução de dimensão (ex. PCA) : buscar um conjunto mínimo que “caracterize” os dados
- Estimação de densidade (density estimation) : extrair dos dados sua densidade de probabilidade, o que permite gerar “novos dados”.

k – means clustering

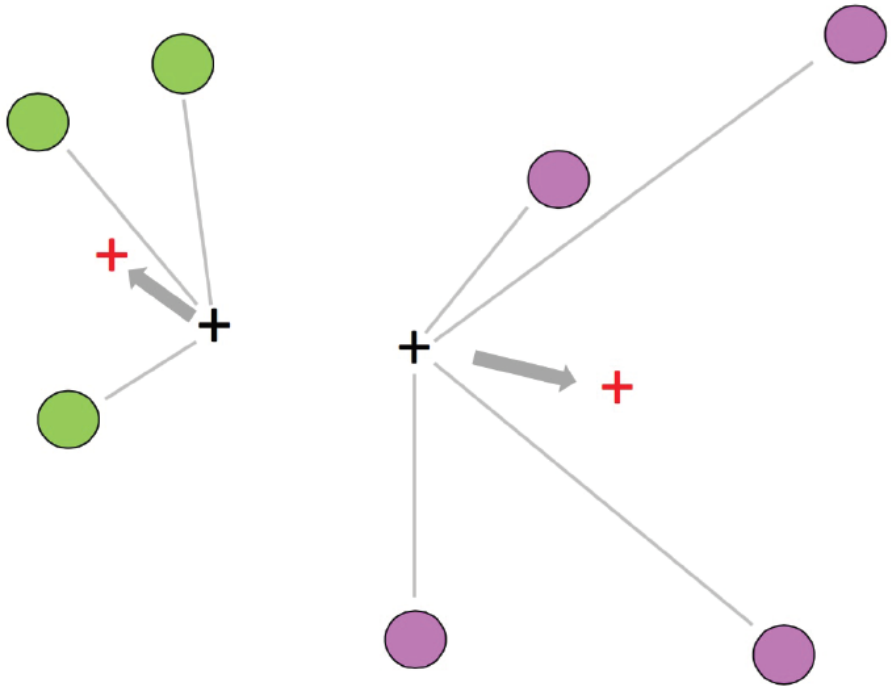
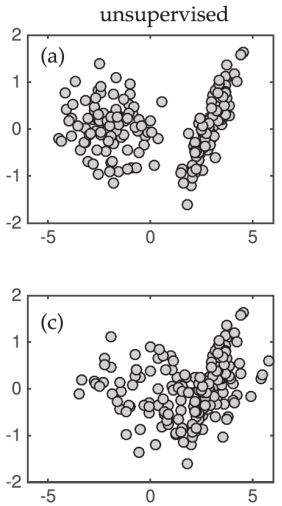


Figure 5.9 Illustration of the k-means algorithm for $k = 2$. Two initial starting values of the man are given (black +). Each point is labeled as belonging to one of the two means. The green balls are thus labeled as part of the cluster with the left + and the magenta balls are labeled as part of the right +. Once labeled, the mean of the two clusters is recomputed (red +). The process is repeated until the means converge.

$$\underset{\boldsymbol{\mu}_j}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{\mathbf{x}_j \in \mathcal{D}'_j} \|\mathbf{x}_j - \boldsymbol{\mu}_j\|^2$$

NP - hard



k-means : uma abordagem heurística

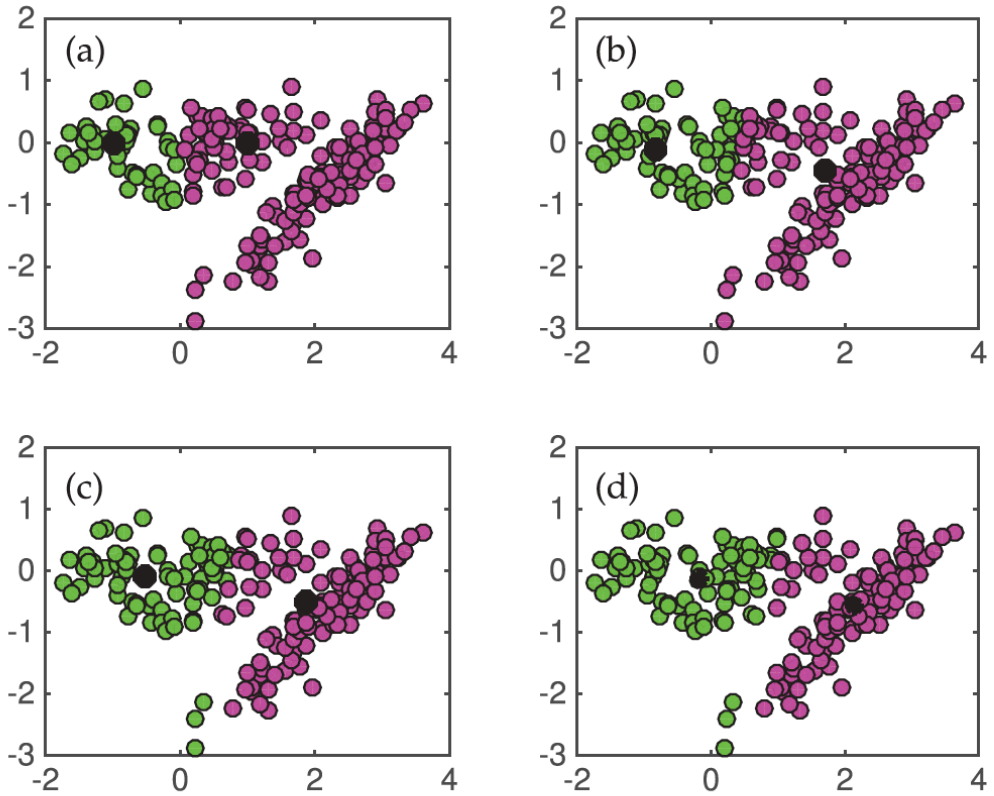


Figure 5.10 Illustration of the k-means iteration procedure based upon Lloyd's algorithm [339]. Two clusters are sought so that $k = 2$. The initial guesses (black circles in panel (a)) are used to initially label all the data according to their distance from each initial guess for the mean. The means are then updated by computing the means of the newly labeled data. This two-stage heuristic converges after approximately four iterations.

Code 5.6 Lloyd algorithm for k-means.

```
g1=[-1 0]; g2=[1 0]; %  Initial guess
for j=1:4
    class1=[]; class2=[];
    for jj=1:length(Y)
        d1=norm(g1-Y(jj,:));
        d2=norm(g2-Y(jj,:));
        if d1<d2
          class1=[class1; [Y(jj,1) Y(jj,2)]];
        else
          class2=[class2; [Y(jj,1) Y(jj,2)]];
        end
    end
    g1=[mean(class1(1:end,1)) mean(class1(1:end,2))];
    g2=[mean(class2(1:end,1)) mean(class2(1:end,2))];
end
```

Density estimation – Gaussian Mixtures

$$p(x) = \sum_{i=1}^{k} \pi_k \, \mathcal{N}(x|\mu_k, \Sigma_k)$$

Encontrar

$$\boldsymbol{\theta} = \{\pi_k, \mu_k, \Sigma_k\}$$

que maximizam (MLE)

$$\mathcal{L} = \sum_{j=1}^{n} log(p(x_n|\pi_k, \mu_k, \Sigma_k))$$

condição de ótimo:

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0$$