



UNIVERSIDADE
FEDERAL DO CEARÁ



Aprendizagem de Máquina

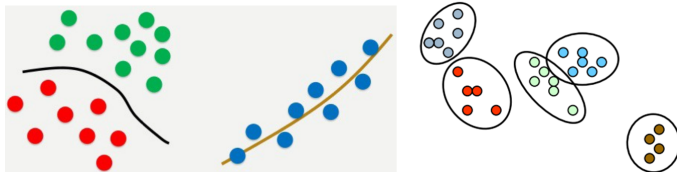
César Lincoln Cavalcante Mattos

2020

Agenda

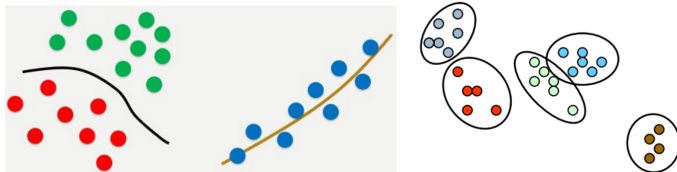
- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

Aprendizagem não-supervisionada



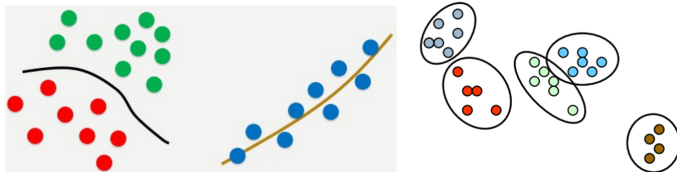
- Não há uma relação entre entradas e saídas observadas.
- **Descobrimento de estruturas/padrões** nos dados disponíveis.

Aprendizagem não-supervisionada



- Não há uma relação entre entradas e saídas observadas.
- **Descobrimento de estruturas/padrões** nos dados disponíveis.
- **Exemplos de tarefas não-supervisionadas:**
 - Agrupamento (*clustering*) de dados.
 - Redução de dimensionalidade.
 - Modelar a densidade de probabilidade dos dados.
 - Obter causas ocultas (não disponíveis) para os dados.

Aprendizagem não-supervisionada



- Não há uma relação entre entradas e saídas observadas.
- **Descobrimento de estruturas/padrões** nos dados disponíveis.
- **Exemplos de tarefas não-supervisionadas:**
 - Agrupamento (*clustering*) de dados.
 - Redução de dimensionalidade.
 - Modelar a densidade de probabilidade dos dados.
 - Obter causas ocultas (não disponíveis) para os dados.
- **Exemplos de aplicações:**
 - Compressão de dados.
 - Detecção de dados discrepantes (*outliers*).
 - Auxiliar outros algoritmos de aprendizagem.

Aprendizagem não-supervisionada

- **Principais abordagens:**

- **Agrupamento (clustering):** Cada padrão de entrada é representado por um protótipo (**algoritmo k-médias**, modelos de misturas...)
- **Redução de dimensionalidade:** Representa os padrões de entrada com uma menor quantidade de atributos (**PCA**, factor analysis, ICA...)
- **Modelar densidades de probabilidade:** Estimar distribuições de probabilidade no espaço de dados disponíveis.

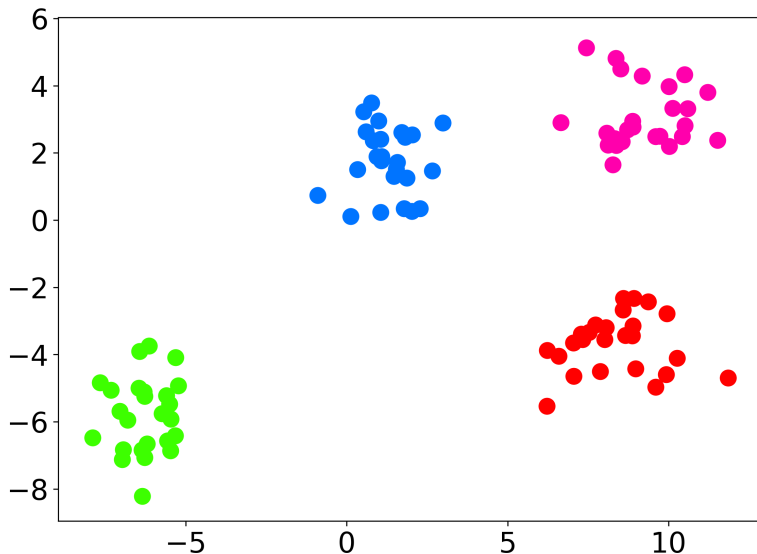
Agenda

- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

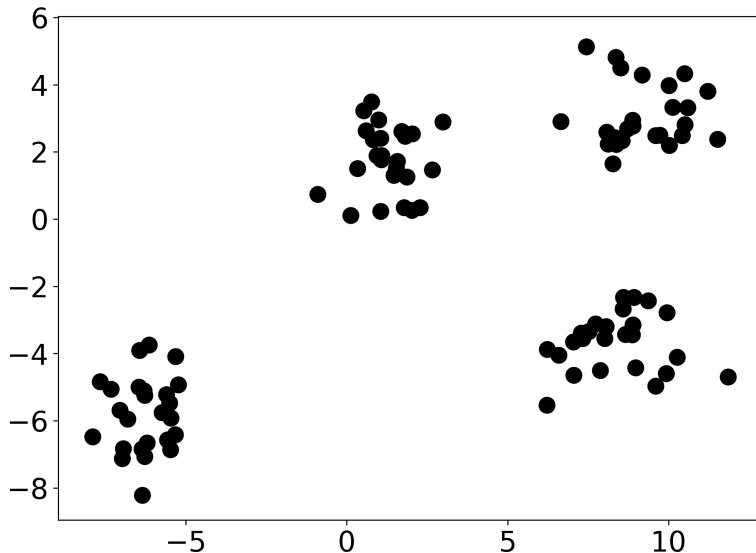
Agrupamento (*clustering*) de dados

- **Problema:** Como agrupar N exemplos em K grupos (*clusters*)?
- **Motivação:** Realizar predições, compressão com perdas, detecção de *outliers*.
- **Suposição:** Os dados foram gerados por K fontes/classes diferentes.

Agrupamento (*clustering*) de dados



Agrupamento (*clustering*) de dados



Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.
 - **Divisivo** (*top-down*): Particionamento recursivo, de grupos maiores para menores.

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.
 - **Divisivo** (*top-down*): Particionamento recursivo, de grupos maiores para menores.
 - Constrói um **dendrograma** (árvore de grupos) com o agrupamento dos padrões.

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.
 - **Divisivo** (*top-down*): Particionamento recursivo, de grupos maiores para menores.
 - Constrói um **dendrograma** (árvore de grupos) com o agrupamento dos padrões.
- **Agrupamento não-hierárquico:**
 - **Particionamento rígido:** Cada ponto pertence a um grupo.

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.
 - **Divisivo** (*top-down*): Particionamento recursivo, de grupos maiores para menores.
 - Constrói um **dendrograma** (árvore de grupos) com o agrupamento dos padrões.
- **Agrupamento não-hierárquico:**
 - **Particionamento rígido:** Cada ponto pertence a um grupo.
 - **Particionamento suave:** Cada ponto pode pertencer parcialmente a múltiplos grupos.

Agrupamento (*clustering*) de dados

- **Problemas:** Quantos grupos existem? Quais padrões pertencem a cada grupo? Como escolher um agrupamento adequado?
- **Agrupamento hierárquico:**
 - **Aglomerativo** (*bottom-up*): Grupos semelhantes são reunidos em grupos de mais alta hierarquia.
 - **Divisivo** (*top-down*): Particionamento recursivo, de grupos maiores para menores.
 - Constrói um **dendrograma** (árvore de grupos) com o agrupamento dos padrões.
- **Agrupamento não-hierárquico:**
 - **Particionamento rígido:** Cada ponto pertence a um grupo.
 - **Particionamento suave:** Cada ponto pode pertencer parcialmente a múltiplos grupos.
 - Realocação de padrões em grupos específicos de acordo com algum critério.

Critérios de dissimilaridade

- **Distância Euclidiana:**

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2 = \sqrt{\sum_{d=1}^D (x_{id} - x_{jd})^2}.$$

- **Distância de Manhattan:**

$$\|\mathbf{x}_i - \mathbf{x}_j\|_1 = \sum_{d=1}^D |x_{id} - x_{jd}|.$$

- **Distância de Mahalanobis:**

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)},$$

em que Σ é matriz de covariância dos dados de treinamento.

Exemplo de agrupamento aglomerativo

- **Dados:** $1 : [1 \ 2], 2 : [1 \ 1], 3 : [3 \ 3], 4 : [4 \ 3]$
- **Matriz de distâncias:**

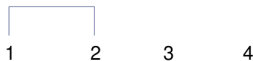
	1	2	3	4
1	0	1	5	10
2	1	0	8	13
3	5	8	0	1
4	10	13	1	0

Exemplo de agrupamento aglomerativo

- **Dados:** 1 : [1 2], 2 : [1 1], 3 : [3 3], 4 : [4 3]
- **Matriz de distâncias:**

	1	2	3	4
1	0	1	5	10
2	1	0	8	13
3	5	8	0	1
4	10	13	1	0

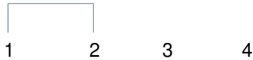
$$c_{1,2} = [1 \ 1.5]$$



Exemplo de agrupamento aglomerativo

- **Dados:** 1, 2 : $[1 \ 1.5]$, 3 : $[3 \ 3]$, 4 : $[4 \ 3]$
- **Matriz de distâncias:**

	1,2	3	4
1,2	0	6.25	11.25
3	6.25	0	1
4	11.25	1	0



1 2 3 4

Exemplo de agrupamento aglomerativo

- **Dados:** 1, 2 : [1 1.5], 3 : [3 3], 4 : [4 3]
- **Matriz de distâncias:**

	1,2	3	4
1,2	0	6.25	11.25
3	6.25	0	1
4	11.25	1	0

$$c_{3,4} = [3.5 \ 3]$$



Exemplo de agrupamento aglomerativo

- **Dados:** 1, 2 : [1 1.5], 3, 4 : [3.5 3]
- **Matriz de distâncias:**

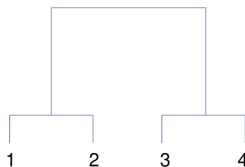
	1,2	3,4
1,2	0	8
3,4	8	0



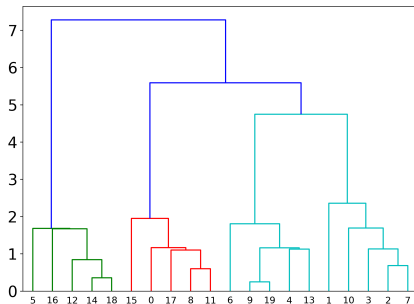
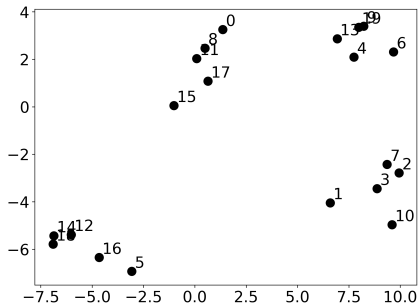
Exemplo de agrupamento aglomerativo

- **Dados:** 1, 2 : $[1 \ 1.5]$, 3, 4 : $[3.5 \ 3]$
- **Matriz de distâncias:**

	1,2	3,4
1,2	0	8
3,4	8	0



Exemplo de agrupamento aglomerativo



Agenda

- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

Algoritmo K-Médias

- Seja o conjunto de dados $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$.
- Devemos particionar o conjunto \mathcal{X} em K clusters $\mathcal{C} = \{C_k\}_{k=1}^K$.

Algoritmo K-Médias

- Seja o conjunto de dados $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$.
- Devemos particionar o conjunto \mathcal{X} em K clusters $\mathcal{C} = \{C_k\}_{k=1}^K$.
- Padrões em um **mesmo cluster** devem ter **alta similaridade**.
- Padrões em **clusters diferentes** devem ter **baixa similaridade**.

Algoritmo K-Médias

- Seja o conjunto de dados $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$.
- Devemos particionar o conjunto \mathcal{X} em K clusters $\mathcal{C} = \{C_k\}_{k=1}^K$.
- Padrões em um **mesmo cluster** devem ter **alta similaridade**.
- Padrões em **clusters diferentes** devem ter **baixa similaridade**.
- Agrupamento como um problema de otimização:

Otimize $f(\mathcal{X}, \mathcal{C})$,

$$\text{s.a. } C_k \neq \emptyset, \quad \forall k,$$

$$C_k \cap C_{k'} = \emptyset, \quad \forall k \neq k',$$

$$\bigcup_{k=1}^K C_k = \mathcal{X},$$

em que $f(\cdot)$ é uma **função de similaridade** ou **dissimilaridade**.

Algoritmo K-Médias

Algoritmo de Lloyd

- Defina o **erro de quantização** ou **erro de reconstrução** por:

$$\mathcal{J}(\mathcal{C}) = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2.$$

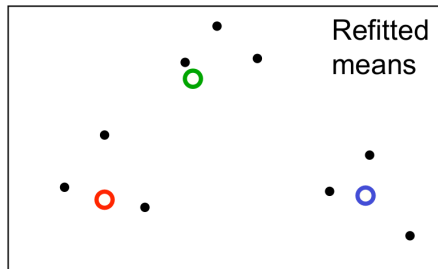
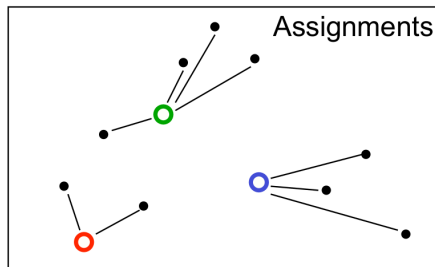
- Escolha um particionamento inicial definido pelos centróides $\mathbf{m}_k \in \mathbb{R}^D, 1 \leq k \leq K$.
- Execute os seguintes passos abaixo iterativamente:
 - Encontre todas as partições $C_k, k = 1, \dots, K$:

$$C_k = \{\mathbf{x}_i \in \mathbb{R}^D \mid \|\mathbf{x}_i - \mathbf{m}_k\|^2 < \|\mathbf{x}_i - \mathbf{m}_j\|^2, \forall j \neq k\}.$$

- Recalcule os centróides dos clusters: $\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i, \forall k$.

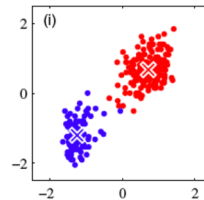
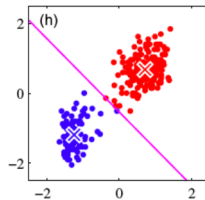
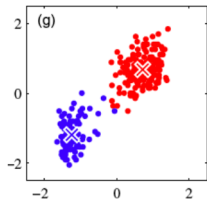
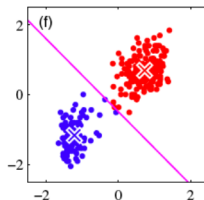
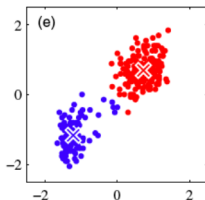
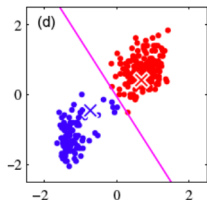
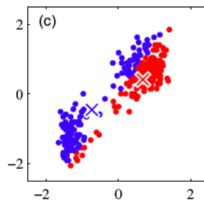
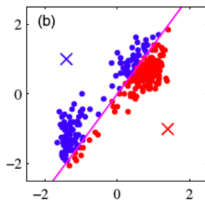
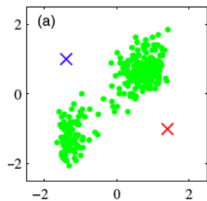
- Repita até os centróides não apresentarem grandes modificações.

Algoritmo K-Médias



- Note que as posições dos centróides são ajustadas ao longo das iterações visando a redução do erro de reconstrução.
- **Otimização coordenada:**
 - Fixe os centróides, ajuste a designação de cada padrão.
 - Fixe as designações, ajuste as posições dos centróides.
- Caso particular do algoritmo *Expectation-Maximization* (EM).

Algoritmo K-Médias



Algoritmo K-Médias

Algoritmo K-Médias generalizado

- Uso de funções de dissimilaridade ou similaridade diferentes.
 - Distância de Manhattan (menos sensível a *outliers*).
 - Distância de Mahalanobis (obtem clusters elípticos).
 - Similaridade por funções de kernel (dados não numéricos).
- O algoritmo de obtenção dos centróides é semelhante ao K-Médias convencional.
 - A métrica alternativa influencia somente a etapa de localização das partições.

Algoritmo K-Médias

- **Problema:** O algoritmo K-médias resolve uma **otimização não-convexa** (presença de mínimos locais).

Algoritmo K-Médias

- **Problema:** O algoritmo K-médias resolve uma **otimização não-convexa** (presença de mínimos locais).
- **Ideias:** Múltiplas inicializações, inicialização cuidadosa.

Algoritmo K-Médias

- **Problema:** O algoritmo K-médias resolve uma **otimização não-convexa** (presença de mínimos locais).
- **Ideias:** Múltiplas inicializações, inicialização cuidadosa.
- **Problema:** Como escolher um valor de K adequado?

Algoritmo K-Médias

- **Problema:** O algoritmo K-médias resolve uma **otimização não-convexa** (presença de mínimos locais).
- **Ideias:** Múltiplas inicializações, inicialização cuidadosa.
- **Problema:** Como escolher um valor de K adequado?
- **Ideia:** Usar métricas de qualidade do particionamento.

Métricas de qualidade do particionamento

- Índice Davies-Bouldin (DB)

$$\text{DB}(\mathcal{C}) = \frac{1}{K} \sum_{k=1}^K \max_{k \neq k'} \left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}} \right).$$

→ δ_k : espalhamento intra-agrupamento (*within cluster scatter*)

→ $\Delta_{kk'}$: espalhamento entre grupos (*between cluster distance*)

$$\delta_k = \frac{1}{N_k} \sum_{\mathbf{x}_n \in C_k} \|\mathbf{x}_n - \mathbf{m}_k\|,$$

$$\Delta_{kk'} = \|\mathbf{m}_k - \mathbf{m}_{k'}\|,$$

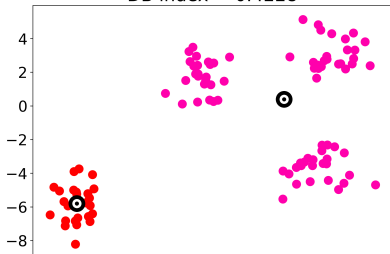
$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i.$$

→ Equilibra soluções com clusters compactos e separados entre si.

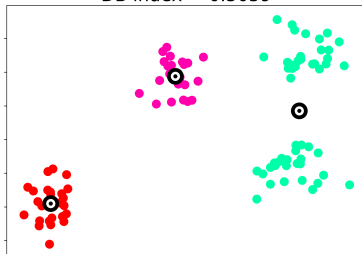
→ Quanto menor seu valor, melhor a solução.

Exemplo de aplicação do algoritmo K-Means

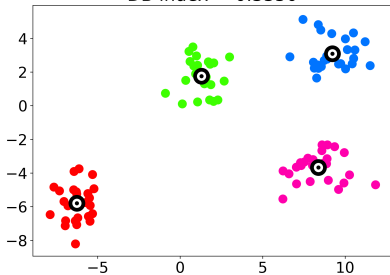
K = 2, Reconstruction error = 1792.47
DB index = 0.4228



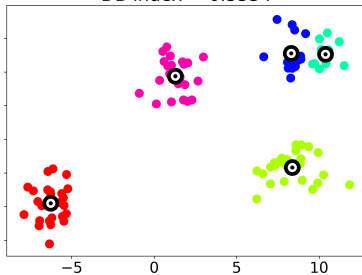
K = 3, Reconstruction error = 782.23
DB index = 0.5059



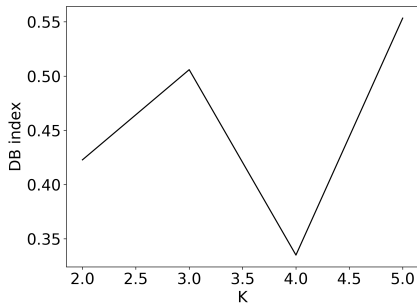
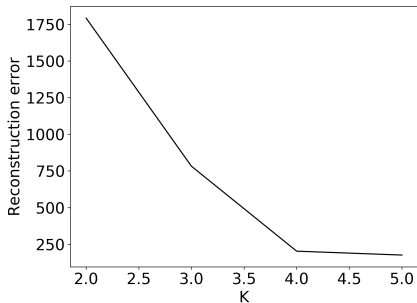
K = 4, Reconstruction error = 202.66
DB index = 0.3350



K = 5, Reconstruction error = 176.16
DB index = 0.5534



Exemplo de aplicação do algoritmo K-Means



Agenda

- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

Redes Neurais *Radial Basis Function*

Redes RBF

- Uma camada oculta com função de ativação radial:

$$z_0 = 1, \quad z_j = \phi_1(\mathbf{w}_j, \mathbf{x}_i) = \rho(\|\mathbf{x}_i - \mathbf{w}_j\|), \quad 1 \leq j \leq N_H$$

- Exemplo: função de base Gaussiana com hiperparâmetro $\gamma > 0$:

$$\rho(\|\mathbf{x}_i - \mathbf{w}_j\|) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{w}_j\|^2)$$

- A saída do modelo é dada por:

$$\hat{y}_i = \phi_2(\mathbf{M}\mathbf{z})$$

- **Vantagem:** Somente os pesos da camada de saída \mathbf{M} são atualizados, por exemplo, via SGD.
- Os vetores $\mathbf{w}_j|_{j=1}^{N_H}$ podem ser obtidos via clustering de $\mathbf{x}_i|_{i=1}^N$.

Agenda

- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

Tópicos adicionais

- Algoritmo Expectation-Maximization (EM).
 - Permite inferência em modelos com variáveis não-observadas.
- GMM (Gaussian Mixture Model):

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Quantização vetorial.
- Mapas de Kohonen (Self-Organizing Maps - SOM).

Agenda

- ① Aprendizagem não-supervisionada
- ② Agrupamento (clustering) de dados
- ③ Algoritmo K-Médias
- ④ Exemplo de aplicação: Redes RBF
- ⑤ Tópicos adicionais
- ⑥ Referências

Referências bibliográficas

- **Cap. 11** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 9** - BISHOP, C. **Pattern recognition and machine learning**, 2006.