Machine Learning
13-Gen-2021 (90 minuti)

Matricola: ____________________
Cognome: __________ Nome: ____________

1) Cosa si intende per iperparametri? Fornire esempi pratici di iperparametri. Come si ottimizzano?

Dispense “Fondamenti”

2) Con quali tecniche si può estendere SVM da 2 a più classi?

Dispense “Classificazione (2)”

3) Cosa si intende per Clustering esclusivo e Clustering soft (o Fuzzy). Quest’ultimo che vantaggi può avere?

Dispense “Clustering”

4) Che cosa codifica la funzione Q nell’ambito dell’approccio Q-learning? Quali sono gli input che ne determinano il valore?

Dispense “Deep Learning (2)”

---

5) Date due distribuzioni multinormali identificate dai seguenti parametri:

$\boldsymbol{\mu}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$

$P(w_0) = 0{,}5$

$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$

$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$

$P(w_1) = 0{,}5$

Nell’ipotesi dell’impiego di un classificatore di Bayes multinormale, calcolare per il punto $\boldsymbol{x} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$:

- le densità di probabilità condizionali;
- l’indice della classe restituita in output.

Si ricorda che:

- la densità di probabilità, nel caso della distribuzione multinormale è:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\cdot(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{t}}\cdot\boldsymbol{\Sigma}^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu})}$$

- l’inversa di una matrice diagonale si ottiene invertendo i singoli elementi;
- il determinante di una matrice diagonale si ottiene moltiplicando gli elementi della diagonale.

Svolgimento

$$\boldsymbol{x} - \boldsymbol{\mu}_0 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_0^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$|\boldsymbol{\Sigma}_0| = 0{,}5$$

$$p(\boldsymbol{x}|w_0) = \frac{1}{2\pi \cdot \sqrt{0{,}5}} \cdot e^{-\frac{1}{2}\cdot\begin{bmatrix} -2 & -1 \end{bmatrix}\cdot\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\cdot\begin{bmatrix} -2 \\ -1 \end{bmatrix}}$$

$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \cdot (-2) + 0 \cdot (-1) \\ 0 \cdot (-2) + 1 \cdot (-1) \end{bmatrix} = \begin{bmatrix} -4 \\ -1 \end{bmatrix}$

$[-2 \quad -1] \cdot \begin{bmatrix} -4 \\ -1 \end{bmatrix} = (-2) \cdot (-4) + (-1) \cdot (-1) = 9$

$p(\boldsymbol{x}|w_0) = 0{,}225 \cdot e^{-\frac{9}{2}} = 0{,}0025$ (Densità di probabilità condizionale di $\boldsymbol{x}$ data w_0)

$\boldsymbol{x} - \boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} - \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$\boldsymbol{\Sigma}_1^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

$|\boldsymbol{\Sigma}_1| = 0{,}5$

$p(\boldsymbol{x}|w_1) = \dfrac{1}{2\pi \cdot \sqrt{0{,}5}} \cdot e^{-\frac{1}{2} \cdot [0 \quad 1] \cdot \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}}$

$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 0 + 0 \cdot 1 \\ 0 \cdot 0 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

$[0 \quad 1] \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix} = 0 \cdot 0 + 1 \cdot 2 = 2$

$p(\boldsymbol{x}|w_1) = 0{,}225 \cdot e^{-\frac{2}{2}} = 0{,}082802$ (Densità di probabilità condizionale di $\boldsymbol{x}$ data w_1)

Visto che la probabilità a priori delle due classi è uguale, la classe predetta sarà quella con densità di probabilità condizionale maggiore.

Indice della classe restituita: **1**

6) Un classificatore *Nearest Neighbor* (NN), con un *training set* (TS) composto da $n = 1000$ pattern di dimensionalità $d = 5$, utilizza come metrica la *distanza euclidea*. Calcolare il numero di somme, sottrazioni e moltiplicazioni necessarie (trascurando la radice quadrata) per effettuare la classificazione di un pattern x supponendo che non vengano utilizzate strutture spaziali specifiche per indicizzare il TS.

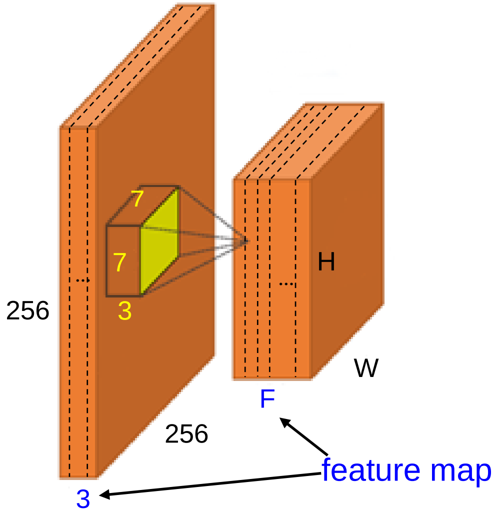
Svolgimento

Per classificare un pattern è necessario calcolare le distanze da tutti gli n pattern del TS e selezionare la classe del pattern con distanza minima. Sapendo che la distanza euclidea tra due punti d-dimensionali è definita come $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_d - q_d)^2}$, per calcolare ognuna delle n distanze saranno necessarie $(d - 1)$ somme, d sottrazioni e d moltiplicazioni (elevamento al quadrato).
Con un TS composto da 1000 pattern 5-dimensionali saranno necessarie:

- $(5 - 1) \cdot 1000 = 4000$ somme;
- $5 \cdot 1000 = 5000$ sottrazioni;
- $5 \cdot 1000 = 5000$ moltiplicazioni.

7) Dati un volume di input e un livello di convoluzione in una CNN, aventi le seguenti caratteristiche:

- *Volume Input*: $3 \times 256 \times 256$
- 64 filtri di dimensione $3 \times 7 \times 7$
- *Stride: 3*
- *Padding: 0*



Si calcoli motivando la risposta:

- La dimensione del volume di output: $F \times W \times H$;
- il numero totale di connessioni e di pesi del livello (senza considerare i bias).

Svolgimento

Il numero di feature map F del volume di output è pari al numero di filtri utilizzati: 64.
Essendo larghezza e altezza uguali tra loro sia nel volume di input che nei filtri, le dimensioni spaziali del volume di output (W e H) saranno anch'esse uguali tra loro.

$$W = H = \frac{D_{IN} - D_F + 2 \cdot Padding}{Stride} + 1 = \frac{256 - 7 + 2 \cdot 0}{3} + 1 = \frac{249}{3} + 1 = 84$$

Ogni neurone del livello di output $(64 \times 84 \times 84)$ è connesso con tanti neuroni del livello di input pari alla dimensione del filtro $(3 \times 7 \times 7)$.
Pertanto, **il numero totale di connessioni** è $(64 \times 84 \times 84) \cdot (3 \times 7 \times 7) = 66\,382\,848$.

Il **numero totale di pesi**, invece, risulta molto più piccolo giacché in una CNN i pesi di ciascun filtro sono condivisi da tutti i neuroni contenuti in una stessa feature map. Visto che il numero di feature map è uguale a 64, e il numero di elementi di ciascun filtro è pari a $(3 \times 7 \times 7)$, il numero totale di pesi (non considerando il bias) è $(3 \times 7 \times 7) \times 64 = 9408$.