UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

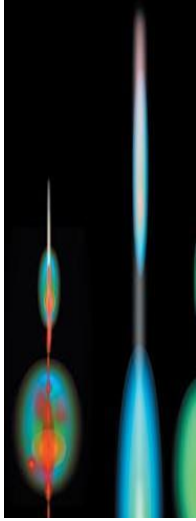
Artificial Intelligence
at Roma Tor Vergata

Introduzione al Test in Itinere

Roberto Basili

Università di Roma, Tor Vergata

Esempi svolti:



- Clustering

Segnalare tra le seguenti quali sono le affermazioni corrette riguardo ad un processo di *Text Clustering*:

(A) L'algoritmo K-means costruisce una tassonomia delle classi di documenti.

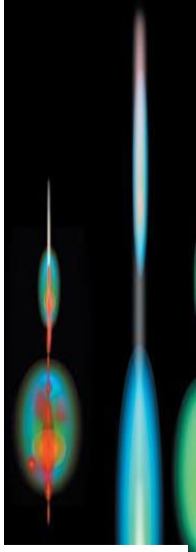
(B) Un algoritmo di tipo *Hierarchical Agglomerative Clustering* puo' applicare ad ogni passo una metrica di tipo *Single Link* tra documenti per la scelta del migliore raggruppamento.

(C) Una metrica di tipo *Single Link* esprime la migliore distanza tra classi di documenti per algoritmi agglomerativi.

(D) Negli algoritmi agglomerativi, una metrica di tipo *Single Link* determina classi di tipo sferico tra i documenti.

Esempi

- SVM



51. Se $\vec{x}_i$ è un support vector ottenuto con l'algoritmo delle hard-margin SVMs quale affermazione risulta falsa?

(A) $y_i(\vec{w} \cdot \vec{x_i} + b) - 1 < 0$.

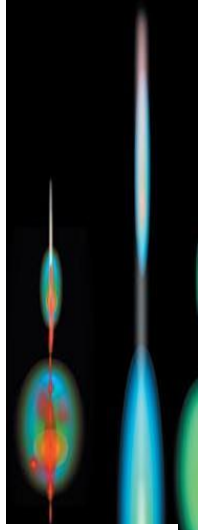
(B) Il moltiplicatore di Lagrange associato $\alpha_i \neq 0$.

(C) Se $\vec{x}_j$ è un'altro support vector con $y_j = -y_i$ allora $b = -\frac{\vec{w} \cdot \vec{x}_i + \vec{w} \cdot \vec{x}_j}{2}$

(D) Il margine geometrico del training set è $y_i(\vec{w} \cdot \vec{x_i} + b)$

Esempi

- Soft margin SVM



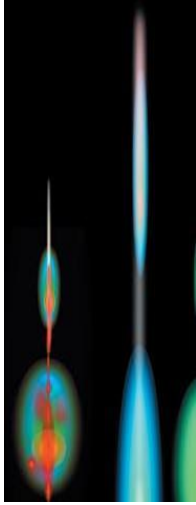
57. Individuare l'affermazione *errata* rispetto al seguente sistema:

$$\begin{cases} min \quad ||\vec{w}|| + C\sum_{i=1}^{m} \xi_i^2 \\ y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i, \quad \forall i = 1, .., m \\ \xi_i \geq 0, \quad i = 1, .., m \end{cases}$$

(A) Se il parametro C tende a 0 i vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i$ tendono ad essere equivalenti ai vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$
(B) $\sum_{i=1}^{m} \xi_i^2$ non conta esattamente il numero degli errori commessi dal iperpiano di separazione.

(C) Se esiste $\xi_i > 1$ il punto $\vec{x_i}$ non è classificato correttamente.
(D) $\sum_{i=1}^{m} \xi_i$ è una misura alternative dell'errore.

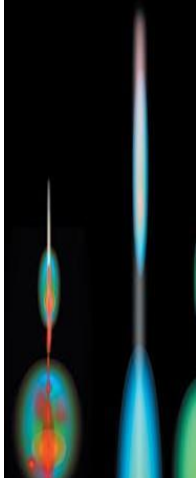
Esempi



- Rocchio

75. Data una classeC_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|C_i|} \vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
segnalare la affermazione corretta?
(A) È un algoritmo quadratico.
(B) È un iperpiano di separazione che divide perfettamente gli esempi di training.
(C) È un iperpiano di separazione il cui gradiente è la differenza tra la media degli esempi positivi e la media degli esempi negativi.
(D) È un iperpiano di separazione simile a quello espresso dal percettrone.

Esempi



- Valutazione delle Prestazioni

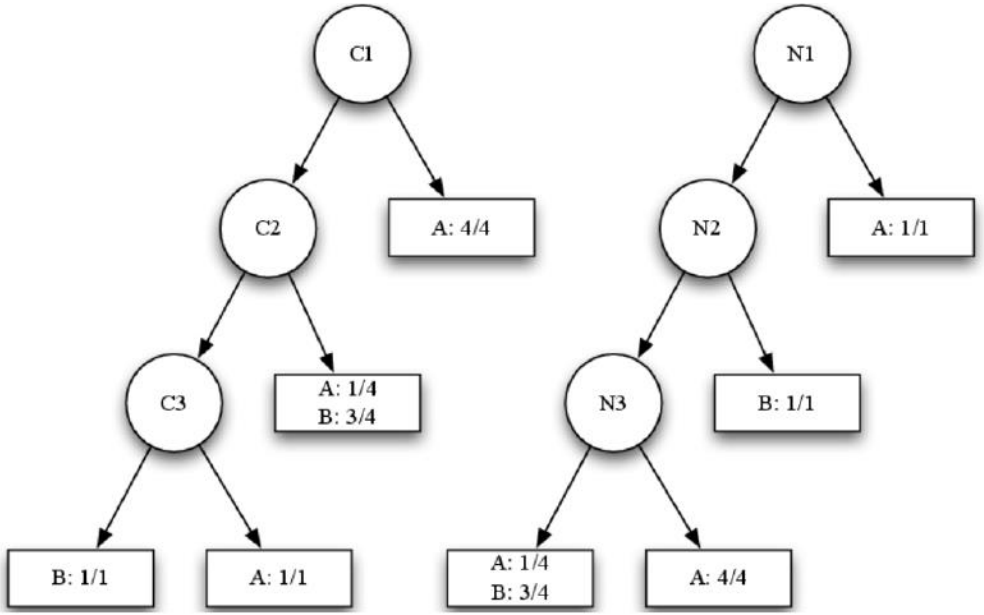
78. Cosa s'intende per n-fold cross validation?
(A) Dati degli esempi di training e di testing si apprendono i modelli sul training e si testano sul testing.

(B) Dati degli esempi di training e di testing si apprendono i modelli sul testing e si testano sul training.

(C) Si divide il corpus di documenti in n parti; a rotazione una viene usata per il testing e $n-1$ sono usate per il training.
(D) Si divide il training in n parti e si addestra il classificatore n volte; ogni volta si misura la performance sul test-set.

Esempi Domande Calcolo

5. Dati gli alberi in figura scegliere le affermazioni più corretta:



(A) La probabilità del SOLO nodo C1 di individuare la classe A correttamente è maggiore del nodo N1 [ ]

(B) La probabilià del SOLO nodo C1 di individuare la classe A correttamente è uguale del nodo N1 [ ]

(C) La classe A viene sempre correttamente riconosciuta in entrambi gli alberi [ ]

(D) L'Information Gain del nodo N2 e' maggiore del nodo C2 [ ]

Esercizo di Modellazione

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Domanda

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

- A. Determinare:
  - L'equazione dell'iperpiano di una SVM lineare con hard margin, cioè i valori attesi di $\underline{w}$ e b.
  - Scrivere la funzione di classificazione tra Blue e Red
  - ed il corrispondente valore del margine
- B. Introdurre un ulteriore punto P1 nel dataset che mantenga l'equazione invariata
- C. Introdurre un punto P2 per il quale la soluzione hard margin deve essere cambiata e calcolare la nuova soluzione ed il nuovo margine.

Temi d' Esame: Domanda aperta

Discutere la applicazione di una modellazione markoviana ai task di tipo *sequence labeling*.
(E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di *Part-Of-Speech tagging* di frasi in linguaggio naturale)

- Definire le assunzioni di base,
- La nozione di stato, transizione ed emissione
- Le equazioni generali del modello
- I metodi di soluzione
- Possibili misure di valutazione

Variante

- Utilizzare una tecnica di tipo HMM per il problema della *tokenizzazione* di un testo libero.
- Si usino come etichette di stato le etichette IOB che stabiliscono l'inizio (B), l'interno (I) e la uscita (O) da un *token*.
- Si definiscano l'alfabeto degli stati e quello delle osservazioni, le matrici di transizione e di emissione. Si discuta infine la possibile tecnica di stima dei parametri applicabile al task, e gli eventuali problemi ad essa connessi.

Soluzioni domande a risposta multipla