UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

Artificial Intelligence
at Roma Tor Vergata

Introduzione al Test in Itinere

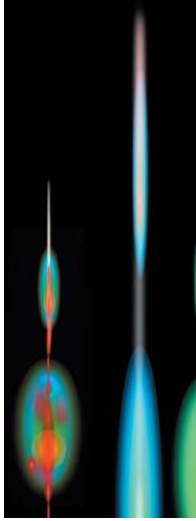
Roberto Basili
Università di Roma, Tor Vergata

Argomenti oggetto di esame

- Rappresentazioni vettoriali per la classificazione
- Clustering
- Algoritmi di apprendimento automatico per la classificazione
  - K-NN
  - DTs
  - NB
  - Rocchio
- Valutazione dei sistemi di classificazione
- Modelli Markoviani
  - Language models & HMMs
  - Example: POS tagging
- Statistical Learning Theory:
  - PAC-learning
  - VC dimension
  - SVMs
  - Kernels
- Online learning
- Latent Semantic Analysis

Esempi domande d’esame

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.

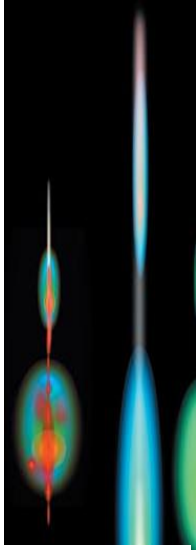
(B) Un algoritmo di tipo *Hierarchical Agglomerative Clustering* puo' applicare ad ogni passo una metrica di tipo *Single Link* tra documenti per la scelta del migliore raggruppamento.

(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

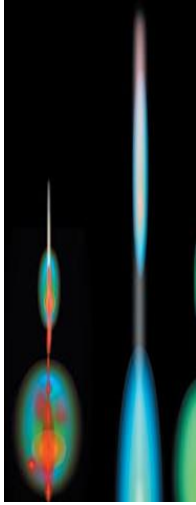
57. Individuare l'affermazione *errata* rispetto al seguente sistema:

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

(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.

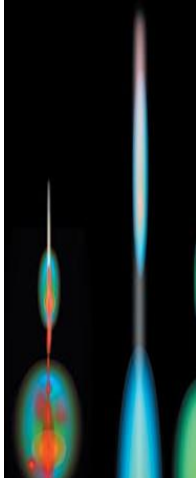
Esempi



- 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.

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

Temi d'Esame: Domanda aperta

Discutere la differenza tra un modello multivariato (binomiale) ed un modello multinomiale nei processi di classificazione *bayesiana*.
(E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di *classificazione di documenti*)

- Definire le assunzioni di base,
- La nozione di evento, spazio campione e caso possibile
- Le equazioni generali del modello
- I metodi di soluzione
- Possibili misure di valutazione

Temi d'Esame: Domanda aperta

Discutere un algoritmo di *clustering* a scelta tra quelli trattati a lezione e la sua applicazione ad un insieme di dati sintetici (ad esempio un insieme di 20 punti rappresentati in uno spazio bidimensionale)

- Definire le assunzioni di base dell'algoritmo
  - Le equazioni generali del modello
- Sviluppare uno pseudo-algoritmo per descrivere l'approccio utilizzato
- Mostrare la applicazione dell'algoritmo rispetto ai dati forniti
- Discutere possibili misure di valutazione