



WM&R: INTRODUZIONE AL TEST IN ITINERE

Roberto Basili

Università di Roma, Tor Vergata

1° MIDTERM TOPICS

- Introduction to ML: Vector Representations for Machine Learning
- Automatic Classification: standard Machine Learning approaches
 - K-NN
 - Decision Trees
 - Naive Bayes classifiers
 - The Rocchio model
- Evaluation of Machine Learning models
- Text and ML: linguistic information, syntax and semantics of NLPs
- Probabilistic Approaches to Text Classification
 - The binomial Bernoulli model
 - The multinomial model: language models
- Markov Models and language modelling
 - Visible and Hidden Markov processes
 - Fundamental Tasks for HMMs
 - Example: POS tagging
- Clustering: k-mean, HAC
- Statistical Learning Theory:
 - PAC-learning, VC dimension
 - The Perceptron



MidTerm open questions: some examples

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.

TOPICS: OPEN QUESTION

- Discuss the application of a Markov modeling to a sequence labeling task.
- (It is welcomed in the answer the presentation of an explicit application such as Part-Of-Speech tagging of natural language sentences)
 - Define basic assumptions of the model,
 - Define the notion of states, transitions and emissions
 - Define the general equations of the model
 - The solution methods
 - Possible measures of performance

VARIANT

- Use an HMM for solving the problem of text *tokenization*.
- (Make use of the IOB state labels that determine the beginning (B), the inner (I) and the outing (O) of a *token*).
- Define the state and observation vocabularies as well as the transition and emission matrices. Finally, discuss the main challenges and solution methods of the parameter estimation problem.

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: DOMANDE APERTE

- Discutere il processo di analisi linguistica di un testo in processi di Information Extraction
- Definire in particolare le differenze tra le diverse informazioni estratte e la nozione di semantica lessicale di un testo libero.
- Discutere un approccio logico alla analisi semantica dei testi (testo, grammatica e forma logica). Esempio su una certa frase fornita in ingresso.
- Definire le diverse potenziali applicazioni della analisi linguistica quali:
 - Analisi semantica dei testi: ad es. Framenet labeling
 - Word Sense Disambiguation nei testi Web

DOMANDA APERTA (SOLI INFORMATICI)

- Definire un modello markoviano che esprima un modello probabilistico del linguaggio:

$$a^n b^m c^k \quad \text{con } n, m, k > 0$$

- che esprime stringhe del tipo

abc, aaabcc, abbbbc, aabbccc,

- e non stringhe del tipo

cbba, cbbc, aaacc, bba, ...

- Si definiscano i parametri del modello in modo tale che valga $p(abcc) > 2p(abbc)$

OPEN QUESTION

- Define a Markov model of the language:

$$a^n b^m c^k \quad \text{with } n, m, k > 0$$

- The language includes strings such as:

abc, aaabcc, abbbbc, aabbccc,

- and excludes strings such as:

cbba, cbbc, aaacc, bba, ...

- Define also the model parameters λ such that:

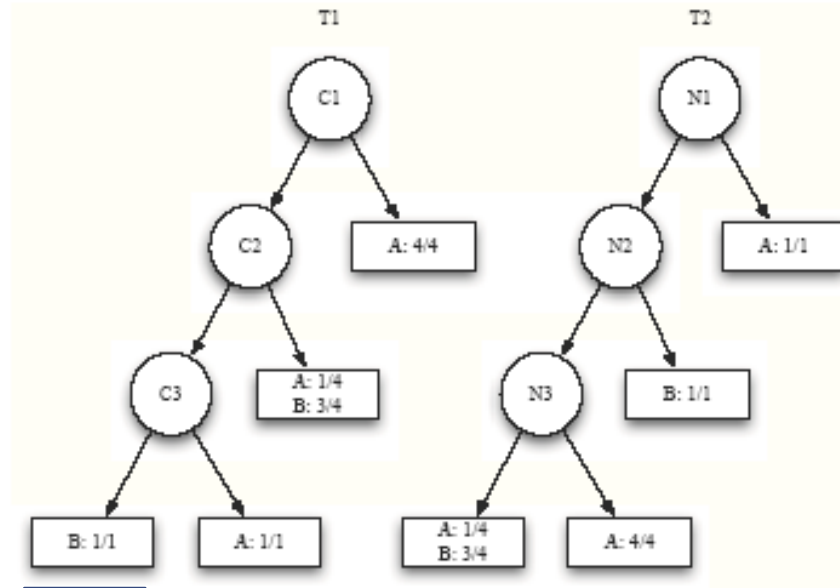
$$p(abcc \mid \lambda) > 2p(abbc \mid \lambda)$$



SOLUZIONI DOMANDE A RISPOSTA MULTIPLA

ESEMPI DOMANDE CALCOLO

5. Dati gli alberi in figura scegliere **le affermazioni più corrette**



- (A) La probabilità del **solo** nodo C1 di individuare la classe A correttamente è maggiore del **solo** nodo N1 []
- (B) La probabilità del 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 C2 è maggiore del nodo N2 []

Mulquestion: è possibile marcare più risposte

ESEMPI DI DOMANDE TRM

- Qual è la differenza tra un algoritmo di *clustering* ed un algoritmo di classificazione
 - Nessuna, l'algoritmo K-mean si applica ad entrambi
 - Il clustering è un processo/task supervised mentre la classificazione non lo è
 - *Nessuna delle altre risposte*
 - Nessuna differenza entrambi si basano su una metrica quantitativa della nozione di similarità tra esempi
 - Sono diversi poiché il clustering genera delle classi mentre la classificazione genera esempi

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

- Rocchio

75. Data una classe C_i ed il classificatore seguente (Rocchio) ,

$$\left(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|C_i|} \vec{d} \right) \cdot \vec{x} - \tau > 0$$

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 perceptrone.

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 DI ALTRE DOMANDE

- Overfitting
 - Natura del problema
 - Definizione
- Tecniche di NLP
 - Parsing vs. analisi di tipo *bag-of-word*
 - Ruolo delle grammatiche nella specifica della sintassi di una lingua
 - Relazione tra semantica e sintassi per le lingue naturali
 - Risorse: cos'è un lessico semantico, come Wordnet
 - Risorse: cos'è Framenet
 - Applicazioni

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:

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

(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.



RISPOSTE

ALTRE DOMANDE

- Data una versione parametrica della equazione di Rocchio, per quali valori dei parametri un certo documento (fatto id un s.i. delle parole del dizionario) ottiene:
 - Una certa categoria X (data la sua eq. di Rocchio)
 - Una probabilità di una categoria X maggiore della probabilità di quella di un'altra Y (disequazione in alfa e beta)
- Per quali valori del parametro ρ_i la forma parametrica di una categoria C_i supera la soglia τ_i di quella categoria C_i

DOMANDA APERTA

- Applicare una HMM ad un linguaggio

Consider a two-bit register. The register has four possible states: 00, 01, 10 and 11. Initially, at time 0, the contents of the register is chosen at random to be one of these four states, each with equal probability. At each time step, beginning at time 1, the register is randomly manipulated as follows: with probability 0.5, the register is left unchanged; with probability 0.25, the two bits of the register are exchanged (e.g., 01 becomes 10); and with probability 0.25, the right bit is flipped (e.g., 01 becomes 00). After the register has been manipulated in this fashion, the left bit is observed. Suppose that on the first three time steps, we observe 0, 0, 1.

(a) Formulate the register as an HMM. What is the probability of transitioning from every state to every other state? What is the probability of observing each output (0 or 1) in each state?

(b) What is the most likely sequence of states given all three observed bits? (Be sure to include the initial state at time 0 in your sequence)