Introduzione al Test in Itinere

Roberto Basili

Università di Roma, Tor Vergata

MidTerm Topics

- Introduction to ML: Vector Representations for Machine Learning
- Clustering Algorithms
- Automatic Classification: Machine Learning approaches
  - K-NN
  - Decision Trees
  - Naive Bayes classifiers
  - The Rocchio model
- Evaluation of Machine Learning models
- Text Representations for IR: linguistic properties and ML features
- Markov Models
  - Language modeling & HMMs
  - Example: POS tagging
- Statistical Learning Theory:
  - PAC-learning
  - VC dimension
  - Perceptrons

MidTerm 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: 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

Domanda Aperta

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

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

- 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:

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

- 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: $p(abcc|\lambda) > 2p(abbc|\lambda)$

Soluzioni domande a risposta multipla