UNIVERSITA' DEGLI STUDI DI FIRENZE

FACOLTA' DI INGEGNERIA – DIPARTIMENTO DI SISTEMI E INFORMATICA

Tesi di Laurea in Ingegneria Informatica

ESTENSIONE DEL CLASSIFICATORE NAIVE BAYES PER LA CATEGORIZZAZIONE DEL TESTO CON DATI PARZIALMENTE ETICHETTATI

Candidato

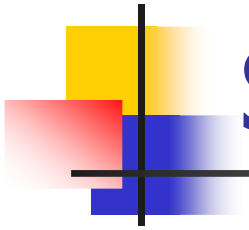
Sauro Menchetti

Relatori

Prof. Paolo Frasconi

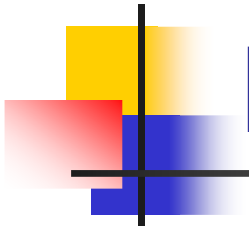
Prof. Giovanni Soda

ANNO ACCADEMICO 1999-2000



Sommario

- La Categorizzazione del Testo
- Rappresentazione del Testo
- Il Classificatore Naive Bayes
- I Dati non Etichettati e l'Algoritmo EM
- Il Nuovo Modello Esteso
- Risultati Sperimentali
- Conclusioni e Prospettive Future

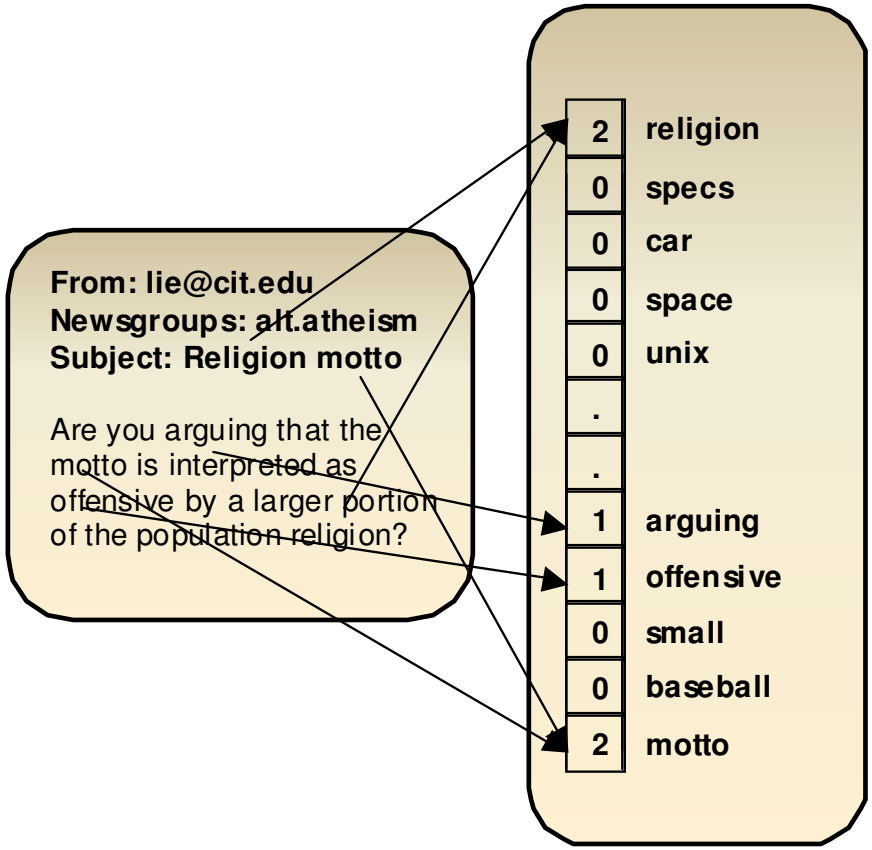


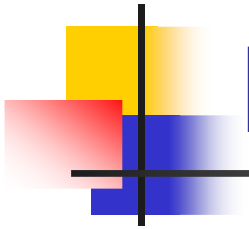
La Categorizzazione del Testo

- Attività rivolta alla costruzione automatica, attraverso tecniche di *Machine Learning*, di classificatori di documenti testuali, ovvero di programmi capaci di etichettare i documenti scritti in linguaggio naturale in un insieme predefinito di categorie

Rappresentazione del Testo

- Le parole costituiscono le unità di rappresentazione (feature)
- Ipotesi semplificative:
  - Viene trascurato l'ordine sequenziale
  - Le parole sono indipendenti dalla loro posizione
- Questa rappresentazione viene detta *Bag Of Words*
- Le parole presenti negli esempi di addestramento formano il vocabolario



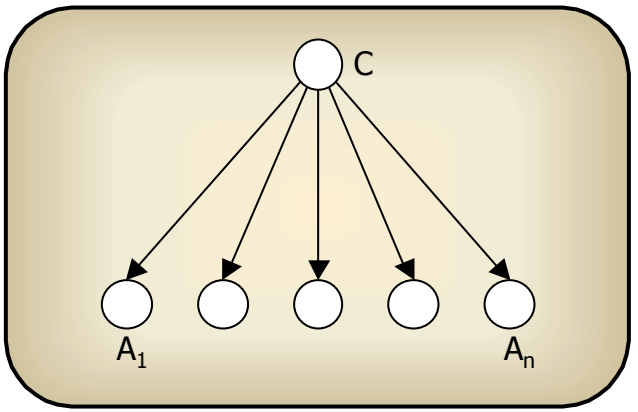


Feature Selection

- Ridurre la dimensione dello spazio delle feature:
  - Rendere più efficiente l'addestramento
  - Incrementare la capacità di generalizzazione del modello
- Tecniche utilizzate:
  - Stop Words: a, and, do, she, with, etc.
  - Stemming: rimuovere il suffisso dalle parole che derivano da una radice comune
  - Pruning basato sulla frequenza delle parole
  - Information Gain: selezionare le parole che hanno la più alta informazione mutua con la classe

Il Classificatore Naive Bayes

- Classificatore basato su un modello probabilistico generativo del documento
- Le istanze sono descritte da un insieme A di attributi
- La classe C può assumere un numero finito di valori
- Ipotesi semplificativa: gli attributi sono condizionalmente indipendenti dato il valore della classe
- Nella categorizzazione del testo gli attributi sono le parole (feature) che compaiono nel documento



$$c_{NB} = \arg\max_{c_j \in C} P(c_j|a_1, \ldots, a_n)$$

$$= \arg\max_{c_j \in C} P(c_j) \prod_{i=1}^{n} P(a_i|c_j)$$

I Dati Non Etichettati e l'Algoritmo EM

- Dati etichettati: risorsa costosa
- Dati non etichettati: sorgente di informazione complementare a basso costo a cui si può ricorrere per incrementare l'accuratezza di classificazione
- L'algoritmo EM fa parte di una classe di algoritmi iterativi per stime ML o MAP nei problemi con dati incompleti
- Permette di utilizzare dati etichettati e non etichettati per stimare i parametri del modello

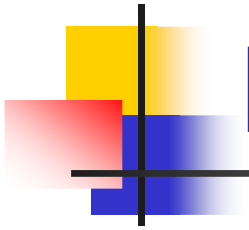
- Massimi locali
- Dati incompleti:
  - Categoria dei documenti: parzialmente osservata
  - Variabili nascoste intermedie associate con le sezioni: mai osservate

Step 1 : Expectation (E) Step

$$Q(\theta \mid \hat{\theta}_i) \leftarrow E_{P(Z \mid D, \hat{\theta}_i)}(\log P(D, Z \mid \theta))$$

Step 2 : Maximization (M) Step

$$\hat{\theta}_{i+1} \leftarrow \arg\max_{\theta} Q(\theta \mid \hat{\theta}_i)$$

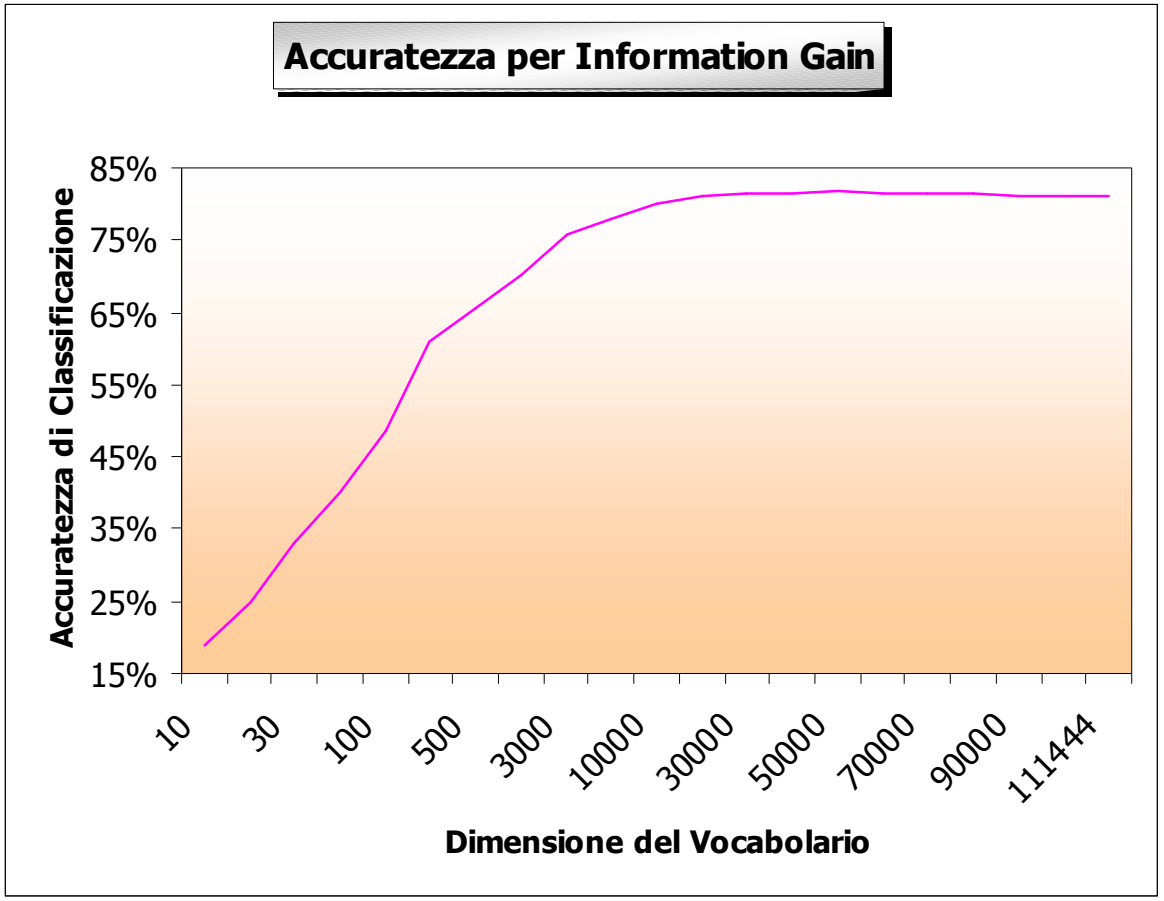


Risultati Sperimentali

- Data Set 20 Newsgroups (Ken Lang 1993)
  - 20000 articoli etichettati suddivisi in 20 diversi gruppi di discussione (1000 per gruppo)
  - Le categorie coincidono con i 20 gruppi di discussione
  - Alcune categorie trattano argomenti dal contenuto simile
- Scopo: classificare ogni articolo nel corrispondente gruppo di discussione a cui è stato inviato
- Training Set e Test Set
  - 4000 articoli costituiscono il test set (200 per classe)
  - 16000 articoli formano il training set (800 per classe) di cui
    - 6000 etichettati e 10000 non etichettati oppure
    - Tutti e 16000 etichettati

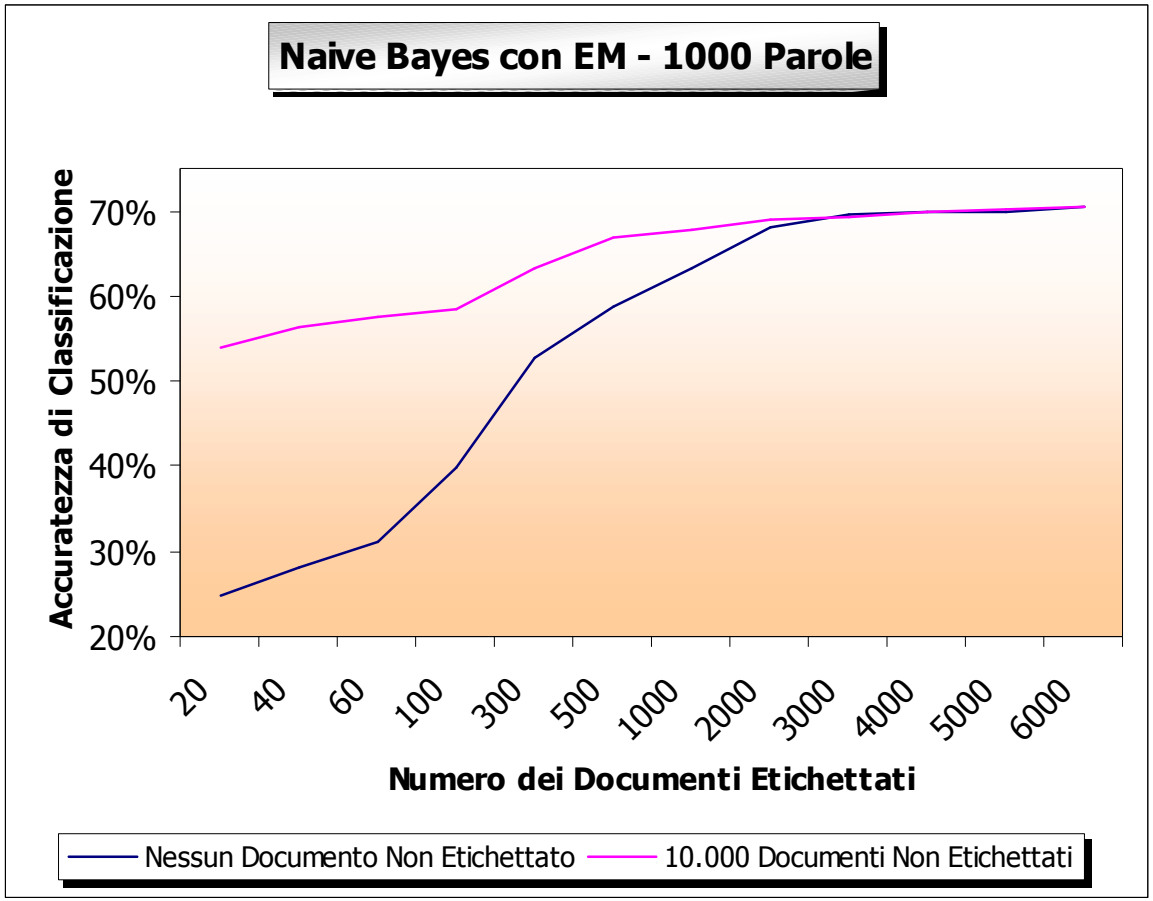
Naive Bayes e Feature Selection

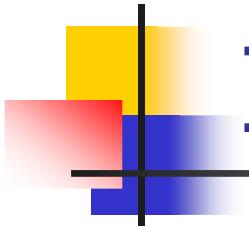
- Feature selection:
  - Stop words
  - Information gain
- L'accuratezza si stabilizza se usiamo più di 10000 parole:
  - 10000: 80%
  - 60000: 81,2%



Naive Bayes e Algoritmo EM

- Incremento consistente soprattutto con pochi dati etichettati
- 20 etichettati (1 per classe):
  - NB: 24,85%
  - EM: 53,93%
  - Riduzione del 38,7% dell'errore di classificazione
- Con più di 2000 dati etichettati si ottengono le stesse prestazioni



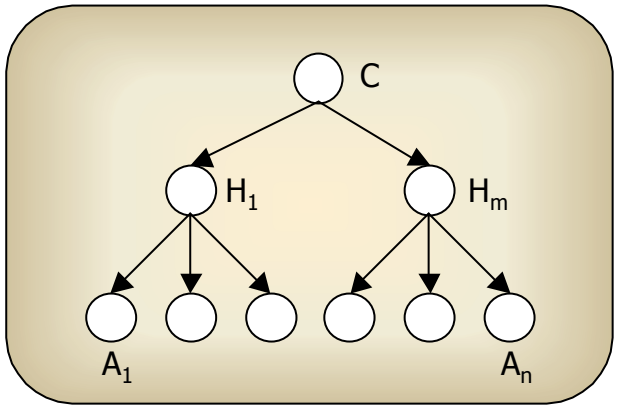


Il Nuovo Modello Esteso (1/2)

- Idea: rilassare l'ipotesi semplificativa del NB al fine di migliorare la modellazione della probabilità generativa del documento
- Si ricorre alla struttura del documento per suddividere l'istanza in unità sintattiche semanticamente significative, dette sezioni (frasi, paragrafi):
  - Le sezioni esprimono meglio delle parole i concetti contenuti nel documento
  - Le sezioni hanno un grado minore di ambiguità rispetto alle parole che le costituiscono
- Si modella la presenza di più sottoargomenti tramite una rappresentazione gerarchica del documento
- Ipotesi sulle sezioni:
  - Le sezioni sono indipendenti tra di loro
  - Le sezioni sono indipendenti dalla posizione che occupano
- Valgono le stesse ipotesi per le parole di una sezione

Il Nuovo Modello Esteso (2/2)

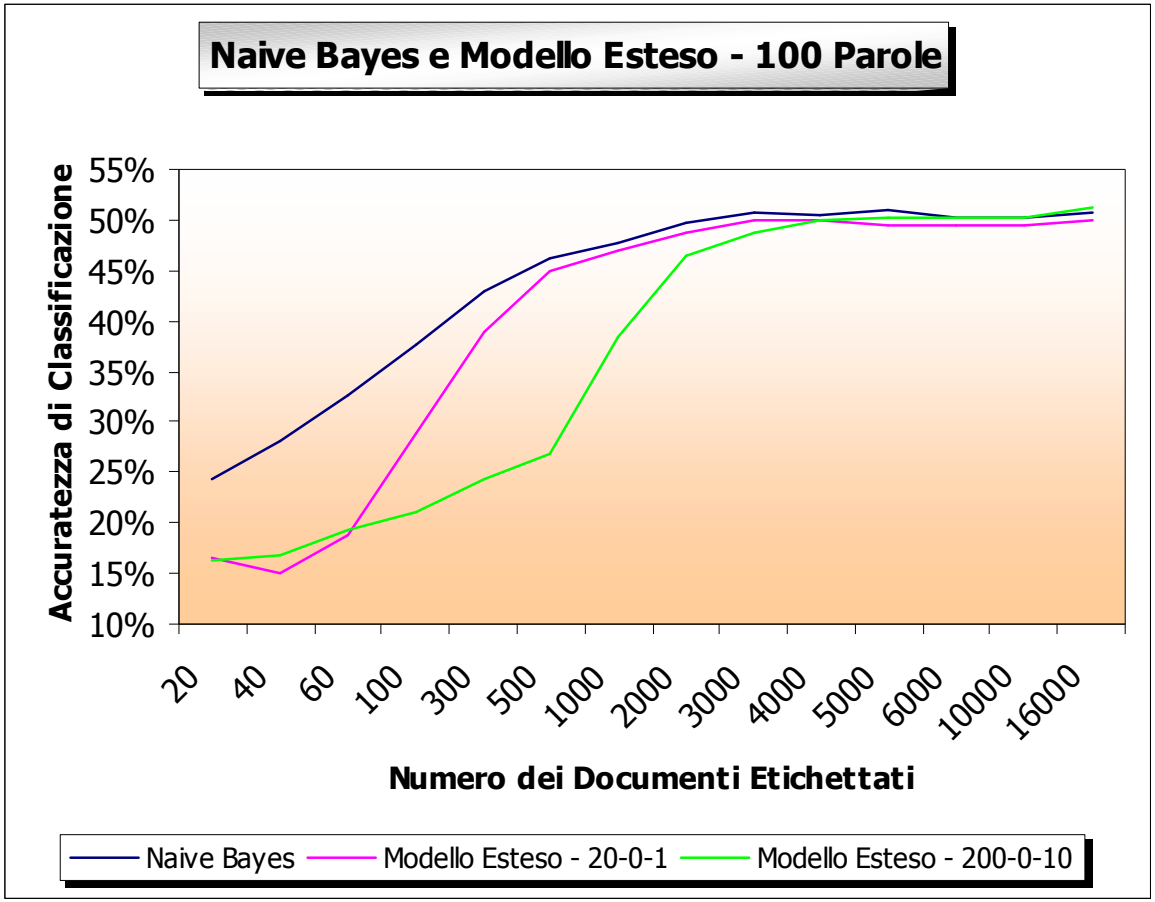
- Si replica la struttura del NB su due livelli
- Le sezioni sono indipendenti data la categoria del documento
- Le parole sono indipendenti data la sottoclasse della sezione
- Le sezioni sono associate con le variabili nascoste intermedie
- Più parametri a disposizione
- *Bag of Bag Of Words*
- Stabilire il numero e la distribuzione tra le classi degli stati delle variabili nascoste che rappresentano le sottoclassi in cui vengono classificate le sezioni



- Apprendimento: si utilizza l'algoritmo EM per stimare i parametri del modello ricorrendo a dati etichettati e non etichettati

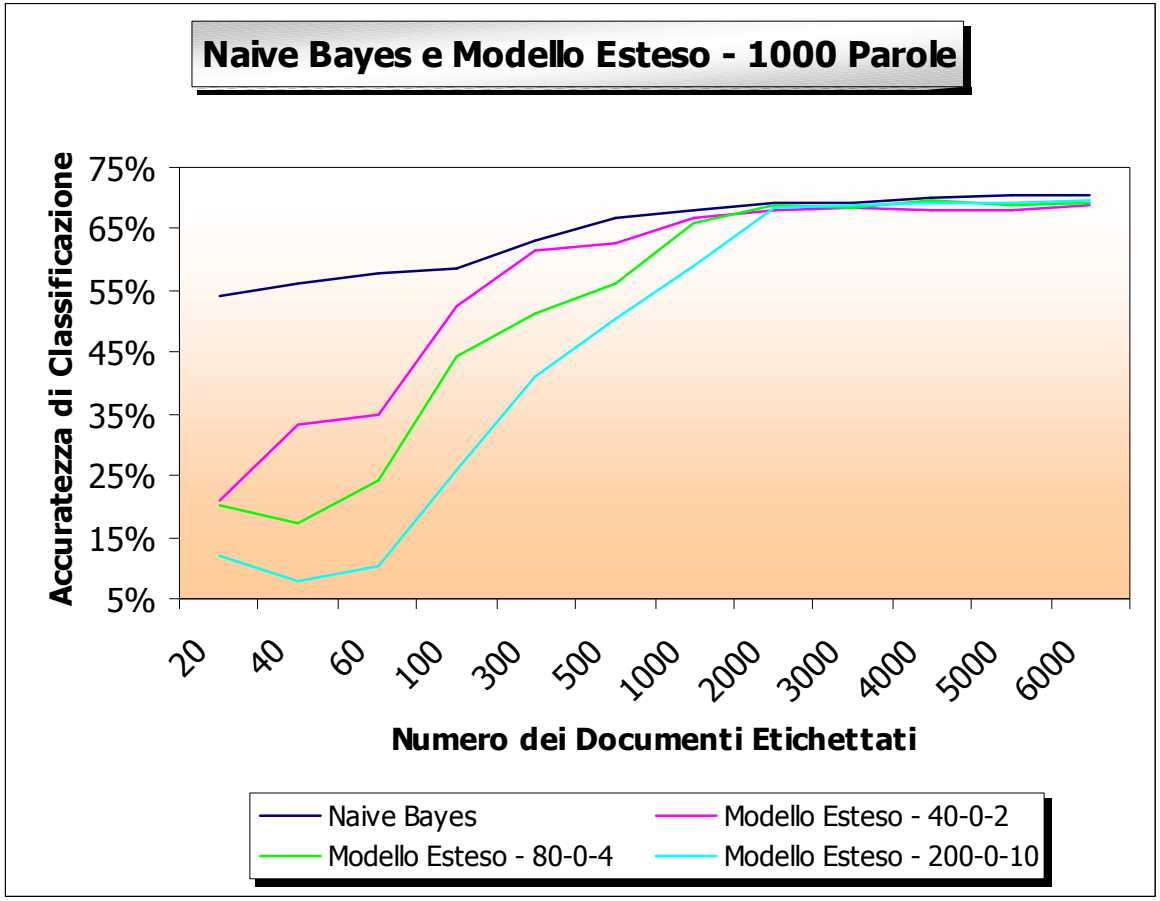
Naive Bayes e Modello Esteso

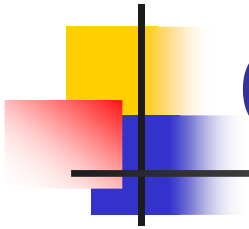
- Solo dati etichettati
- Si sono variati il numero e la distribuzione degli stati delle variabili nascoste
- Prestazioni simili con più di 1000 documenti etichettati
- Piccolo incremento dell'accuratezza con molti dati etichettati



Naive Bayes e Modello Esteso

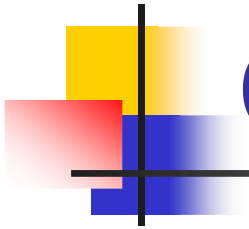
- 10000 documenti non etichettati
- Si sono variati il numero e la distribuzione degli stati delle variabili nascoste
- Prestazioni simili con più di 1000 dati etichettati
- I dati non etichettati forniscono un incremento inferiore delle prestazioni nel ME





Conclusioni (1/2)

- Efficacia della feature selection
- Valore dei dati non etichettati
- ME inferiore al NB con pochi dati etichettati:
  - Troppi parametri rispetto ai dati
  - Inizializzazione casuale dei parametri
- ME paragonabile al NB con un numero consistente di dati etichettati
- In un caso ME migliore di NB
- Un'evidenza anche parziale sulle variabili nascoste porterebbe ad un incremento dell'accuratezza



Conclusioni (2/2)

- Altri lavori hanno tentato di utilizzare delle feature più strutturate delle sole parole con dei risultati piuttosto deludenti
- Il tentativo descritto in questa sede ha cercato di andare oltre, modellando le relazioni tra le parole presenti in una stessa frase
- Prospettive future:
  - Cambiare data set
  - Determinare dai dati il numero e la ripartizione degli stati delle variabili nascoste
  - Inizializzazione guidata dei parametri