Versione Italiana: Cherri Andrea - 200071

Automatic Clasification: Naïve Bayes

Wm&R a.a. 2009/10

R. Basili

(slides borrowed by: H. Schutze)

Dipartimento di Informatica Sistemi e produzione
Università di Roma "Tor Vergata"
Email: basili@info.uniroma2.it

Sommario

- Algoritmo Probabilistico per la Classificazione Automatica (AC)
  - Naive Bayes
  - Lazy Learning (k-NN)
- Stima Parametri
- Valutare un sistema AC

Classificazione Testo: Classificazione del Testo di Naïve Bayes

- Oggi:
  - Introduzione alla Classificazione del Testo
  - Modelli di Linguaggio Probabilistico
  - Categorizzazione del testo di Naïve Bayes

E' Spam?

From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

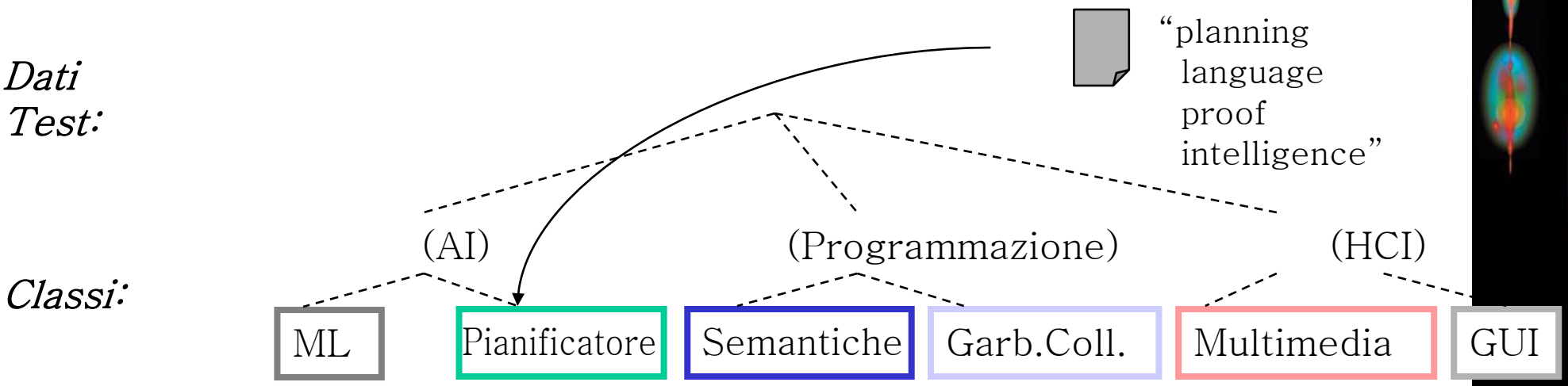
==========
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
==========

Categorizzazione / Classificazione

- Data:
  - Una descrizione di una istanza, $x \in X$, dove X è il *linguaggio istanza* o lo *spazio istanza*.
    - Problema: come rappresentare documenti di testo.
  - Un insieme fisso di categorie:

    $C = \{c_1, c_2, \ldots, c_n\}$
- Determinare:
  - La categoria di x: $c(x) \in C$, dove $c(x)$ è una *funzione di categorizzazione* il cui dominio è X e il cui raggio è C.
    - Vogliamo sapere come costruire funzioni di categorizzazione ("classificatori").

Classificazione Documenti



Dati Training:					
	apprendimento	pianifica	programma	collezione	...
	intelligenza	ragionamento	semantiche	garbage	
	algoritmi	temporale	prova	ottimizzazione	
	rinforzo	pianifica	linguaggio	regione	
	rete	linguaggio		memoria...	

Nota: nella vita reale esiste spesso una gerarchia, non presente nei problemi sopra; e puoi ottenere documenti sugli approcci ML riguardati il Garbage Collector)

Esempi di Categorizzazione Testo

Assegna etichette ad ogni documento o pagina web:

- Le etichette sono spesso argomenti come categorie Yahoo
  - *es., "finance," "sports," "news>world>asia>business"*
- Le etichette possono essere generi
  - *es., "editorials" "movie -reviews" "news"*
- Le etichette possono essere opinioni
  - *e.g., "like", "hate", "neutral"*
- Le etichette possono essere binari specifici del dominio
  - *es., "interesting -to-me" : "not-interesting-to-me", "spam" : "not-spam", "contains adult language" :"doesn't"*

Metodi di Classificazione (1)

- Classificazione Manuale
  - Usata da Yahoo!, Looksmart, about.com, ODP, Medline
  - Molto accurata quando il lavoro è fatto da esperti
  - Consistente quando la grandezza del problema e la squadra sono piccoli
  - Difficoltoso e costoso da scalare

Metodi di Classificazione (2)

- Classificazione automatica di documenti
  - Sistemi basati su regole e codifiche a mano
    - Una tecnica usata dal filtro spam del dipartimento CS, Reuters, CIA, Verity, ...
    - Es., assegnare la categoria se il documenti contiene una data combinazione booleana di parole
    - query: I sistemi commerciali hanno linguaggi query complessi (tutto sui linguaggi di query IR + accumulatori)
    - L'accuratezza è spesso molto alta se una regola è stata raffinata con attenzione nel tempo da un esperto nel campo
    - Costruire e mantenere queste basi di regole è costoso

Metodi di Classificazione (3)

- Apprendimento supervisionato di una funzione di assegnamento etichette ad un documento
  - Molti sistemi parzialmente si basano sull'apprendimento macchina (Autonomy, MSN, Verity, Enkata, Yahoo!, ...)
    - k-Nearest Neighbors (semplice, potente)
    - Naive Bayes (semplice, comune)
    - Macchine a supporto vettore (nuovo, molto potente)
    - ... e molti altri metodi
    - Nessun pranzo gratuito: richiede dati di apprendimento classificati a mano
    - Ma i dati possono essere costruiti (e raffinati) da amatori
- Nota che molti sistemi commerciali usano una mistura di metodi

Metodi Bayesiani

- Metodi di apprendimento e classificazione basati sulla teoria delle probabilità.
- Il teorema di Bayes gioca un ruolo critico nell'apprendimento probabilistico e nella classificazione.
- Costruire un *modello generativo* che approssima come i dati sono prodotti
- Usa probabilità *a priori* di ogni categoria data nessuna informazioni su un oggetto.
- La categorizzazione produce una distribuzione di probabilità *a posteriori* su una possibile categoria data una descrizione di un oggetto.

Regola di Bayes

$$P(C, X) = P(C \mid X)P(X) = P(X \mid C)P(C)$$

$$\boxed{P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}}$$

Ipotesi di *Massimo a posteriori*

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}}\ P(h \mid D)$$

$$= \underset{h \in H}{\operatorname{argmax}}\ \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}}\ P(D \mid h)P(h)$$

Come *P(D)* è costante

Ipotesi di *Massima somiglianza*

Se tutte le ipotesi sono a priori equalmente somiglianti noi dobbiamo solo considerare il termine $P(D|h)$:

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}} \; P(D \mid h)$$

Classificatori Naive Bayes

Compito: Classifica una nuova istanza D basata su una tupla di valori di attributi $D = \langle x_1, x_2, \ldots, x_n \rangle$ in una delle classi $c_j \in C$

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} \; P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \; \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \; P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

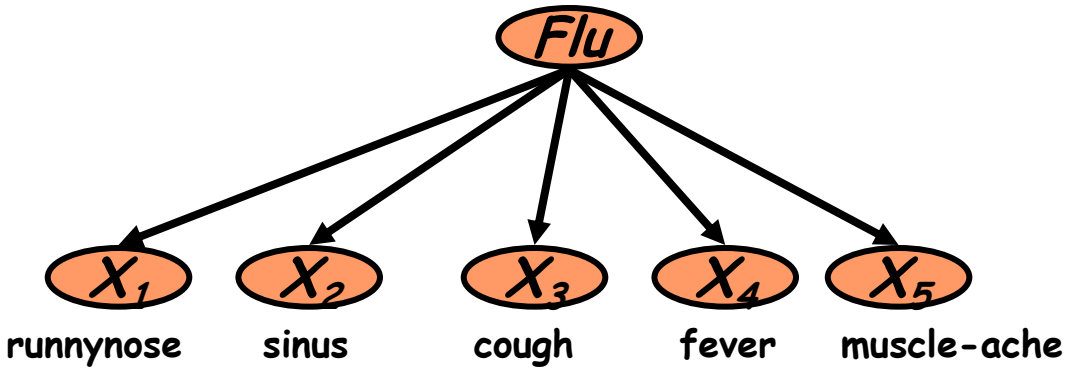
Classificatore di Naïve Bayes
Assunzione di Naïve Bayes

- $P(c_j)$
  - Puo' essere stimato dalla frequenza delle classi nell'esempio di training.
- $P(x_1,x_2,\ldots,x_n|c_j)$
  - $O(|X|^n \bullet |C|)$ parametri
  - Puo' essere solo stimato se è disponibile un numero molto grande di esempi di training.

Assunzione di Indipendenza Condizionale di Naïve Bayes

- Assume che la probabilità di osservare la congiunzione di attributi è uguale al prodotto della probabilità individuale $P(x_i|c_j)$.

Il classificatore di Naïve Bayes

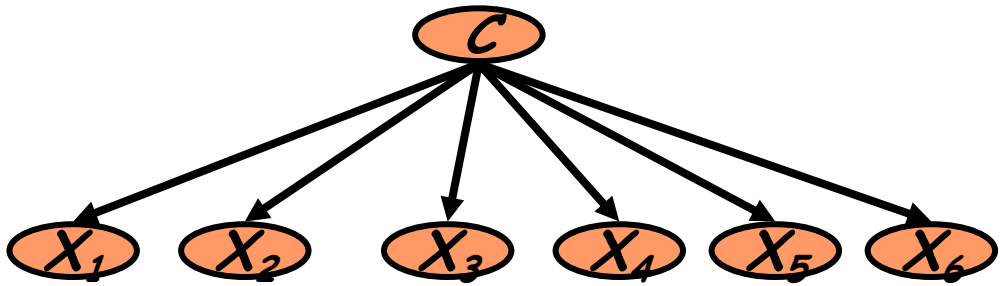


- **Assunzione di Indipendenza Condizionale:** coinvolge l'individuazione della presenza di termini ed è indipendente da ogni altro, data la classe:

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

- questo modello è appropriato per le variabili binarie
  - Modello binomiale multivariato

Imparare il Modello



- Primo tentativo: stima della massima somiglianza
  - semplicemente usa le frequenze nei dati

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

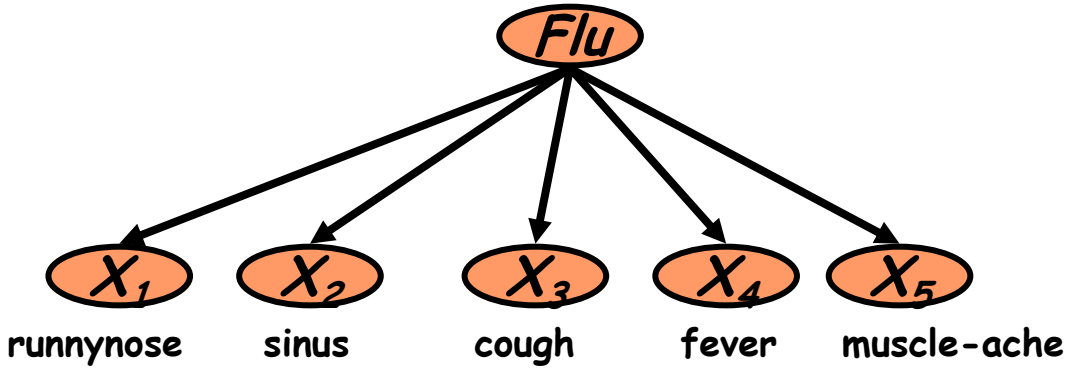
NB Bernoulli: Apprendimento

```
TRAINBERNOULLINB(ℂ, 𝔻)
1  V ← EXTRACTVOCABULARY(𝔻)
2  N ← COUNTDOCS(𝔻)
3  for each c ∈ ℂ
4  do N_c ← COUNTDOCSINCLASS(𝔻, c)
5     prior[c] ← N_c / N
6     for each t ∈ V
7     do N_ct ← COUNTDOCSINCLASSCONTAININGTERM(𝔻, c, t)
8        condprob[t][c] ← (N_ct + 1) / (N_c + 2)
9  return V, prior, condprob
```

NB Modello di Bernoulli: Classificazione

```
APPLYBERNOULLINB(C, V, prior, condprob, d)
1  V_d ← EXTRACTTERMSFROMDOC(V, d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4      for each t ∈ V
5      do if t ∈ V_d
6             then score[c] += log condprob[t][c]
7             else score[c] += log(1 − condprob[t][c])
8  return arg max_{c∈C} score[c]
```

Problema con la Massima Somiglianza



$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

- Cosa succede se non vediamo casi di training dove il paziente non ha sintomi e dolore muscolare?

$$\hat{P}(X_5 = t \mid C = nf) = \frac{N(X_5 = t, C = nf)}{N(C = nf)} = 0$$

- Zero probabilità non possono essere decondizionate, non importa l'altra evidenza!

$$\ell = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Scorrevolezza

- *Scorrevolezza di Laplace*
  - ogni caratteristica ha una probabilità a priori p,
  - E' assunto che è stato osservato in un numero di m esempi virtuali.

$$P(x_j \mid c_i) = \frac{n_{ij} + mp}{n_i + m}$$

- In genere
  - Una distribuzione uniforme di tutte le parole è assunta cosi' che $p = 1/|V|$ e $m = |V|$
  - E' equivalente ad osservare ogni parola nel dizionario una volta per ogni categoria.

Scorrevolezza per evitare l'overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

della diff. valori di X_i

- In qualche modo piu' subdola versione:

k esprime il dato differente bins

frazione generale nei dati dove $X_i = x_{i,k}$

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

estensione di scorrevolezza numero di binari

Modelli di Linguaggio Stocastico

- Modella la *probabilita* di generare stringhe (ogni parola a turno) nel linguaggio (comunemente tutte le stringhe su ∑). Es., modello unigramma

Modello M

0.2	all / la
0.1	un
0.01	uomo
0.01	donna
0.03	detto
0.02	piace
…	

all	uomo	piace	la	donna
0.2	0.01	0.02	0.2	0.01

moltiplica

P(s | M) = 0.00000008

Modelli di Linguaggio Stocastico

- Modella la *probabilita* di generare qualsiasi stringa

Modello M1	
0.2	alla
0.01	classe
0.0001	dice
0.0001	piace
0.0001	yon
0.0005	maiden
0.01	donna

Model M2	
0.2	la
0.0001	classe
0.03	dice
0.02	piace
0.1	yon
0.01	maiden
0.0001	donna

alla	classe	piace	yon	maiden
0.2	0.01	0.0001	0.0001	0.0005
0.2	0.0001	0.02	0.1	0.01

$$P(s|M2) > P(s|M1)$$

Unigramma e modelli di ordine maggiore

- **P(● ○ ● ●)**

  = P(●) P(○ | ●) P(● | ● ○) P(● | ● ○ ●)

- Modelli di Linguaggio Unigramma

  P(●) P(○) P(●) P(●)

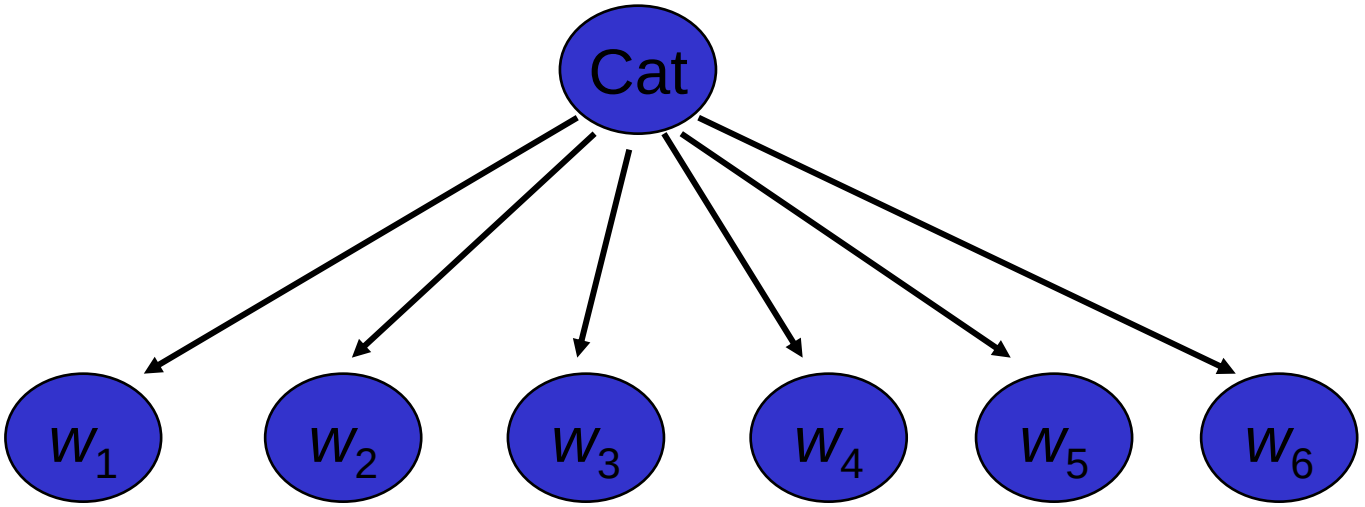
  facile. Efficace!

- Modello di linguaggio a Bigramma (generalmente, n-gramma)

  P(●) P(○ | ●) P(● | ○) P(● | ●)

- Altri Modelli di Linguaggio
  - Modelli basati sulla grammatica (PCFG), ecc.
    - Probabilmente non la prima cosa da provare in IR

Naïve Bayes attraverso un linguaggio a classe condizionale
modello = NB multinomiale



- Effettivamente, la probabilità di ogni classe è formata come un modello di linguaggio ad unigramma specifico per classe

Usare il Classificatore Multinomiale di Naive Bayes per Classificare Testo: Metodo Base

- Gli attributi sono posizioni di testo, i valori sono parole.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j)\prod_i P(x_i \mid c_j)$$

$$= \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j)P(x_1 = \text{"our"} \mid c_j)\cdots P(x_n = \text{"text"} \mid c_j)$$

- Ancora troppe possibilità
- Assume che la classificazione è *indipendente* dalla posizione delle parole
  - Usa stessi parametri per ogni posizione
  - Il risultato è una borsa di modelli di parole (su token, non su tipi)

Naïve Bayes: Apprendimento

- Dal corpo di training, estrae *Vocabulary*
- Calcola i richiesti $P(c_j)$ e i termini $P(x_k \mid c_j)$
  - Per ogni c_j in C esegui
    - $docs_j \leftarrow$ sottoinsieme di documenti per cui la classe obiettivo è c_j
    - $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{documenti \# totali}|}$$
    - $Text_j \leftarrow$ singolo documento contenente tutti i $docs_j$
    - per ogni parola x_k in *Vocabulary*
      - $n_k \leftarrow$ numero di occorrenze di x_k in $Text_j$
      - $$P(x_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Naïve Bayes: Classificazione

- posizioni ← tutte le posizioni delle parole nel documento corrente che contiene tokens trovati in *Vocabulary*
- Ritorna c_{NB}, dove

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \; P(c_j) \prod_{i \in posizioni} P(x_i \mid c_j)$$

Naive Bayes: Complessità di Tempo

- Tempo di Training: $O(|D|L_d + |C||V|))$
  dove L_d è la lunghezza media di un documento in D.
  - Assume V e tutti i D_i, n_i, e n_{ij} pre-compilati in tempo $O(|D|L_d)$ durante un passo lungo tutti i dati.
  - Generalmente $O(|D|L_d)$ quando in genere $C||V| < |D|L_d$
- Test del Tempo: $O(|C| L_t)$
  dove L_t è la lunghezza media di un documento di test.
- In generale molto efficiente, linearmente proporzionale al tempo necessario per leggere tutti i dati.

NB Multinomiale: Algoritmo di Apprendimento

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)

1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

NB Multinomiale: Algoritmo di Classificazione

```
APPLYMULTINOMIALNB(ℂ, V, prior, condprob, d)
1  W ← EXTRACTTOKENSFROMDOC(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4      for each t ∈ W
5      do score[c] += log condprob[t][c]
6  return arg max_{c∈ℂ} score[c]
```

Prevenzione degli underflow

- Moltiplicare molte probabilità, che sono tra o ed 1 per definizione, puo' portare ad un underflow di punti float
- Dato che $\log(xy) = \log(x) + \log(y)$, è meglio eseguire tutte le computazioni sommando log di probabilità invece di moltiplicare probabilità.

- Una classe con il log non-normalizzato finale con il punteggio di probabilità piu alto è ancora la piu probabile.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in posizioni} \log P(x_i \mid c_j)$$

Nota: Due Modelli

- Modello 1: Multivariazione binomiale
  - Una caratteristica X_w per ogni parola nel dizionario
  - X_w = *true* nel documento d se w appare in d
  - Assunzione di Naive Bayes:
    - Dato l argomento del documento, l apparizione di una parola nel documento non ci dice nulla sulla possibilità che un altra parola apparirà
- questo è il modello usato nel modello di indipendenza binaria nei classici probabilistici feedback di rilevanza nei dati classificati a mano (Maron nell'IR è stato uno dei primi utenti di NB).

Due Modelli

- Modello 2: Multinomiale = Unigramma condizionale di classe
  - Una caratteristica X_i per ogni posizione di parola nel documento
    - i valori delle caratteristiche sono tutte parole nel dizionario
  - Valore di X_i è la parola in posizione i
  - Assunzione di Naïve Bayes:
    - Dato l argomento del documento, una parola in una posizione nel documento non ci dice nulla sulle parole in altre posizioni
  - Seconda assunzione:
    - L'apparizione delle parole non dipende dalla posizione

$$P(X_i = w \mid c) = P(X_j = w \mid c)$$

per tutte le posizioni i,j, parola w, e classe c

    - Devi avere una caratteristica multinomiale predicendo tutte le parole

Stima dei parametri

- Modello Binomiale

$$\hat{P}(X_w = true \mid c_j) =$$

frazione di documenti di argomento c_j in cui la parola *w* appare

- Modello Multinomiale:

$$\hat{P}(X_i = w \mid c_j) =$$

frazione di tempio n cui la parola *w* appare lungo tutti i documenti di argomento c_j

  - Puo' creare un mega doeumento per l'argomento *j* concatenando tutti i documenti in questo argomento
  - Usa la frequenza di *w* nel mega-documento

Classificazione

- Multinomiale vs Multivariazione binomiale?
  - Multinomiale è generalmente meglio
    - Guarda i risultati figurati successivamente

Esempio NB

- Dati: 4 documenti
  - D1 (sport): Calcio Cinese
  - D2 (sport): Baseball Giapponese
  - D3 (politica): Commercio Cinese
  - D4 (politica): Esportazione Giapponese
- Classifica:
  - D5: calcio
  - D6: Giappone
- Usa
    - Aggiunge una scorrevolezza
  - Modello Multinomiale
  - Modello Binomiale a Multivariazione

Un esempio di Naïve Bayes

- C = {allergia, freddo, bene}
- e_1 = raffreddore; e_2 = tosse; e_3 = febbre
- E = {raffreddore, tosse,¬febbre}

Problema	Bene	Freddo	Allergia
$P(c_i)$	0.9	0.05	0.05
$P(\text{raffreddore}\|c_i)$	0.1	0.9	0.9
$P(\text{tosse}\|c_i)$	0.1	0.8	0.7
$P(\text{febbre}\|c_i)$	0.01	0.7	0.4

Un esempio di Naïve Bayes (cont.)

Probabilità	Bene	Freddo	Allergia
$P(c_i)$	0.9	0.05	0.05
$P(\text{raffreddore} \mid c_i)$	0.1	0.9	0.9
$P(\text{tosse} \mid c_i)$	0.1	0.8	0.7
$P(\text{febbre} \mid c_i)$	0.01	0.7	0.4

E={raffreddore,tosse,¬febbre}

P(bene | E) = (0.9)(0.1)(0.1)(0.99)/P(E)=0.0089/P(E)
P(freddo | E) = (0.05)(0.9)(0.8)(0.3)/P(E)=0.01/P(E)
P(allergia | E) = (0.05)(0.9)(0.7)(0.6)/P(E)=0.019/P(E)

La classe piu' frequente è allergia in quanto:
P(E) = 0.0089 + 0.01 + 0.019 = 0.0379
P(bene | E) = 0.23, P(freddo | E) = 0.26, P(allergia | E) = 0.50

Selezione Caratteristiche: Perchè?

- La collezione di testi ha un grande numero di caratteristiche
  - 10,000 – 1,000,000 parole uniche ...e ancora piu.
- Selezione Caratteristica:
  - è **il processo per cui un grande insieme di caratteristiche disponibili sono negate durante la classificazione**
  - Non affidabile, non bene stimato, non utile
- Puo essere fatto usando un particolare classificatore fattibile, es. riduce il tempo di training
  - Alcuni classificatori non possono gestire 100.000 caratteristiche
  - Il tempo di Training per alcuni metodi è quadratico o peggio nel numero di caratteristiche
- Puo' migliorare la generalizzazione (performance)
  - Elimina le caratteristiche di rumore + Evita l'overfitting

Selezione caratteristiche: come?

- Due idee:
  - Statistiche di test ipotesi:
    - Sappiamo che il valore di una variabile categorica è associata con il valore di un altra
    - Test Chi-Square
  - Teoria dell'Informazione:
    - quanta informazione ti dà il valore di una variabile categorica sul valore di un'altra
    - Informazione Mutuale

- Sono simili, ma χ^2 misura la confidenza in associazione, (basata sulle statistiche disponibili), mentre MI misura estensioni di associazione (assumendo conoscenza perfetta delle probabilità)

Statistiche χ^2 (CHI)

- Il Chi-Square di Pearson è spesso usato per creare un test di indipendenza.
- Un test di indipendenza è creato quando osservazioni accoppiate su due variabili, espresse in una tabella di contingenza sono indipendenti da ogni altra – per esempio, quando documenti in classi differenti differiscono nell'osservazione di una caratteristica data (es. parola).
- Es. di una tabella di contingenza:

	Termine = jaguar	*Termine ≠ jaguar*
Classe = auto	2	500
Classe ≠ auto	3	9500

Statistiche χ^2 (CHI)

- χ^2 è interessato in $(Obs - Exp)^2/Exp$ sommato su tutte le entrate di tabella: è il numero osservato cio che ti aspetti dati i marginali?
- Valori Aspettati (assumendo piena indipendenza), es. la "frequenza teorica" per una cella, data l'ipotesi di indipendenza

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N}$$

- Valore di χ^2:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Statistiche χ^2 (CHI)

$$E_{1,1} = \frac{1}{N}(O_{1,1}(O_{1,1}+O_{1,2})+O_{1,2}(O_{1,1}+O_{1,2})) =$$
$$= \frac{1}{10005}(2(2+3)+500(2+3)) = 0.25$$

	Termine = jaguar	*Termine ≠ jaguar*
Classe = auto	2 *(0.25)*	500 *(502)*
Classe ≠ auto	3 *(4.75)*	9500 *(9498)*

aspettato: *E*

osservato: *O*

- L'ipotesi nulla è rigettata con confidenza .999,
- dato che 12.9 > 10.83 (il valore per la confidenza .999)

Statistica χ^2 (CHI)

Cè una formula piu semplice per 2x2 χ^2 :

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A = #(t,c)	C = #(¬t,c)
B = #(t,¬c)	D = #(¬t, ¬c)

$$N = A + B + C + D$$

Valore per la completa indipendenza di termini e categoria?

Selezione caratteristiche attraverso Informazione Mutuale

- Nell'insieme di Training, scegli k parole che meglio discriminano (danno maggiori informazioni su) la categoria.
- La Mutuale Informazione tra una parola w e una classe c è:

$$I(w,c) = \sum_{e_w \in \{0,1\}} \sum_{e_c \in \{0,1\}} p(e_w, e_c) \log \frac{p(e_w, e_c)}{p(e_w)p(e_c)}$$

  - Per ogni parola w e ogni categoria c

Selezione di Caratteristiche attraverso MI (cont.)

- Per ogni categoria costruiamo una lista di k termini piu discriminanti.
- Per esempio (su 20 Newsgroups):
  - ***sci.electronics:*** circuiti, voltaggio, amp, terra, copia, batteria, elettronica, raffreddamento, ...
  - ***rec.autos:*** macchina, macchine, motore, ford, venditore, mustang, olio, collisione, auto, gomme, toyota, ...
- Greedy: non conta per le correlazioni tra termini
- Perchè?

Selezione Caratteristiche

- Informazione Mutuale
  - Chiara interpretazione di informazione teorica
  - Puo' selezionare rari termini non informativi
- Chi-square
  - Fondazione Statistica
  - Puo' selezionare piu frequentemente termini informativi che non sono molto utili per la classificazione

- Usare solo il termine piu comune?
  - Nessuna particolare fondazione
  - In pratica, questo è circa il 90% buono

Selezione Caratteristiche per NB

- Nella caratteristicha generale la selezione è *necessaria* per NB binomiali.
- Altrimenti soffrirai per rumore, multi-conteggio

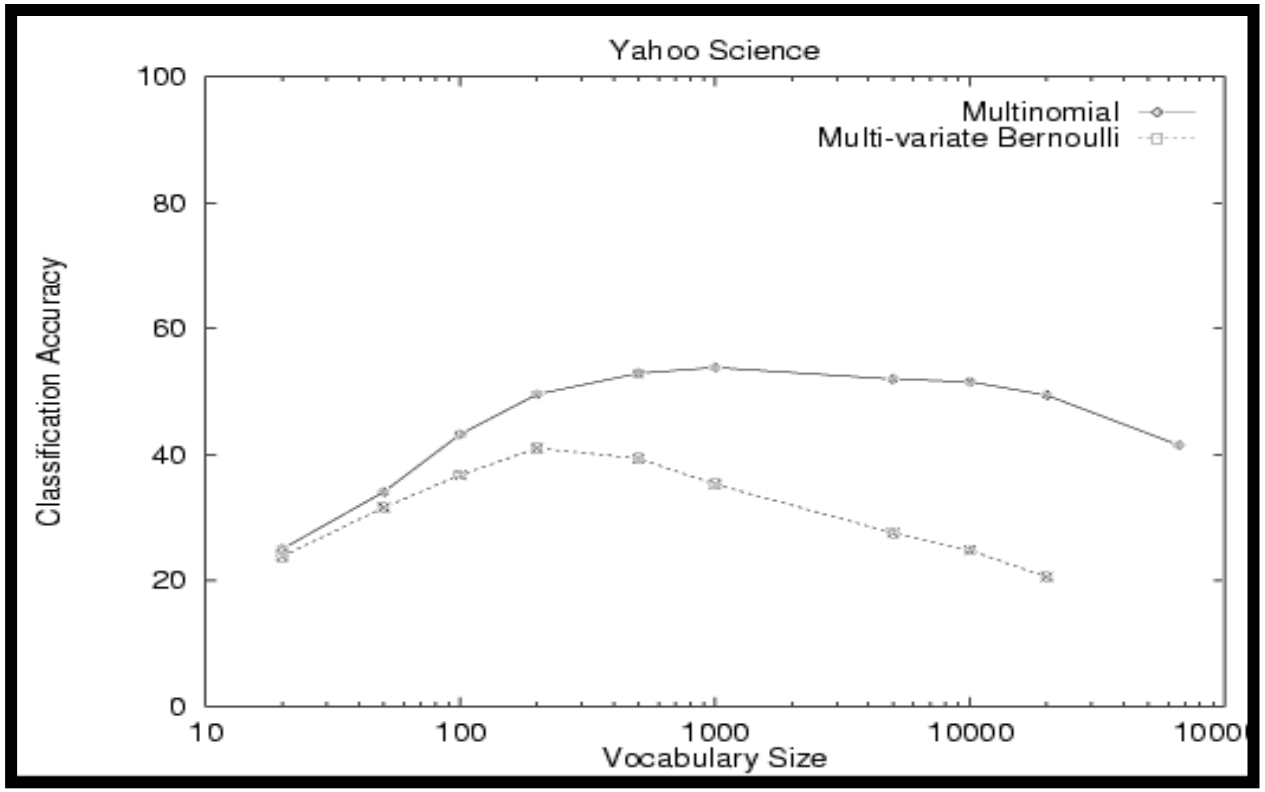
- "Selezione Caratteristica" davvero significa qualcosa di diverso per NB multinomiali. Significa troncaggio di dizionario
  - Il modello NB multinomiale ha solo 1 caratteristica
- questa "selezione di caratteristiche" normalmente non è necessaria per NB nultinomiali, ma puo aiutare una frazione con quantità che sono male stimate

Valutare la Categorizzazione

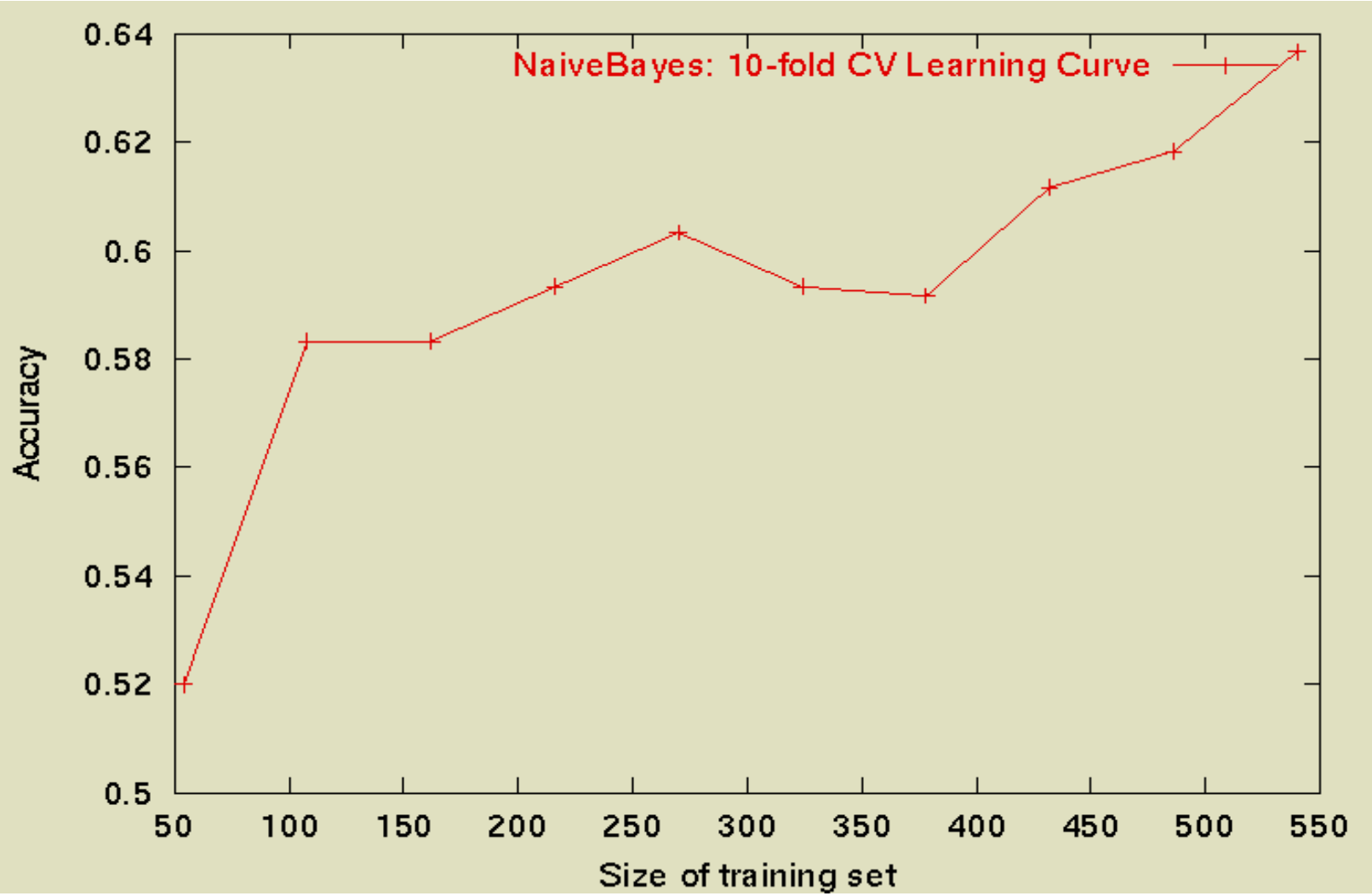
- La valutazione deve essere fatta su dati di test che sono indipendenti dai dati di training (generalmente un insieme disgiunto di istanze).
- *Classificazione di accuratezza*: c/n dove n è il numero totale di istanze di test e c è il numero di istanze di test correttamente classificate dal sistema.
- I risultati possono variare in base agli errori di campionamento a causa di differenti insiemi di training e di test.
- Risultati medi su training multipli e insiemi di test (divisioni dei dati globali) per il miglior risultato.

Esempio: AutoYahoo!

- Classifica 13,589 pagine Yahoo! nel sottoalbero "Scienze" in 95 differenti argomenti (profondità gerarchica 2)



Curva di Apprendimento di Esempio (Dati Scientifici di Yahoo): ne servono di piu!



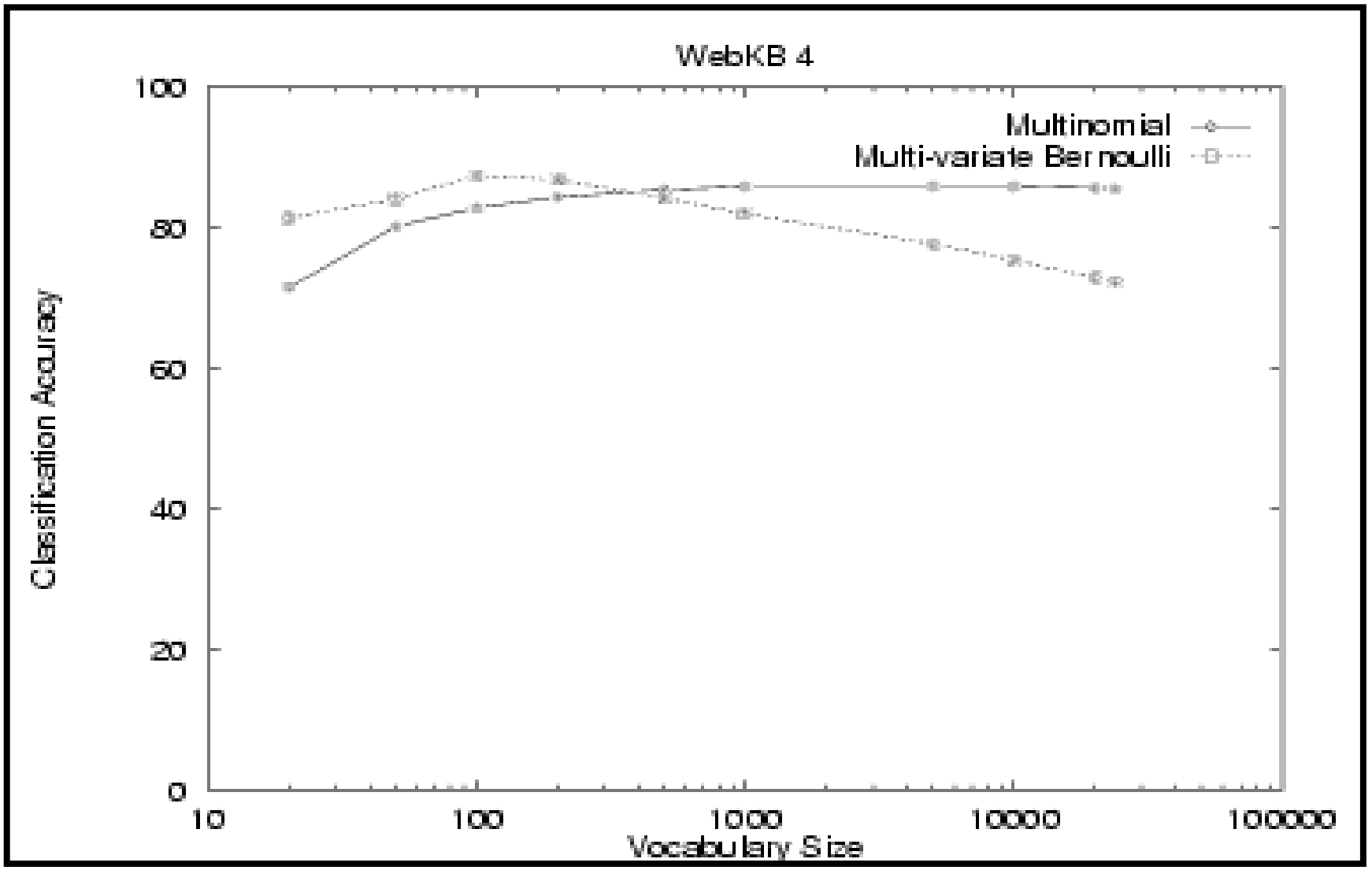
Espeerimento WebKB

- Classifica le pagine web dal dipartimento CS in:
  - studente, facoltà, corso, progetto
- Apprende su ~5,000 pagine web etichettate manualmente
  - Cornell, Washington, U.Texas, Wisconsin
- Scansiona e classifica un nuovo sito (CMU)

- Risultati:

	Studente	Facoltà	Persona	Progetto	Corso	Dipartimento
Estratti	180	66	246	99	28	1
Corretti	130	28	194	72	25	1
Accuratezza	72%	42%	79%	73%	89%	100%

Comparazione Modello NB



Faculty

associate	0.00417
chair	0.00303
member	0.00288
ph	0.00287
director	0.00282
fax	0.00279
journal	0.00271
recent	0.00260
received	0.00258
award	0.00250

Students

resume	0.00516
advisor	0.00456
student	0.00387
working	0.00361
stuff	0.00359
links	0.00355
homepage	0.00345
interests	0.00332
personal	0.00332
favorite	0.00310

Courses

homework	0.00413
syllabus	0.00399
assignments	0.00388
exam	0.00385
grading	0.00381
midterm	0.00374
pm	0.00371
instructor	0.00370
due	0.00364
final	0.00355

Departments

departmental	0.01246
colloquia	0.01076
epartment	0.01045
seminars	0.00997
schedules	0.00879
webmaster	0.00879
events	0.00826
facilities	0.00807
eople	0.00772
postgraduate	0.00764

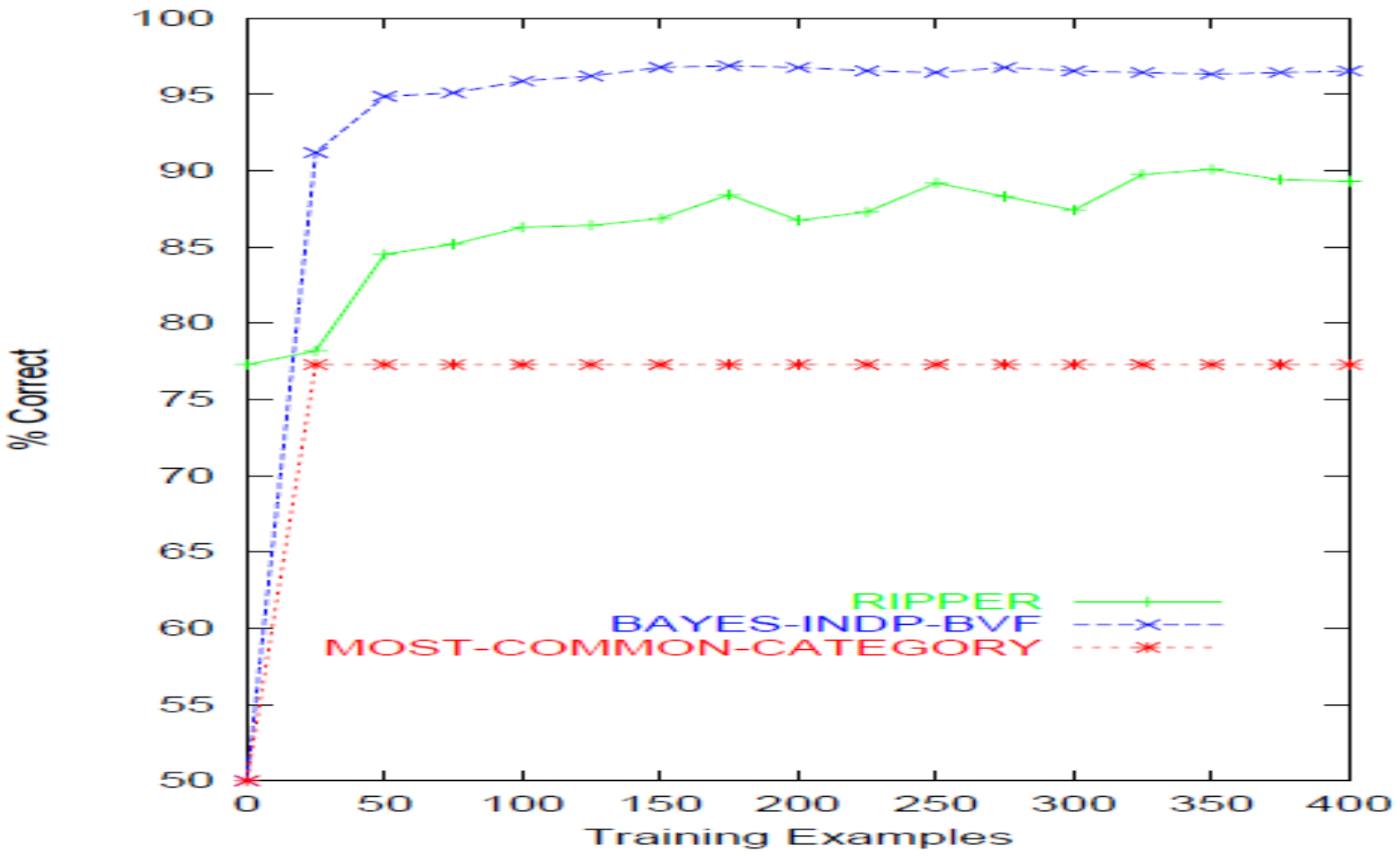
Research Projects

investigators	0.00256
group	0.00250
members	0.00242
researchers	0.00241
laboratory	0.00238
develop	0.00201
related	0.00200
arpa	0.00187
affiliated	0.00184
project	0.00183

Others

type	0.00164
jan	0.00148
enter	0.00145
random	0.00142
program	0.00136
net	0.00128
time	0.00128
format	0.00124
access	0.00117
begin	0.00116

Naïve Bayes sullo Spam E-Mail



SpamAssassin

- Naïve Bayes ha trovato una casa per il filtraggio dello spam
  - *A Plan for Spam* di Graham
    - E la sua prole mutante...
  - I classificatori come Naive Bayes con stima di strani parametri
  - Largamente usato nei filtri spam
    - Il Naive Bayes classico è superiore quando usato bene
    - In accordo con David D. Lewis
- Molti filtri email usano i classificatori NB
  - Ma anche molti altri elementi: liste buchi neri, ecc.

Violazioni di Assunzioni NB

- Indipendenza Condizionale
- Indipendenza Posizionale
- Esempi?

Probabilità Posteriori di Naïve Bayes

- I risultati della classificazione di naïve Bayes (la classe con massima probabilità posteriore) sono generalmente abbastanza accurati.
- Ad ogni modo, a causa dell'inadeguatezza dell' assunzione condizionale di indipendenza, le attuali stime numeriche posteriori-probabilità non lo sono
  - Le probabilità in output sono generalmente molto vicine a 0 o 1.

quando deve lavorare Naive Bayes?

- Alcune volte NB si comporta bene anche se l'Assunzione Condizionale di Indipendenza è altamente violata.
- La Classificazione riguarda la predizione della corretta etichetta della classe e NON riguarda la stima accurata delle probabilità.

Assume due classi c_1 e c_2. Un nuovo caso A arriva.

NB classificherà A su c_1 se:

$$P(A, c_1) > P(A, c_2)$$

	$P(A,c_1)$	$P(A,c_2)$	Classe di A
Attuale Probabilità	0.1	0.01	c_1
Probabilità Stimata da NB	0.08	0.07	c_1

Dietro il grande errore nella stima delle probabilità la classificazione è ancora corretta.

Stima corretta $\Rightarrow$ predizione accurata

ma NON

predizione accurata $\Rightarrow$ Stima corretta

Naive Bayes è non troppo Naive

- Naïve Bayes: Primo e Secondo posto nella competizione KDD-COPPA 97 su 16 (poi) algoritmi dello stato dell'arte

  Obiettivo: modello di predizione per responso diretto delle mail per industrie di servizi finanziari
  I destinatari delle mail rispondono attualmente alla pubblicità– 750,000 record.

- Robusto verso Caratteristiche Irrilevanti

  Le Caratteristiche Irrilevanti cancellano ogni altra senza affliggere i risultati
  Invece gli Alberi di Decisione possono altamente soffrire di cio.

- Molto buono in domini con molte caratteristiche equalmente importanti

  Gli Alberi di Decisione soffrono di *frammentazione* in questi casi – specialmente se pochi dati

- Un ottima linea di base per la classificazione del testo (ma non la migliore)!
- Ottima se l'Assunzione di Indipendenza rimane: Se si assume che l'indipendenza è corretta, allora il problema è il Classificatore Ottimo di Bayes
- Molto Veloce:Apprendere con un passo sui dati; testing lineare nei numeri di attributi, e nella grandezza della collezione dei documenti
- Pochi Requisiti di Spazio

Risorse

- IIR 13
- Fabrizio Sebastiani. **Machine Learning in Automated Text Categorization.** *ACM Computing Surveys*, 34(1):1-47, 2002. (http://faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS01/ACMCS01.pdf)
- Andrew McCallum and Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41-48.
- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.
  - Clear simple explanation
- Yiming Yang & Xin Liu, A re-examination of text categorization methods. *Proceedings of SIGIR*, 1999.

Sommario

- Un tipo di apprendimento di base è quello probabilistico dove apprendere significa
  - Descrivere il problema mediante un modello generativo che mette in relazione le variabili in input (e.g. sintomi) e quelle in output (e.g. diagnosi)
  - Determinare i corretto parametri del problema (i.e. le distribuzioni analitiche o la stima delle probabilità discrete)
- Un esempio: classificazione NB (caso discreto)
- Due sono i modelli piu' usati:
  - Binomiale Multivariazione (o Bernoulli) NB
  - NB Multinomiale

Sommario (2)

- Nella stima dei parametri in NB un ruolo centrale è svolto dalle tecniche di *smoothing*: a parità di modello infatti stimatori errati producono risultati insoddisfacenti
- La classificazione mediante NB è preferibile per la relativa robustezza nei casi in cui l'efficienza è fondamentale
- E' invece usato come baseline in molte sperimentazioni