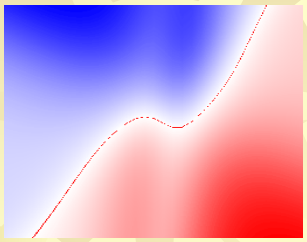
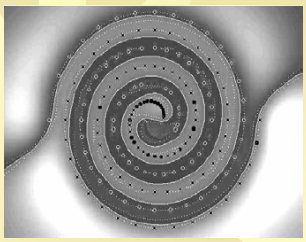
Università Degli Studi Roma Tre



Be MoRe

BEST MODEL RETRIEVAL



Una nuova metodologia di classificazione di testi

Claudio Biancalana
Alessandro Micarelli

Sommario

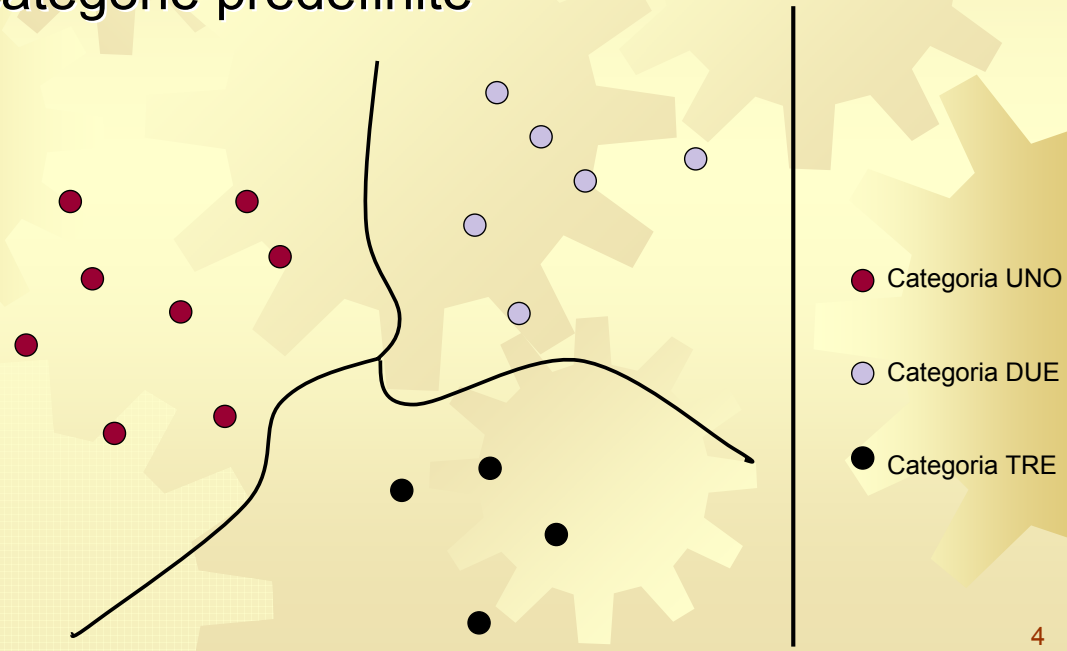
- Definizione del problema
- Be More: la metodologia di classificazione
  - NLP
    - WSD
    - Edit Distance
  - Indicizzazione Log Noise
  - OnLine Hyperplane
- Risultati sperimentali
- Obiettivi raggiunti

Categorizzazione

- Input:
  - Una descrizione di una istanza, $x \in X$, dove X è l'*istanza linguaggio* o *spazio dell'istanza*.
  - Un numero fissato di categorie: $C=\{c_1, c_2,\ldots c_n\}$
- Output:
  - La categoria di x: $c(x) \in C$, dove $c(x)$ è una funzione di catgorizzazione che ha come dominio X e come codominio C.

Text Categorization

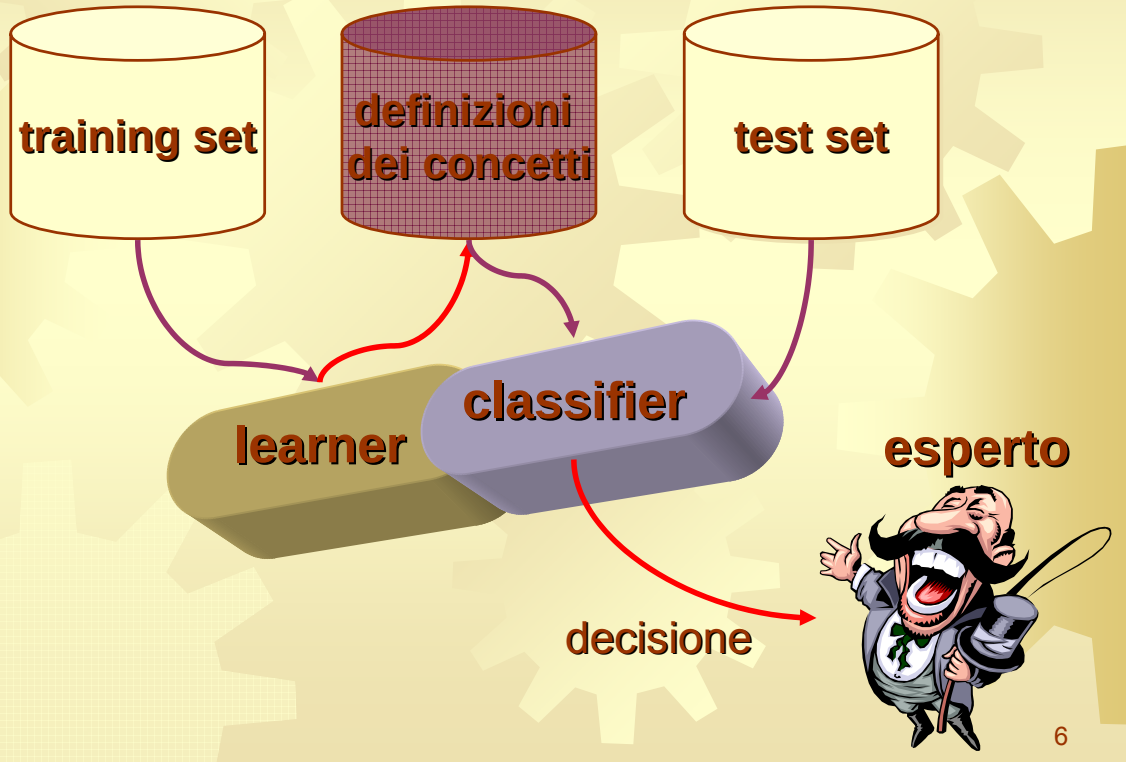
- Text categorization (text classfication): assegnare documenti testuali ad una o più categorie predefinite



Hard Problems

- Spazio delle caratteristiche sparso
- Piccolo insieme di training
- Rumore
- Efficienza computazionale
- Apprendimento complesso

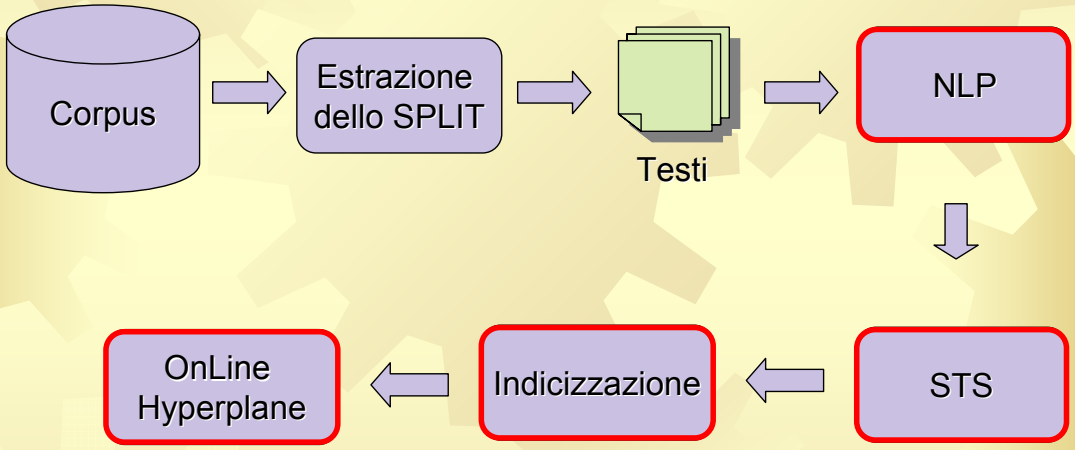
Fasi dell'Apprendimento



Be MoRe

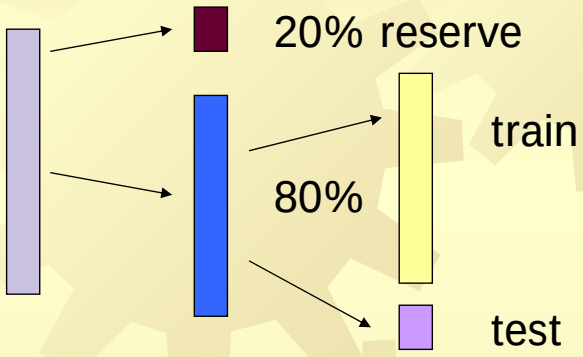
BEST MODEL RETRIEVAL

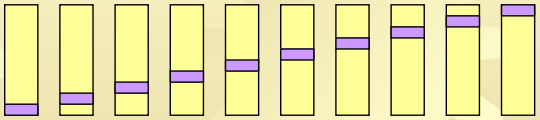
Schema concettuale



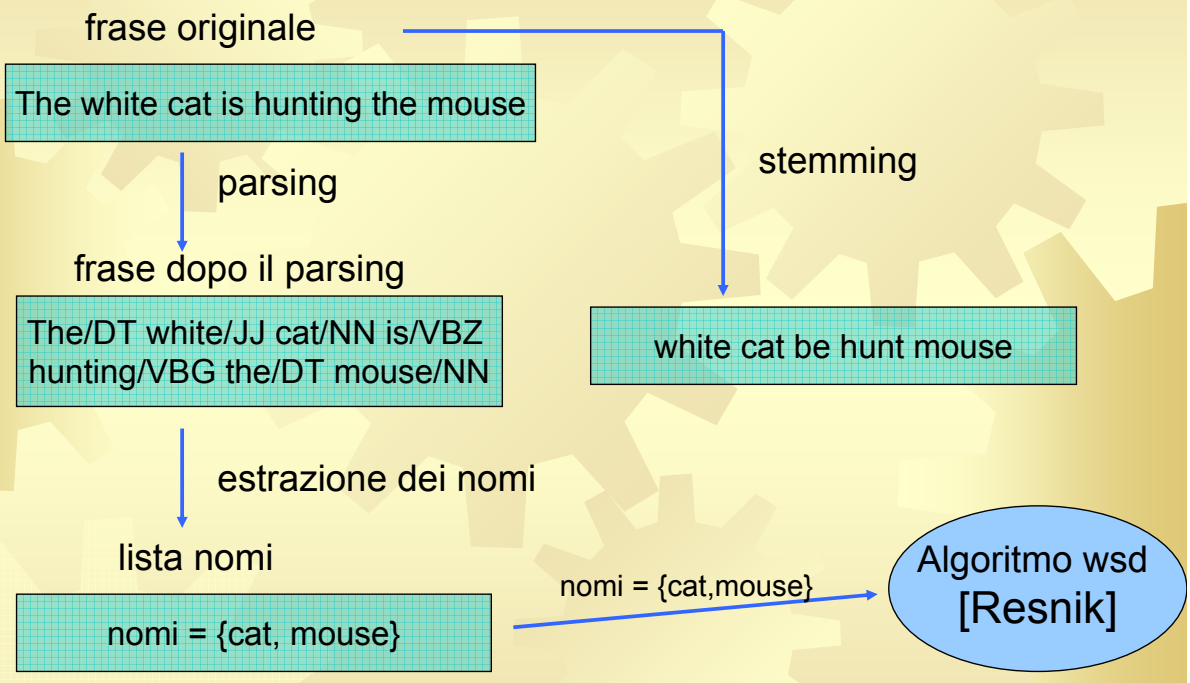
Ricerca del tuning migliore per la classificazione con stima Leave-one-out

Leave one out estimatior





NLP Natural Language Processing



Disambiguazione dei nomi

Il sistema, mediante il dizionario WordNet, assegna ad ogni significato un codice numerico univoco.

Dopo aver eseguito gli algoritmi di word sense disambiguation, ad ogni nome della frase viene sostituito il codice del significato che è risultato essere migliore.

Frase originale:

The white cat is hunting the mouse

Frase dopo lo stemming:

white cat be hunt mouse

Frase con codici:

white 1788952 be hunt 1993014

Frase originale:

The white computer is detecting a mouse

Frase dopo lo stemming:

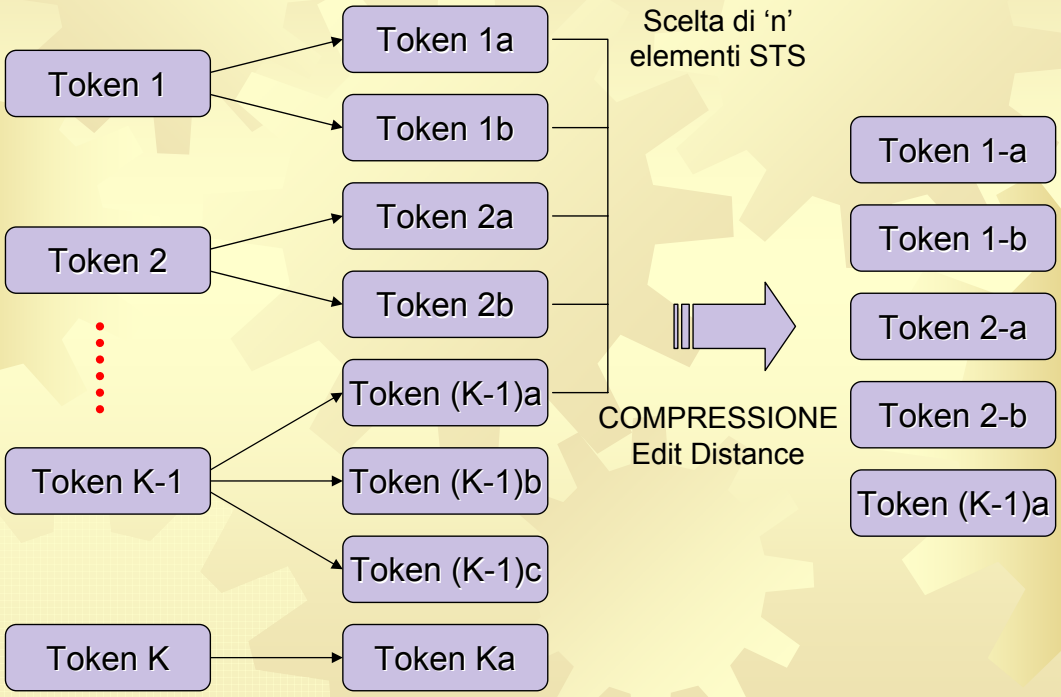
white computer be detect mouse

Frase con codici:

white 7846548 be detect 85657

Espansione e compressione

le innovative fasi del preprocessing



STS Supervised Term Selection

(prima fase della compressione)

Selezione supervisionata dei termini.

$$\xi = \frac{|T| - |T'|}{|T|}$$

Reduction Factor

TEF Term evaluation function (low-noise):

$$1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log(ndocs)} \quad \text{con} \quad p_{ij} = \frac{tf_{ij}}{df_i}$$

tf_{ij}: Frequenza del termine i sul documento j

df_i: Numero dei documenti contenenti il termine i

Obiettivo: MINIMIZZARE IL RUMORE DI INFORMAZIONE

Edit Distance

(seconda fase della compressione)

Edit Distance: indice di somiglianza moforlogica

A C T G T **A C T T T G T A**

$$C_{i,j} = \begin{cases} C_{i-1,j-1} & \text{se } i=0 \text{ o } j=0 \\ C_{i-1,j-1} + 1 & \text{se } a_i = b_j \\ 1 + \text{Max}(C_{i-1,j}, C_{i,j-1}) & \text{altrimenti} \end{cases}$$

Edit Distance

(seconda fase della compressione)

A C T G T **A C T T T G T A**

		A	C	T	T	T	G	T	A
	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1
C	0	1	2	2	2	2	2	2	2
T	0	1	2	3	3	3	3	3	3
G	0	1	2	3	3	3	4	4	4
T	0	1	2	3	4	4	4	5	5

Indicizzazione Log-Noise

Term Weighting $A = [a_{ij}]$

$$a_{ij} \equiv L(i,j) \times G(i)$$ Termine i Documento j

$$L(i,j) \equiv [\log(tf_{ij} + 1)]$$ Indice locale LOG

$$G(i) \equiv 1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log(ndocs)}$$ Indice globale NOISE

$p_{ij} = \frac{tf_{ij}}{df_i}$ → Frequenza del termine i sul documento j

→ Numero dei documenti contenenti il termine i

Dimensionality reduction

- La dimensione dello spazio dei termini può costituire un problema perché:
  - gli algoritmi di learning non scalano facilmente su grandi valori della dimensione
  - se la dimensione è alta spesso si verificano fenomeni di overfitting
- Abbiamo due scelte
  - Riduzione locale (un insieme di termini diverso per ciascuna categoria)
  - Riduzione globale (il set di termini è valido per qualunque categoria)
- La riduzione può essere
  - per selezione
  - per estrazione

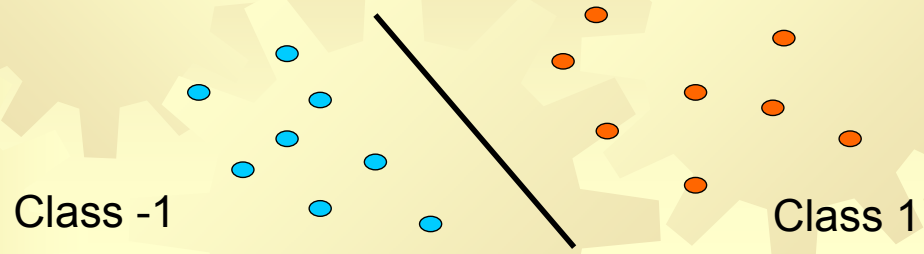
Feature selection

Function	Denoted by	Mathematical form
Document frequency	$\#(t_k, c_i)$	$P(t_k \mid c_i)$
DIA association factor	$z(t_k, c_i)$	$P(c_i \mid t_k)$
Information gain	$IG(t_k, c_i)$	$\sum_{c \in \{c_i, \overline{c}_i\}} \sum_{t \in \{t_k, \overline{t}_k\}} P(t, c) \cdot \log \frac{P(t, c)}{P(t) \cdot P(c)}$
Mutual information	$MI(t_k, c_i)$	$\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$
Chi-square	$\chi^2(t_k, c_i)$	$\frac{\lvert Tr\rvert \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]^2}{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}$
NGL coefficient	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert Tr\rvert} \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$\log \frac{P(t_k \mid c_i) + d}{P(\overline{t}_k \mid \overline{c}_i) + d}$
Odds Ratio	$OR(t_k, c_i)$	$\frac{P(t_k \mid c_i) \cdot (1 - P(t_k \mid \overline{c}_i))}{(1 - P(t_k \mid c_i)) \cdot P(t_k \mid \overline{c}_i)}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)$

ESEMPIO: Noise '3 features'

Category	1st word	2nd word	3rd word	BEP
Earn	vs+	cts+	loss+	93.5%
Acq	shares+	vs-	Inc+	76.3%
Money-fx	dollar+	vs-	exchange+	53.8%
Grain	wheat+	tonnes+	grain+	77.8%
Crude	oil+	bpd+	OPEC+	73.2%
Trade	trade+	vs-	cts-	67.1%
Interest	rates+	rate+	vs-	57.0%
Ship	ships+	vs-	strike+	64.1%
Wheat	wheat+	tonnes+	WHEAT+	87.8%
Corn	corn+	tonnes+	vs-	70.3%

OnLine Hyperplane

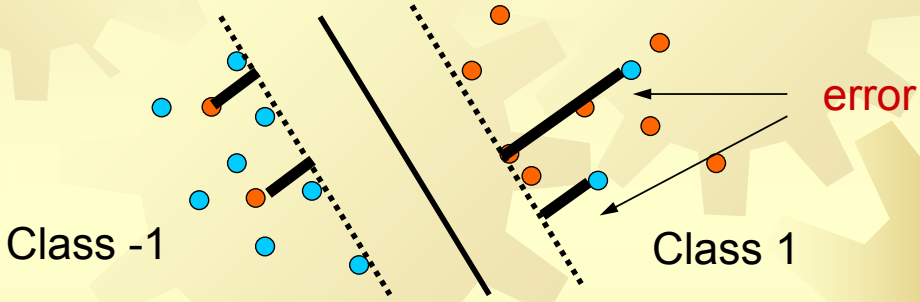


- Obiettivo: Trovare l'iperpiano separatore ottimo dell'insieme di training. Come formalizzare?
  - In due dimensioni, l'equazione della linea è data da: $w_1x + w_2y = b$
  - In 'n' dimensioni :

$$\sum_{i=0}^{n} wi.xi = b \qquad \vec{w}.\vec{x} = b$$

Slack variables

- Se l'iperpiano separatore non esiste?



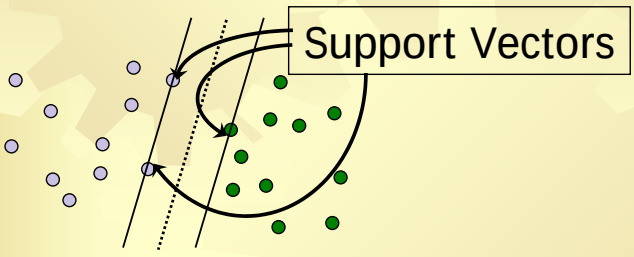
- Cosa si può fare se i dati non sono linearmente separabili per la presenza di rumore?
- *Slack variables*
- Consentono la classificazione non corretta di alcuni punti, tenendo conto del rumore nei dati

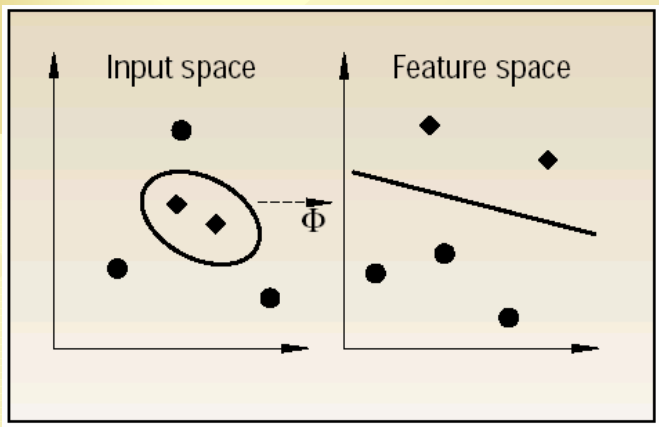
$$\vec{w}.\vec{x}_i \geq b+1 \rightarrow \vec{w}.\vec{x}_i + \xi_i \geq b+1$$

OnLine Hyperplane

- *Ricerca incrementale della soluzione ottima*
- *Elementi positivi e negativi rappresentati con pesi diversi*
- *Loss function a basso costo computazionale*
- *Convergenza garantita dall'**estensione del teorema di Novikoff***

Kernelizing





Ereditando alcune caratteristiche matematiche delle SVM, il sistema si avvale di un modulo di kernel per risolvere il problema di separabilità non lineare.

Kernel supportati:

- Lineare
- Polinomiale
- RBF

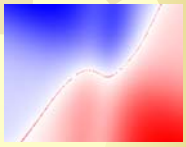
Scelta del kernel migliore attraverso model selection

GET-BOUND

Macchine SVM non lineari

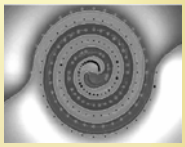
- L'uso della funzione kernel consente di calcolare l'iperpiano di separazione senza bisogno di effettuare esplicitamente il *mapping* in F
- Esempio:

$$k(x,x') = e^{-\|x-x'\|^2}$$

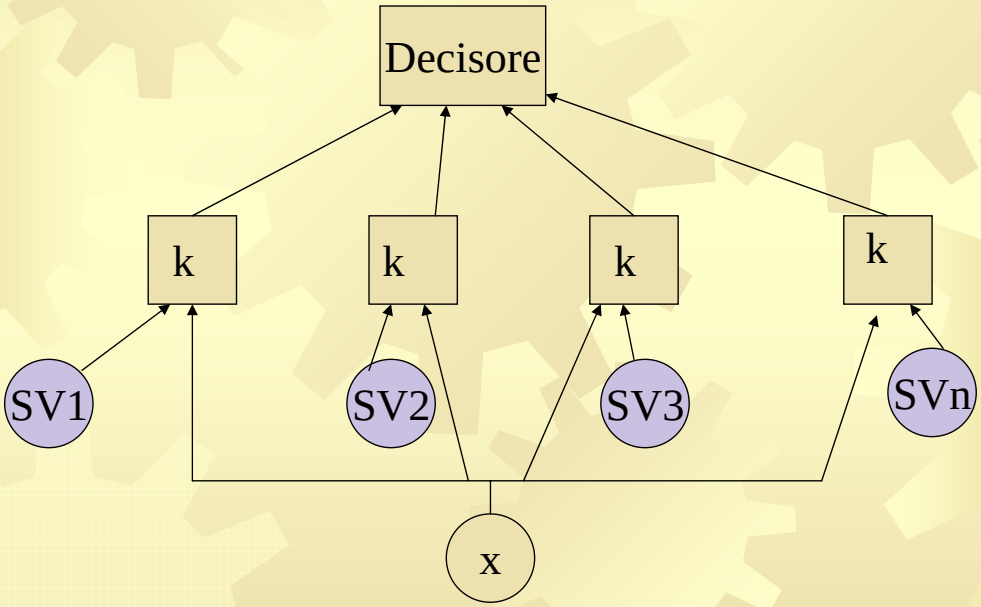


- Altre funzioni kernel:
  - Gaussian RBF $e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$
  - Polinomio di grado d $(x \cdot y + 1)^d$
  - Tangente iperbolica $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

$$z = x \cdot x' - \theta$$



Architettura multi-label



Valutazione del classificatore

- La valutazione è sperimentale perché il problema non ha una specifica formale che consenta un altro tipo di valutazione
- La valutazione deve essere effettuata su dati di test indipendenti dai dati di training (solitamente insiemi disgiunti di istanze)
- I risultati possono variare in base in base all'uso di diversi di training e testing set

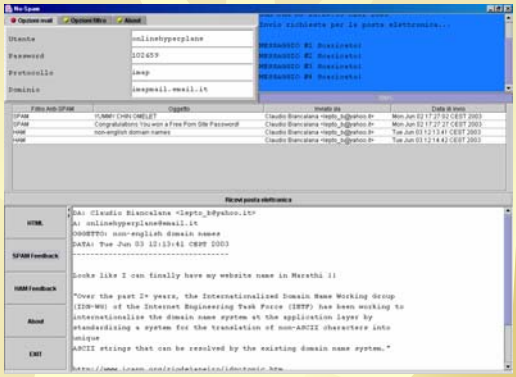
Misure di performance

- Tavola di contingenza

	Truth: Yes	**Truth: No**
System: Yes	*a*	*b*
System: No	*c*	*d*

- Misure di performance per la classificazione binaria
  - error rate = (b+c)/n
  - accuracy = 1 - error rate
  - precision (P) = a/(a+b)
  - recall (R) = a/(a+c)
  - break-even = (P+R)/2
  - F1-measure = 2PR/(P+R)

Benchmark



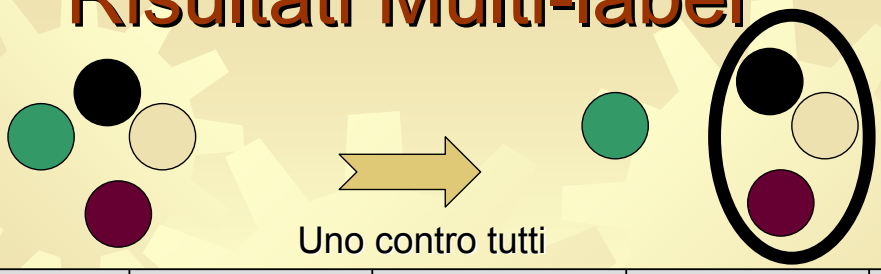
Be MoRe:

- Reuters 21578
  - Split ModAptè[10]
  - Split ModAptè[115]
- RCV1 (800000 docs)
- WebKB (pagine HTML)
- Ohsumed (referti medici)
- LingSpam (Anti-spam)

OnLine Hyperplane:

- MNIST (OCR)
- Copchrom (Corredo cromosomico)

Risultati Multi-label



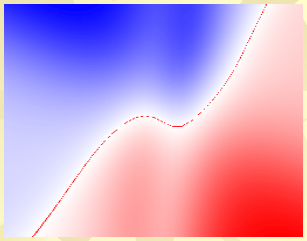
Dominio / Categorizzatore	Reuters ModAptè[10] Micro F1Measure	Reuters ModAptè[115] Micro F1Measure	Ohsumed Micro F1Measure	WebKB BEP
Be MoRe	93%	87.5%	71.6%	89.4%
SVM light (TF-IDF + MI)	92% (Dumais et al.)	86.7% (Joachims)	67.7%	90.3% (Joachims)
IB (TF-IDF)	92.6% (Bekkerman et al.)	…	…	…
Naive Bayes	81% (Dumais et al.)	72.3% (Joachims)	62.4%	82%
kNN (TF-IDF + MI)	…	82.6% (Joachims)	63.4%	80.5%

Complessità computazionale OH vs SVM

- Training time: SVM vs. OH (# features = 100):

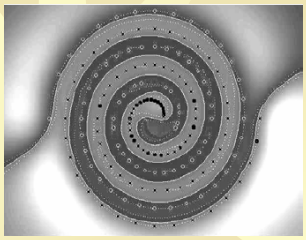
# documents	Training Time per OH	Training Time per SVM
9603	5	8
19206	15	25
28809	27	60
38412	32	120
48015	40	340
57618	50	410
67221	65	498
76824	78	600
86427	100	630

Università Degli Studi Roma Tre



Be MoRe

BEST MODEL RETRIEVAL



Una nuova metodologia di classificazione di testi

Claudio Biancalana
Alessandro Micarelli