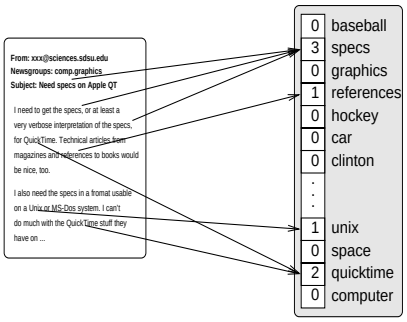
CS630 Representing and Accessing Digital Information

Text Classification: Support Vector Machines

Thorsten Joachims
Cornell University

Representing Text as Attribute Vectors



Attributes: Words (Word-Stems)

Values: Occurrence-Frequencies

==> The ordering of words is ignored!

Generative vs. Discriminative Training

Process:

- Generator: Generates descriptions $\vec{x}$ according to distribution $P(\vec{x})$.
- Teacher: Assigns a value y to each description $\vec{x}$ based on $P(y|\vec{x})$.

=> Training examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(\vec{x}, y) \quad \vec{x}_i \in \Re^N \quad y_i \in \{1, -1\}$

Generative Training

- make assumptions about the parametric form of $P(\vec{x}, y)$.
- estimate the parameters of $P(\vec{x}, y)$ from the training data
- derive optimal classifier using Bayes' rule
- example: naive Bayes

Discriminative Training

- make assumptions about the set H of classifiers
- estimate error of classifiers in H from the training data
- select classifier with lowest error rate
- example: SVM, decision tree

Principle: Empirical Risk Minimization (ERM)

Learning Principle:

Find the decision rule $h^\circ \in H$ for which the training error is minimal:

$$h^\circ = \arg min_{h \in H}\{Err_S(h)\}$$

Training Error:

$$Err_S(h) = \frac{1}{n}\sum_{i=1}^{n}\Delta(y_i \neq h(\vec{x}_i))$$

==> Number of misclassifications on training examples.

Central Problem of Statistical Learning Theory:

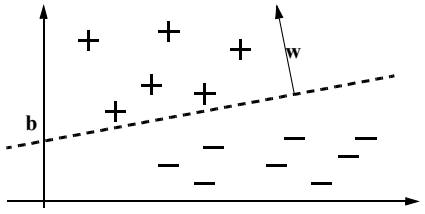
When does a low training error lead to a low generalization error? (i.e. what is the probability of error after n training examples)

Linear Classifiers

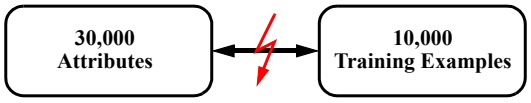
Rules of the Form: weight vector $\vec{w}$, threshold b

$$h(\vec{x}) = sign\left[\sum_{i=1}^{N}\vec{w}_i\vec{x}_i + b\right] = \begin{cases} 1 & if \sum_{i=1}^{N}\vec{w}_i\vec{x}_i + b > 0 \\ -1 & else \end{cases}$$

Geometric Interpretation (Hyperplane):



Paradoxon of Text Classification

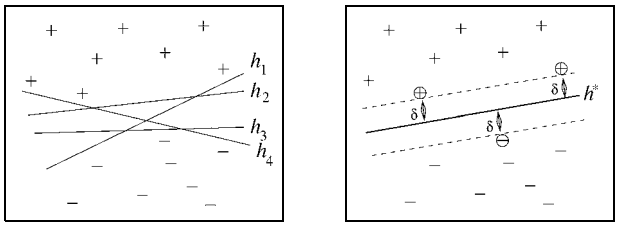


Good News: SVMs can overcome this problem!

Bad News: This does not hold for all high-dimensional problems!

Optimal Hyperplane (SVM Type 1)

Assumption: The training examples are linearly separable.



Support Vectors: Examples with minimal distance (circled).

SVM Primal Optimization Problem

Training Examples: $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \quad \vec{x}_i \in \Re^N \quad y_i \in \{1, -1\}$

Training:

- minimize $P(\vec{w}, b) = \frac{1}{2}\vec{w} \cdot \vec{w}$ over $\vec{w}$ and b

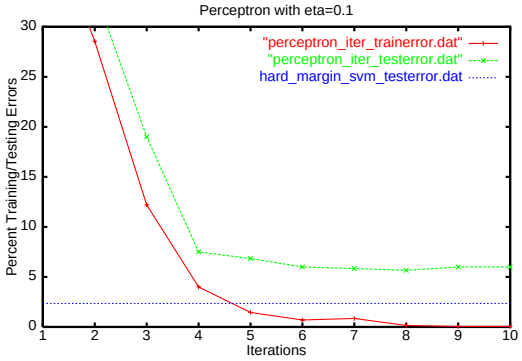
  subject to fulfilling the constraints $\begin{bmatrix} y_1[\vec{w} \cdot \vec{x}_1 + b] \geq 1 \\ \ldots \\ y_n[\vec{w} \cdot \vec{x}_n + b] \geq 1 \end{bmatrix}$

- typically single solution (i. e. $(\vec{w}, b)$ is unique)

Prediction Rule:

$$h(\vec{x}) = sign\left[\sum_{i=1}^{N} w_i x_i + b\right] = sign[\vec{w} \cdot \vec{x} + b]$$

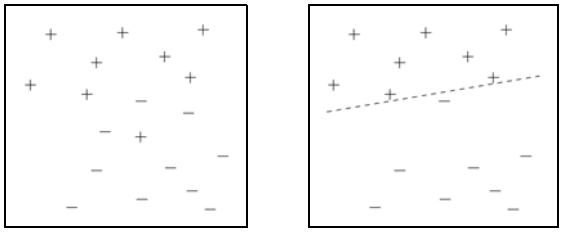
Example: Optimal Hyperplane vs. Perceptron



Train on 1000 pos / 1000 neg examples for " acq" (Reuters-21578).

Non-Separable Training Samples

- For some training samples there is no separating hyperplane!
- Complete separation is suboptimal for many training samples!



=> minimize trade-off between margin and training error.

Soft-Margin Separation

Idea: Maximize margin and minimize training error simultanously.

Hard Margin:

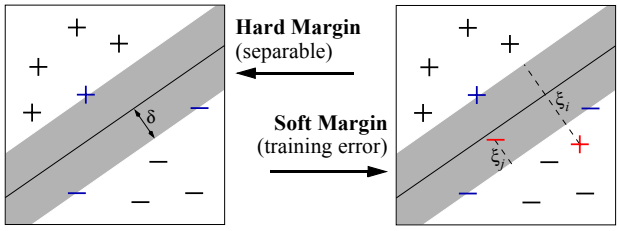
minimize $P(\vec{w}, b) = \frac{1}{2}\vec{w} \cdot \vec{w}$

s. t. $y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1$

Soft Margin:

minimize $P(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w} \cdot \vec{w} + C\sum_{i=1}^{n} \xi_i$

s. t. $y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i$ *and* $\xi_i \geq 0$

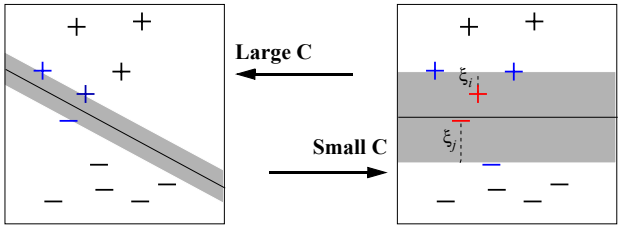


Controlling Soft-Margin Separation

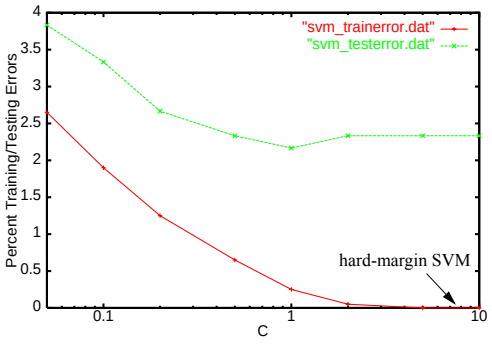
Soft Margin: minimize $P(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w} \cdot \vec{w} + C\sum_{i=1}^{n} \xi_i$

s. t. $y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i$ *and* $\xi_i \geq 0$

- $\sum \xi_i$ is an upper bound on the number of training errors.
- C is a parameter that controls trade-off between margin and error.



Example Reuters " acq" : Varying C



Observation: Typically no local optima, but not necessarily...

Solution as Linear Combination

Primal OP: minimize $P(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w} \cdot \vec{w} + C\sum_{i=1}^{n} \xi_i$

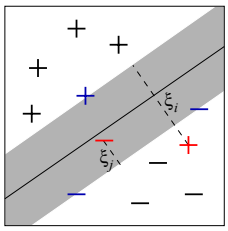
s. t. $y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i$ *and* $\xi_i \geq 0$

Lemma: The solution w° can always be written as a linear combination

$$\vec{w}^\circ = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \qquad 0 \leq \alpha_i \leq C$$

of the training data.

- one factor α_i for each training example
  - " influence" of single training example limited by C
  - $0 < \alpha_i < C$ <=> SV with $\xi_i = 0$
  - $\alpha_i = C$ <=> SV with $\xi_i > 0$
  - $\alpha_i = 0$ else
- SVM-light outputs α_i with option " -a"



Summary
What is a (linear) SVM?

Given:

- Training examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ $\vec{x}_i \in \Re^N$ $y_i \in \{1, -1\}$
- Parameter C for trading-off training error and margin size (often C=1 is a good choice for normalized document vectors)

Training:

- Finds hyperplane.
- The hyperplane has maximum margin with minimal training error (upper bound $\sum \xi_i$) given C.
- The result of training are $\alpha_1, \ldots, \alpha_n$. They determine $(\vec{w}, b)$.

Classification: For new example $h(\vec{x}) = sign\left(\sum_{x_i \in SV} \alpha_i y_i \vec{x}_i \cdot \vec{x} + b\right)$

Leave-One-Out

Training set: $(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), \ldots, (\vec{x}_n, y_n)$

Approach: Repeatedly leave one example out for testing.

train on	test on
$(\vec{x}_2, y_2), (\vec{x}_3, y_3), (\vec{x}_4, y_4), \ldots, (\vec{x}_n, y_n)$	$(\vec{x}_1, y_1)$
$(\vec{x}_1, y_1), (\vec{x}_3, y_3), (\vec{x}_4, y_4), \ldots, (\vec{x}_n, y_n)$	$(\vec{x}_2, y_2)$
$(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_4, y_4), \ldots, (\vec{x}_n, y_n)$	$(\vec{x}_3, y_3)$
...	...
$(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), \ldots, (\vec{x}_{n-1}, y_{n-1})$	$(\vec{x}_n, y_n)$

=> **Error estimate:**

$$Err_{loo}(h) = \frac{1}{n}\sum_{i=1}^{n} |h_i(\vec{x}_i) = y_i|$$

Question: Is there a connection between margin and the estimate?

Necessary Cond. for Leave-One-Out Error of SVM

Lemma: SVM $[h_i(\vec{x}_i) \neq y_i]$ ♥ $[2\alpha_i R^2 + \xi_i \geq 1]$ [Joachims, 2000] [Jaakkola & Haussler, 1999] [Vapnik & Chapelle, 2000]

Input:

- α_i dual variable of example i
- ξ_i slack variable of example i
- $\|\vec{x}\| \leq R$ bound on length

⋁

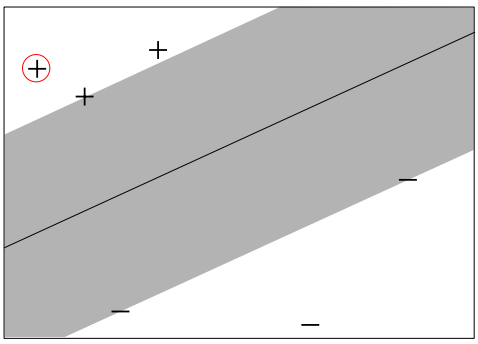
Available after training SVM on the full training data

Example:

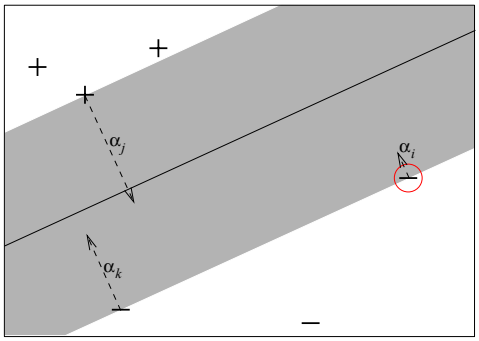
$\rho\alpha_i R^2 + \xi_i$	leave-one-out error
0.0	OK
0.7	OK
3.5	ERROR
0.1	OK
1.3	OK
0.0	OK
0.0	OK
...	...

Case 1: Example is no SV

$(\alpha_i = 0)$ ♥ $(\xi_i = 0)$ ♥ $(2\alpha_i R^2 + \xi_i < 1)$ ♥ *no leave– one– out error*

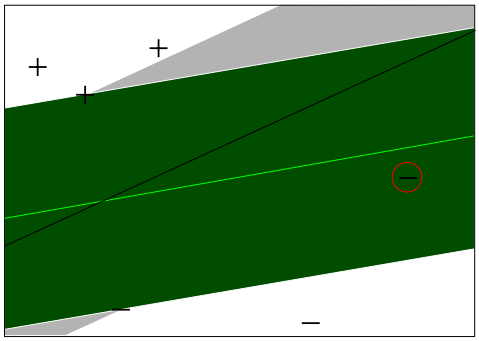


Case 2: Example is SV with Low Influence



$$\left\{ \alpha_i < \frac{0.5}{R^2} < C \right\} \heartsuit \ (\xi_i = 0) \heartsuit \ (2\alpha_i R^2 + \xi_i < 1) \heartsuit \ \textit{no leave-one-out error}$$

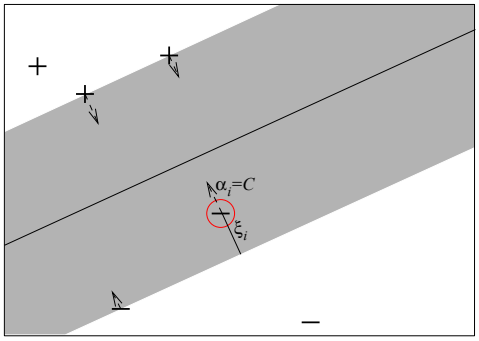
Case 2: Example is SV with Low Influence



$$\left\{ \alpha_i < \frac{0.5}{R^2} < C \right\} \heartsuit \ (\xi_i = 0) \heartsuit \ (2\alpha_i R^2 + \xi_i < 1) \heartsuit \ \textit{no leave-one-out error}$$

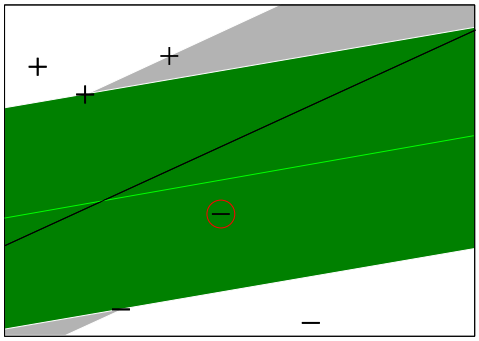
Case 3: Example has Small Training Error

$$(\alpha_i = C) \wedge (\xi_i < 1 - 2CR^2) \heartsuit \ (2\alpha_i R^2 + \xi_i < 1) \heartsuit \ \textit{no leave-one-out error}$$



Case 3: Example has Small Training Error

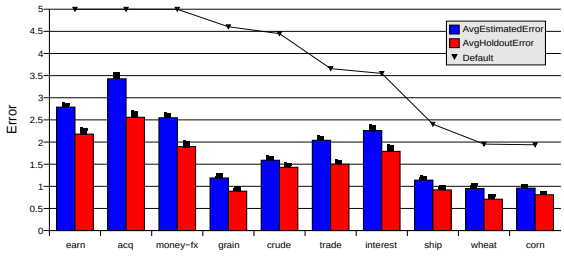
$$(\alpha_i = C) \wedge (\xi_i < 1 - 2CR^2) \heartsuit \ (2\alpha_i R^2 + \xi_i < 1) \heartsuit \ \textit{no leave-one-out error}$$



Experiment: Reuters-21578

- 6451 training examples
- 6451 test examples for holdout testing
- ~27,000 features

Average error estimate over 10 random training/test splits:



=> small bias, variance of estimators is approximately equal

Fast Leave-One-Out Estimation for SVMs

Lemma: Training errors are always leave-out-out errors.

Algorithm:

- (R, α, ξ) = *train_SVM(X,0,0);*
- *for all training examples, do*
  - *if* $\xi_i > 1$ *then loo++;*
  - *else if* $(\rho\alpha_i R^2 + \xi_i < 1)$ *then loo=loo;*
  - *else train_SVM*(X_i, α, ξ)*;*

Experiment:

	Training Examples	Retraining Steps (%)		CPU-Time (sec)	
		$\rho = 1$	$\rho = 2$	$\rho = 1$	$\rho = 2$
Reuters	6451	0.20%	0.58%	11.1	32.3
WebKB	2092	6.78%	20.42%	78.5	235.4
Ohsumed	10000	1.07%	2.56%	433.0	1132.3

Expected Error Rate of SVM

Leave-One-Out Error Estimate: $Err_{loo}(h) = \frac{1}{n} \sum_{i=1}^{n} \left| h_i(\vec{x}_i) = y_i \right|$

For separable problems:

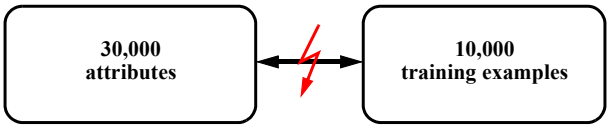
$$\left[h_i(\vec{x}_i) \neq y_i \right] \Leftrightarrow \left[\alpha_i R^2 \geq 1 \right] \qquad \left\| \vec{x} \right\| \leq R$$

$$\Rightarrow \quad Err_{loo}(h_{SVM}) \leq \frac{1}{n} \left| \left[i \mid \alpha_i R^2 \geq 1 \right] \right| \leq \frac{1}{n} \sum_{i=1}^{n} \alpha_i R^2 \leq \frac{1}{n} \frac{R^2}{\delta^2}$$

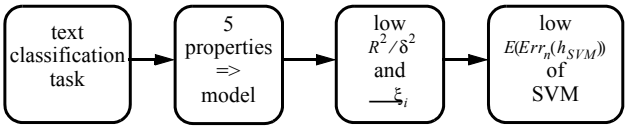
Bound on the expected error rate [Vapnik, 98]:

$$E[Err_{true(n)}(h_{SVM})] = E[Err_{loo(n+1)}(h_{SVM})] \leq \frac{1}{n+1} E\left[\frac{R^2}{\delta^2} \right]$$

Why Do SVMs Work Well for Text Classification?



A statistical learning model of text classification with SVMs:



Margin/Loss Based Bound on the Expected Error

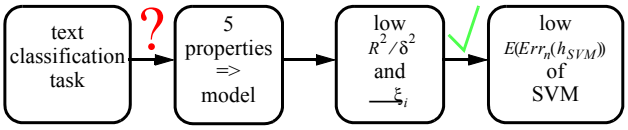
Theorem: The expected error of a soft margin SVM is bounded by

$$E(Err_n(h_{SVM})) \leq \frac{\rho E\left[\frac{R^2}{\delta^2} \right] + \rho C R^2 E\left[\sum_{i=1}^{n+1} \xi_i \right]}{n+1} \qquad C \geq \frac{1}{\rho R^2}$$

$$E(Err_n(h_{SVM})) \leq \frac{\rho E\left[\frac{R^2}{\delta^2} \right] + \rho (C R^2 + 1) E\left[\sum_{i=1}^{n+1} \xi_i \right]}{n+1} \qquad C < \frac{1}{\rho R^2}$$

Where $E\left[\frac{R^2}{\delta^2} \right]$ is the expected soft margin and $E\left[\sum_{i=1}^{n+1} \xi_i \right]$ is the expected training loss on training sets of size $n+1$.

First Step Completed



Properties 1+2: Sparse Examples in High Dimension

- High dimensional feature vectors (30,000 features)
- Sparse document vectors: only a few words of the whole language occur in each document

	Training Examples	Number of Features	Distinct Words (Sparsity)
Reuters Newswire Articles	9,603	27,658	74 (0.27%)
Ohsumed MeSH Abstracts	10,000	38,679	100 (0.26%)
WebKB WWW-Pages	3,957	38,359	130 (0.34%)

Property 3: Heterogeneous Use Of Words

MODULAIRE BUYS BOISE HOMES PROPERTY

Modulaire Industries said it acquired the design library and manufacturing rights of privately-owned Boise Homes for an undisclosed amount of cash. Boise Homes sold commercial and residential prefabricated structures, Modulaire said.

JUSTICE ASKS U.S. DISMISSAL OF TWA FILING

The Justice Department told the Transportation Department it supported a request by USAir Group that the DOT dismiss an application by Trans World Airlines Inc for approval to take control of USAir. ``Our rationale is that we reviewed the application for control filed by TWA with the DOT and ascertained that it did not contain sufficient information upon which to base a competitive review,'' James Weiss, an official in Justice's Antitrust Division, told Reuters.

USX, CONS. NATURAL END TALKS

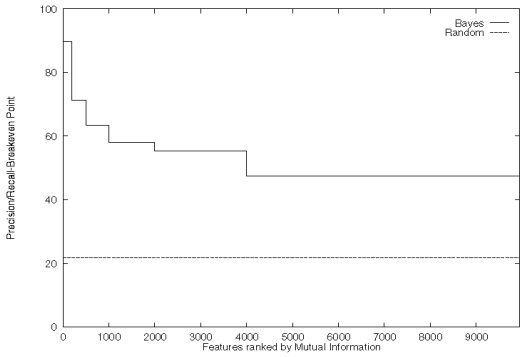
USX Corp's Texas Oil and Gas Corp subsidiary and Consolidated Natural Gas Co have mutually agreed not to pursue further their talks on Consolidated's possible purchase of Apollo Gas Co from Texas Oil. No details were given.

E.D. And F. MAN TO BUY INTO HONG KONG FIRM

The U.K. Based commodity house E.D. And F. Man Ltd and Singapore's Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo's 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

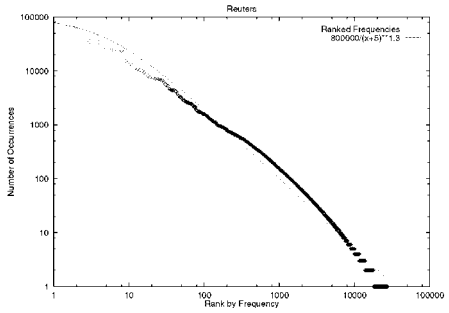
No pair of documents shares any words, but "it", "the", "and", "of", "for", "an", "a", "not", "that", "in".

Property 4: High Level Of Redundancy



=> Few features are irrelevant!

Property 5: " Zipf' s Law"



Zipf' s Law: In text, the i-th frequent word occurs $f_i = \frac{k}{(c+i)^{\Theta}}$ times.

=> Most words occur very infrequently!

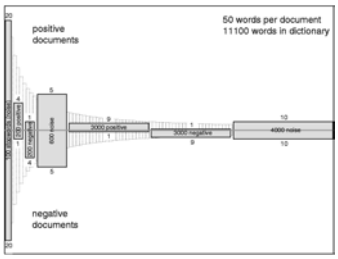
Text Classification Model

Definition: For the TCat-concept there are s disjoint sets of features.

$$TCat([p_1|n_1|f_1], \ldots, [p_s|n_s|f_s])$$

Each positive (negative) example contains p_i (n_i) occurrences from the f_i features in set i.

Example: $TCat$ [20|20|100], [4|1|200], [1|4|200], [5|5|600], [9|1|3000], [1|9|3000], [10|10|4000]



TCat-Concept for WebKB " Course"

$TCat$ [77|29|98], [4|21|52] — *high frequency*
[16|2|431], [1|12|341] — *medium frequency*
[9|1|5045], [1|21|24276] — *low frequency*
[169|191|8116] — *rest*

	high frequency	medium frequency	low frequency
pos	**98 words** all any assignment assignments available bc book c chapter class code course cse description discussion document due each ecs exam exams fall final ... section set should solution solutions spring structures students syllabus ta text textbook there thursday topics tuesday unix use wednesday week will you your	**431 words** account acrobat adapted addison ah ahead aho allowed alternate announced announcement announcements answers appointment approximately ... mesdays runing rum turned ruth txt uldaho unswa ullman understand ungraded units useless upenn use vectors vi walter weaver wed wednesdays weekly weeks weights wesley yurtas	**5045 words** 002cc 009a 00a 00cm 01act 01pm 02pm 03oct 03pm 03sep 04dec ... gradable gradebook gradebooks gradefreq gradefreq2 gradefreq3 graders gradesheet gradients grafica grafik ... zimmermann zinc zipi zipser zj zlorate znol zoran zp zwatch zwhere zwiener zyda
neg	**52 words**	**341 words**	**24276 words**
	acm address am austin ca california center college computational conference contact current currently d department dr faculty fax graduate group he ... me member mv our parallel performance ph pp proceedings professor publications recent research sciences support technical technology university vision was working	anal academy accesses accurate adaptation advisor advisory affiliated affiliations agent agents alberta album alumni amanda america amherst annual ... victoria virginia visiting visitors visualization vita vitae voice wa watson weather webster went west wi wife wireless wisconsin worked workshop workshops wrate yale york	0a 0b 0b1 0e 0f 0r 0software 0x82d4ff 100k 100mhz 100th 1020x620 102k 103k ... bmar burches lunchtime lund lundberg lunch lung luniewski luo luong lupin lupton lure lurker lus ... zuo znowei zurich zvi zw zwaenepoel zwarico zwickau zwilling zygmunt zzlm00
	high frequency	medium frequency	low frequency

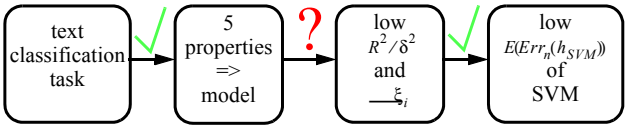
Real Text Classification Tasks as TCat-Concepts

Reuters " Earn" : $TCat$ [33|2|65], [32|65|152] — *high frequency*; [2|1|171], [3|21|974] — *medium frequency*; [3|1|3455], [1|10|17020] — *low frequency*; [78|52|5821] — *rest*

Webkb " Course" : $TCat$ [77|29|98], [4|21|52] — *high frequency*; [16|2|431], [1|12|341] — *medium frequency*; [9|1|5045], [1|21|24276] — *low frequency*; [169|191|8116] — *rest*

Ohsumed " Pathology" : $TCat$ [2|1|10], [1|4|22] — *high frequency*; [2|1|92], [1|2|94] — *medium frequency*; [5|1|4080], [1|10|20922] — *low frequency*; [197|190|13459] — *rest*

Second Step Completed



The Margin δ^2 of TCat-Concepts

Lemma 1: For $TCat([p_1|n_1|f_1], \ldots, [p_s|n_s|f_s])$ -concepts there is always a hyperplane passing through the origin with margin δ^2 at least

$$\delta^2 \geq \frac{ad - b^2}{a + 2b + d} \quad with \quad a = \sum_{i=1}^{s} \frac{p_i^2}{f_i}, \quad d = \sum_{i=1}^{s} \frac{n_i^2}{f_i}, \quad b = \sum_{i=1}^{s} \frac{n_i p_i}{f_i}$$

Example: The previous example WebKB " course" has a margin of at least

$$\delta^2 \geq 0.23$$

The Length R^2 of Document Vectors

Lemma 2: If the ranked term frequencies f_i in a document with l words have the form of the generalized Zipf" s Law

$$f_i = \frac{k}{(c+i)^{\Theta}}$$

based on their frequency rank i, then the Euclidean length of the document vector $\vec{x}$ is bounded by

$$\left\|\vec{x}\right\| \leq \sqrt{\sum_{i=1}^{d} \left(\frac{k}{(c+i)^{\Theta}}\right)^2} \quad with \quad \sum_{i=1}^{d} \frac{k}{(c+i)^{\Theta}} = l$$

Example: For WebKB " course" with

$$f_i = \frac{470000}{(5+i)^{1.25}}$$

follows that $R^2 \leq 1900$.

R^2, δ^2, and $\sum \xi_i$ for Text Classification

Reuters Newswire Stories

- 10 most frequent categories
- 9603 training examples
- 27658 attributes

$$E(Err_P(h_{SVM})) \leq \frac{E\left[R^2\right]\delta^2 + CR^2 E\left[\sum_{i=1}^{n+1} \xi_i\right]}{n+1}$$

	R^2/δ^2	$\sum \xi_i$
earn	1143	0
acq	1848	0
money-fx	1489	27
grain	585	0
crude	810	4

	R^2/δ^2	$\sum \xi_i$
trade	869	9
interest	2082	33
ship	458	0
wheat	405	2
corn	378	0

Learnability of TCat-Concepts

Theorem: For $TCat([p_1|n_1|f_1], \ldots, [p_s|n_s|f_s])$ -concepts and documents with l words that follow the generalized Zipf" s Law $f_i = k/(c+i)^{\Theta}$ the expected generalization error of an unbiased SVM after training on n examples is bounded by

$$E(Err_n(h_{SVM})) \leq \frac{R^2}{n+1} \frac{ad - b^2}{a + 2b + d} \quad with \quad a = \sum_{i=1}^{s} \frac{p_i^2}{f_i}, \quad d = \sum_{i=1}^{s} \frac{n_i^2}{f_i}, \quad b = \sum_{i=1}^{s} \frac{n_i p_i}{f_i}, \quad R^2 \leq \sum_{i=1}^{d} \left(\frac{k}{(c+i)^{\Theta}}\right)^2$$

Comparison Theory vs. Experiments

	Learning Curve Bound	Predicted Bound on Error Rate	Error Rate in Experiment
Reuters " earn"	$E(Err_n(h_{SVM})) \leq \frac{138}{n+1}$	1.5%	1.3%
WebKB " course"	$E(Err_n(h_{SVM})) \leq \frac{443}{n+1}$	11.2%	4.4%
Ohsumed " pathology"	$E(Err_n(h_{SVM})) \leq \frac{9457}{n+1}$	94.6%	23.1%

- Model can differentiate between " difficult" and " easy" tasks
- Predicts and reproduces the effect of information retrieval heuristics (e.g. TFIDF-weighting)

Sensitivity Analysis

What makes a text classification problem suitable for a linear SVM?

High Redundancy:

$$TCat\left(\begin{matrix}[40|40|50] \\ [25|5|1000], [5|25|1000] \\ [30|30|30000]\end{matrix}\right) \quad \begin{matrix}high\ frequency \\ medium\ frequency \\ low\ frequency\end{matrix}$$

High Discriminatory Power:

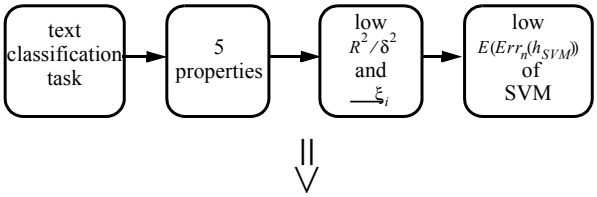
$$TCat\left(\begin{matrix}[40|40|50] \\ [15|0|500], [0|15|500], [15|15|1000] \\ [30|30|30000]\end{matrix}\right) \quad \begin{matrix}high\ frequency \\ medium\ frequency \\ low\ frequency\end{matrix}$$

High Frequency:

$$TCat\left(\begin{matrix}[16|4|10], [4|16|10], [20|20|30] \\ [30|30|2000] \\ [30|30|30000]\end{matrix}\right) \quad \begin{matrix}high\ frequency \\ medium\ frequency \\ low\ frequency\end{matrix}$$

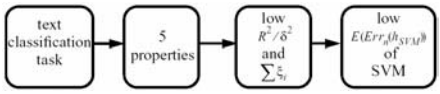
What does this Model Provide?

Connects the statistical properties of text classification tasks with generalization error of SVM!



- Explains the behavior of (linear) SVMs on text classification tasks
- Gives guideline for when to apply (linear) SVMs
- Provides formal basis for developing new methods

Summary
When do (Linear) SVMs Work Well?



Intuition: If the problem can be cast as a TCat-concept with

- high redundancy,
- strongly discriminating features
- particularly in the high frequency region

then linear SVMs achieve a low generalization error.

Assumptions and Restrictions:

- no noise (attribute and classification)
- no variance (only " average" examples)
- only upper bounds, no lower bounds