

Text Categorization

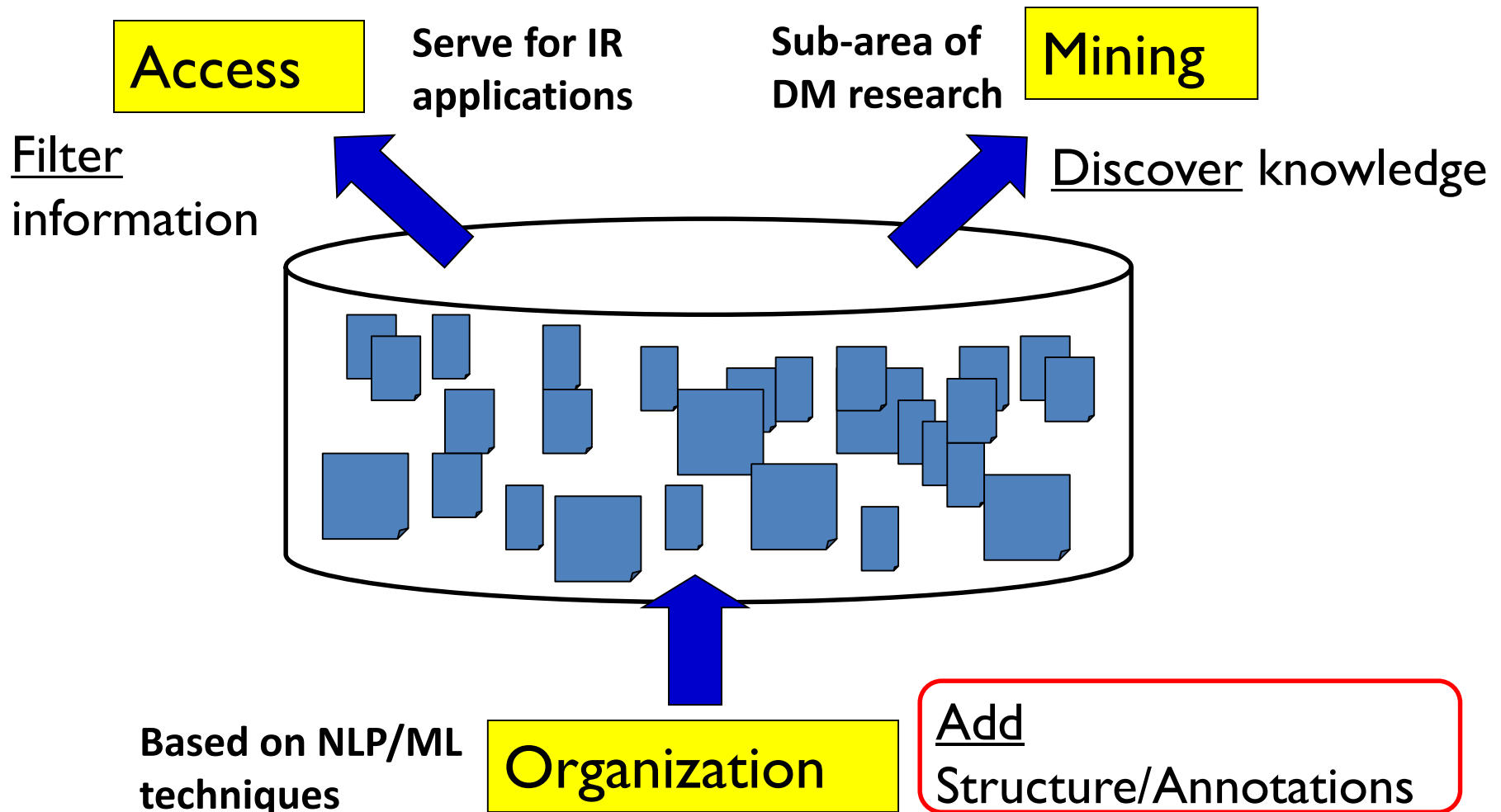
Hongning Wang

CS@UVa

Today's lecture

- Bayes decision theory
- Supervised text categorization
 - General steps for text categorization
 - Feature selection methods
 - Evaluation metrics

Text mining in general



Applications of text categorization

- Automatically classify politic news from sports news

political

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky State of the Union address after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

sports

PRO FOOTBALL | ANALYSIS

Super Bowl 2015: Patriots' Red-Hot Offense Faces Seahawks' Dominant Defense

By CHASE STUART JAN. 20, 2015

Email

Share

Tweet

Pin

Save

More

Last year's Super Bowl pitted one of the greatest single-season offenses in N.F.L. history against one of the greatest single-season defenses. Using slightly different time frames, this year's Super Bowl can boast similar claims.

Both the New England Patriots and the Seattle Seahawks had slow starts in 2014. After New England's 41-14 loss to the Kansas City Chiefs in Week 4, pundits wondered if we were witnessing the end of the Tom Brady/Bill Belichick-era Patriots. But since that game, the offensive line emerged as a cohesive unit, Rob Gronkowski's health improved and Brady became red-hot. Since that



The Seahawks' Richard Sherman intercepting a pass against the Packers in the N.F.C. title game. David J. Phillip/Associated Press

Applications of text categorization

- Recognizing spam emails

```
Received: from 192.168.1.100 ([65.202.85.3]) by pacific-carrier-annex.mit.edu  
(8.9.2/8.9.2) with SMTP id AAA06179;  
Mon, 11 Jun 2001 00:39:32 -0400 (EDT)  
From: [some forged email address]  
Message-ID: <200106110439.AAA06179@pacific-carrier-annex.mit.edu>  
Subject: I am as shocked as you!  
Date: Sun, 10 Jun 01 00:32:35 Pacific Daylight Time  
X-Priority: 3  
X-MSMailPriority: Normal  
Importance: Normal  
MIME-Version: 1.0  
Content-Type: multipart/mixed;  
boundary="-----_NextPart_000_018C_01BD9940.715D52A0"
```

```
<HTML>  
<BODY>
```

Spam=**True**/False

```
<FONT face="MS Sans Serif">  
<FONT size=2> <BR>  
<BR>  
Some of the most beautiful women in the world bare it all for you.Denise Richard  
s, Britney Spears, Jessica Simpson, and many more.<A HREF="http://216.130.166.1  
88/index.html">CLICK HERE FOR NUDE CELEBS</A><BR>  
<BR>  
</FONT></FONT></BODY></HTML>
```

Applications of text categorization

- Sentiment analysis



The best tablet, but not a necessary one., November 25, 2014

By Andy, an Amazon Customer (Fargo, ND) - [See all my reviews](#)

This review is from: Apple iPad Air 2 MH0W2LL/A (16GB, Wi-Fi, Gold) NEWEST VERSION (Personal Computers)

Short version: if you don't have a tablet yet, this is the one to get holiday 2014. If you already have a tablet that you're mostly happy with, whether an iPad or Android version, keep it.

I purchased the new iPad Air 2, in Gold, 16GB capacity about a week ago at Walmart, and I'd like to give a few impressions of the hardware and software here. I had particularly high hopes for this device, and have been waiting a long time to buy one; after holding a friend's brand new 64GB version, and being really impressed by how light the device seemed, I bought one for myself! :)

A little bit of background: My other experience with tablets involves a 2013 Nexus 7 that I use at least weekly; an Asus Transformer Pad, with a Tegra 3 1920x1080 screen, an Acer android tablet whose screen cracked 3 months after purchase; a Kindle Fire HD; I have also used both an iPad 2 and an iPad Mini (original) off and on, but never owned an iPad before. I use an iPhone 5.

The device is extremely light and thin. Its shocking, honestly - its far lighter than my chunky Kindle Fire HD 7. I bought it in gold (because why not live a little?) and it looks really nice. It feels like a premium device. The back is metal, which can be a little cold to the touch, but is smooth and easy to hold. It does get tedious holding it up while lying in bed, however. Probably this is due part to the small side bezels; my palm or thumb was nearly always bumping the screen.

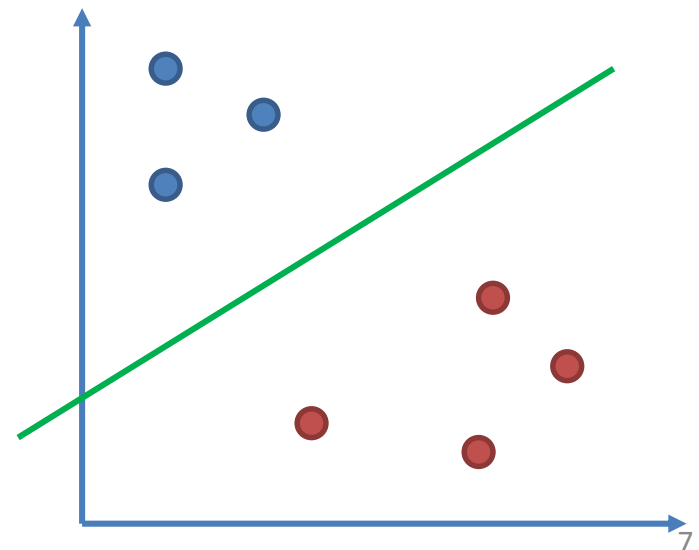
The screen is gorgeous. Bright, easy to read, and I haven't noticed any reflections on it yet, which is fantastic. Honestly, its beautiful. And it shows off photographs really really well. I haven't used it to take any pictures, and probably won't, so I can't really comment on that aspect.

The software is good, but I was honestly expecting something noticeably better than iOS 8 on my iPhone, which just isn't the case. In fact, because of the animations, and the larger screen, it feels almost slower than my two year old iPhone.

Basic notions about categorization

- Data points/Instances
 - X : an m -dimensional feature vector
- Labels
 - y : a categorical value from $\{0, \dots, k - 1\}$
- Classification hyper-plane
 - $f(X) \rightarrow y$

**Key question: how to
find such a mapping?**




Bayes decision theory

- If we know $p(y)$ and $p(X|y)$, the Bayes decision rule is



$$\hat{y} = \operatorname{argmax}_y p(y|X) = \operatorname{argmax}_y \frac{p(X|y)p(y)}{p(X)}$$

Constant with
respect to y



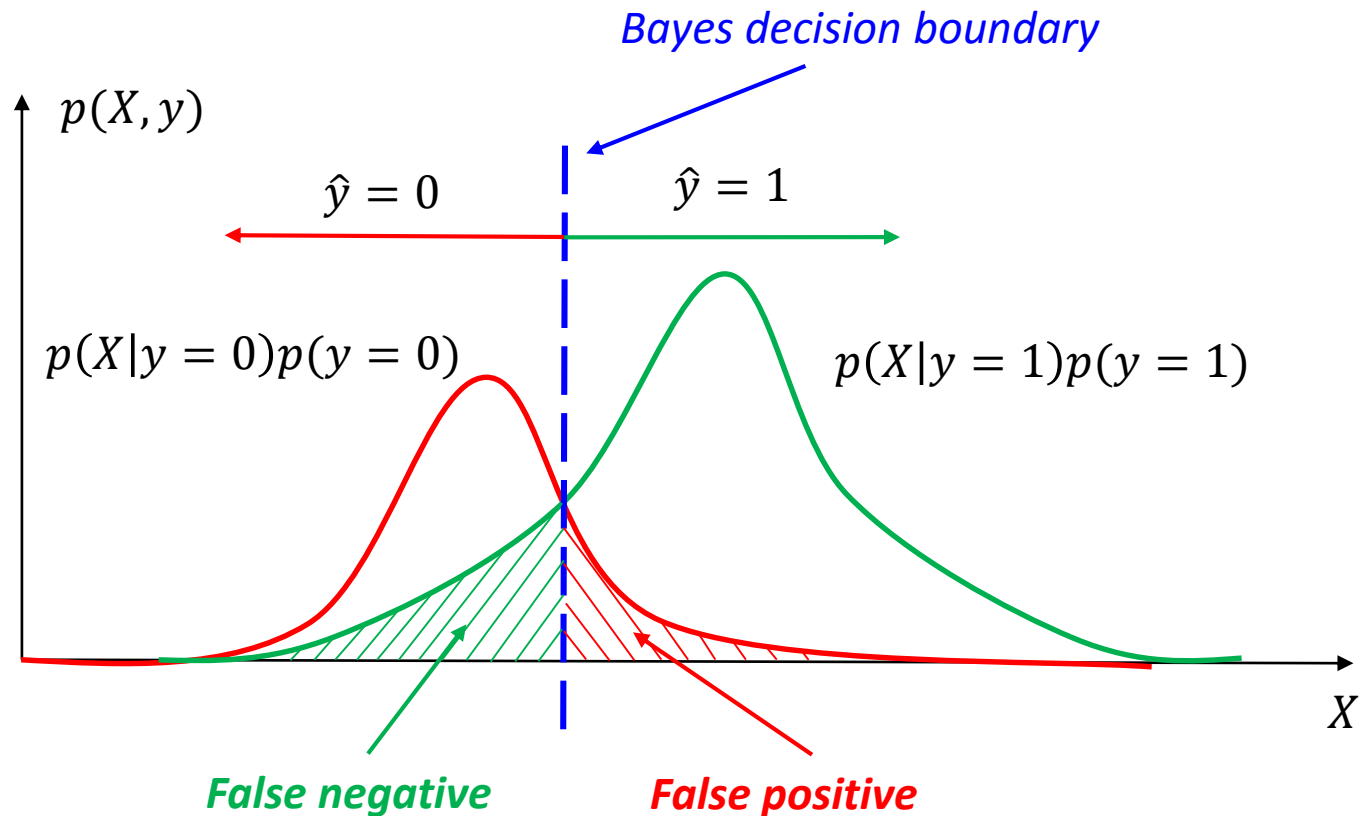
- Example in binary classification
 - $\hat{y} = 1$, if $p(X|y = 1)p(y = 1) > p(X|y = 0)p(y = 0)$
 - $\hat{y} = 0$, otherwise
- This leads to optimal classification result
 - Optimal in the sense of ‘risk’ minimization

Bayes risk

- Risk – assign instance to a wrong class
 - Type I error: $(y^* = 0, \hat{y} = 1)$  *False positive*
 - $p(X|y = 0)p(y = 0)$
 - Type II error: $(y^* = 1, \hat{y} = 0)$  *False negative*
 - $p(X|y = 1)p(y = 1)$
 - Risk by Bayes decision rule
 - $r(X) = \min\{p(X|y = 1)p(y = 1), p(X|y = 0)p(y = 0)\}$
 - It can determine a ‘reject region’

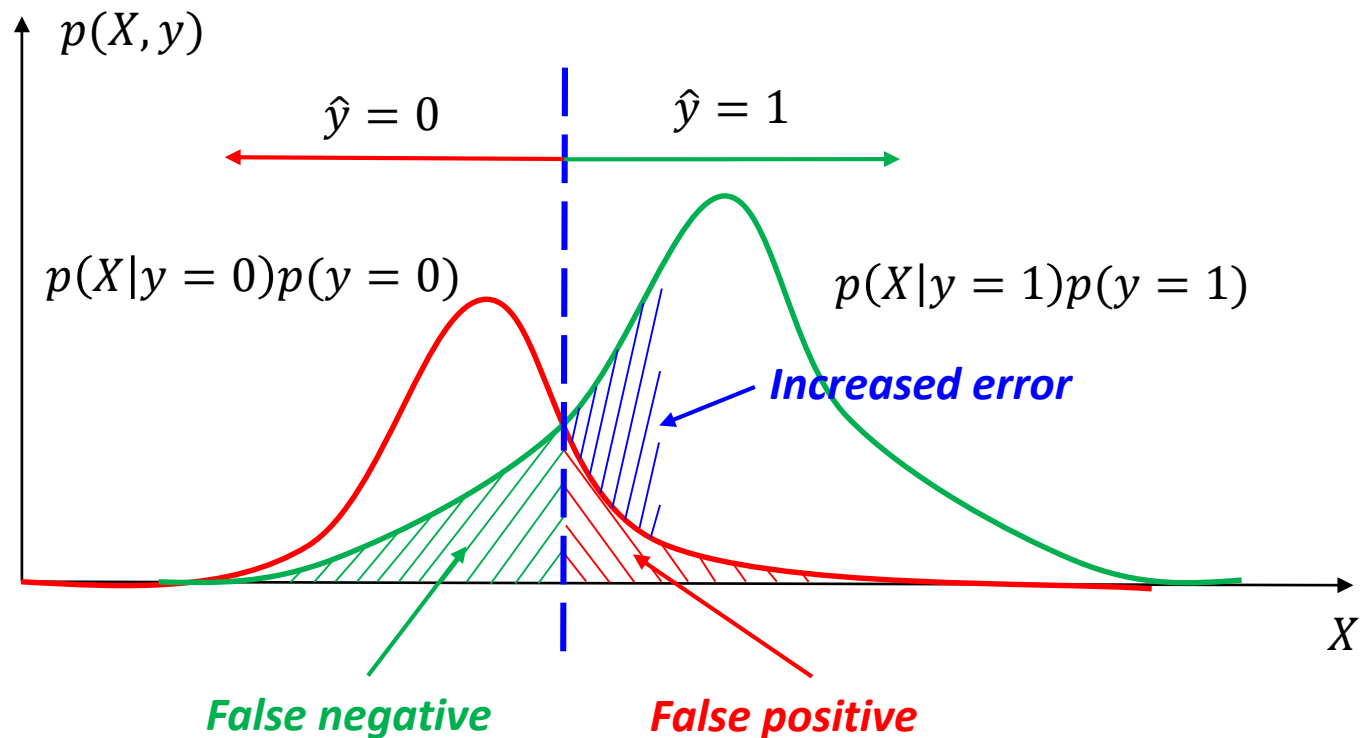
Bayes risk

- Risk – assign instance to a wrong class



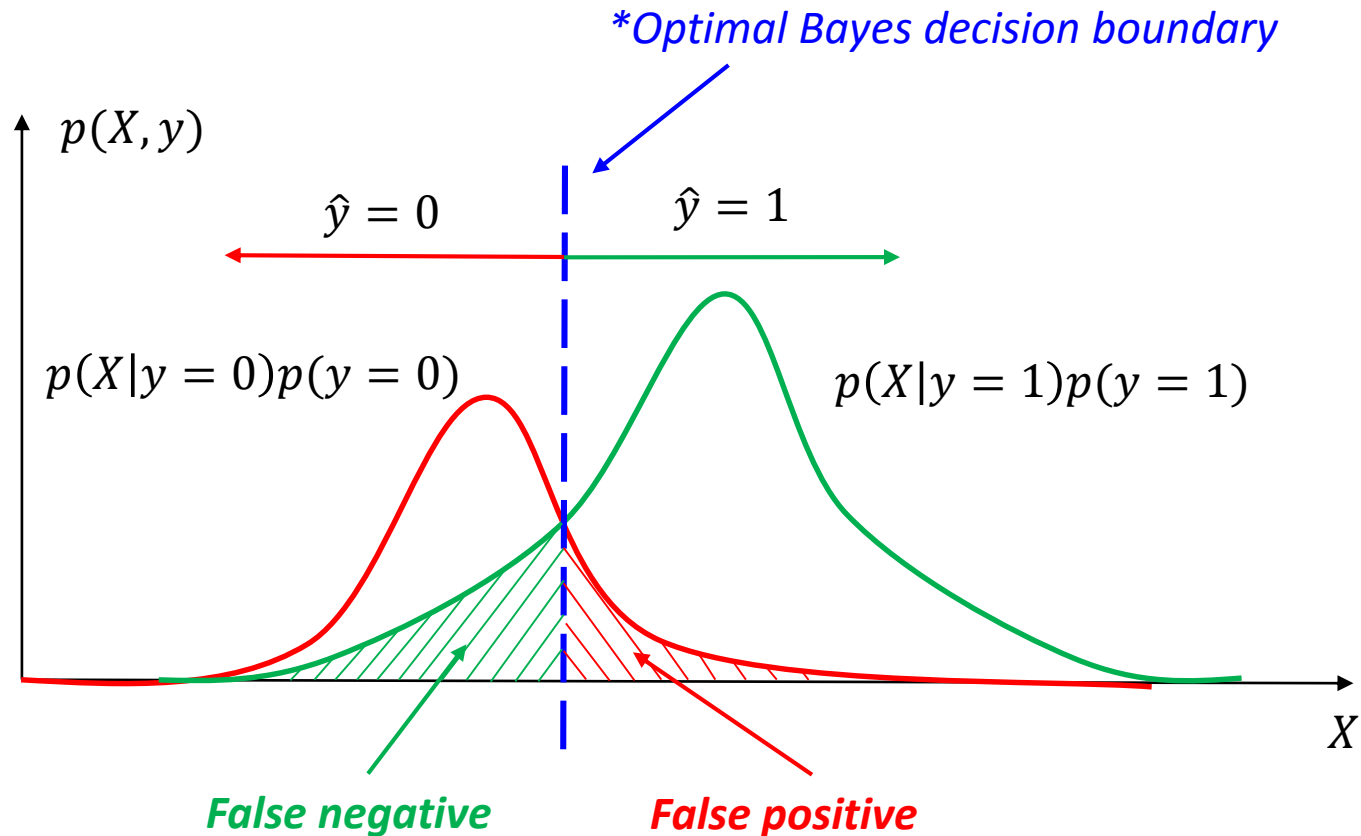
Bayes risk

- Risk – assign instance to a wrong class



Bayes risk

- Risk – assign instance to a wrong class



Bayes risk

- Expected risk

$$E[r(x)] = \int_x p(x)r(x)dx$$

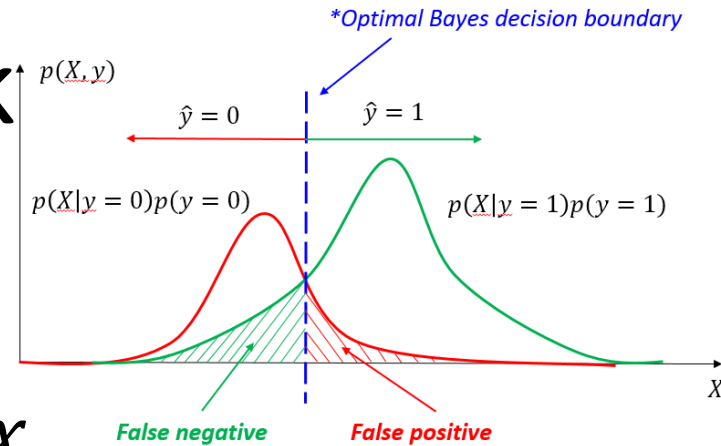
$$= \int_x p(x) \min\{p(x|y=1)p(y=1), p(x|y=0)p(y=0)\}dx$$

$$= p(y=1) \int_{R_0} p(x)p(x|y=1)dx$$

$$+ p(y=0) \int_{R_1} p(x)p(x|y=0)dx$$

Region where we assign x to class 0

Region where we assign x to class 1



Will the error of assigning c_1 to c_0 be always equal to the error of assigning c_0 to c_1 ?

Recap: IBM translation models

- A generative model based on noisy channel framework
 - Generate the translation sentence e with regard to the given sentence f by a stochastic process
 1. Generate the length of f
 2. Generate the **alignment** of e to the target sentence f
 3. Generate the words of f
 - $Eng^* = \operatorname{argmax}_{Eng} p(Fre|Eng)p(Eng)$

Recap: IBM translation models

- Translation model with word alignment

$$- p(Fre|Eng) = \sum_{a \in A(Eng, Fre)} p(Fre, a|Eng)$$

marginalize over all possible alignments a

- Generate the words of f with respect to alignment a

$$p(\mathbf{f}, \mathbf{a} | \mathbf{e}) = p(m | \mathbf{e}) \prod_{j=1}^m p(a_j | a_{1..j-1}, f_{1..j-1}, m, \mathbf{e}) p(f_j | a_{1..j}, f_{1..j-1}, m, \mathbf{e})$$

Length of target sentence f

Word alignment a_j

Translation of f_j

Recap: decoding process in Model 1

$$p(e|f) = 1e^{-55}$$

For a particular English sentence $e = e_1..e_n$ of length n

$p(e)$

0	1	2	3	4	5
NULL	John	flies	across	the	river

Search through all English sentences

1. Choose a length m for the target sentence (e.g $m = 8$)

$$p(m|e) = \epsilon$$

1	2	3	4	5	6	7	8
?	?	?	?	?	?	?	?

Search through all possible alignments

2. Choose an alignment $a = a_1 ... a_m$ for the source sentence

$$p(a|e) = \frac{1}{n}$$

Target Position	1	2	3	4	5	6	7	8
Source Position	1	2	4	5	5	2	0	3

$$\prod_{j=1}^m$$

$$p(f_j|e_{a_j})$$

3. Translate each source word e_{a_j} into the target language

English	John	flies	the	river	river	flies	NULL	across
Alignment	1	2	4	5	5	2	0	3
Encoded	Jean	a	traversé	le	lac	à	la	nage

Order of action

Receiver

Recap: decoding process in Model 1

$$p(e|f) = 1e^{-15}$$

For a particular English sentence $e = e_1..e_n$ of length n

$p(e)$

0	1	2	3	4	5
NULL	John	swam	across	the	lake

Search through all English sentences

1. Choose a length m for the target sentence (e.g $m = 8$)

$$p(m|e) = \epsilon$$

1	2	3	4	5	6	7	8
?	?	?	?	?	?	?	?

Search through all possible alignments

2. Choose an alignment $a = a_1 ... a_m$ for the source sentence

$$p(a|e) = \frac{1}{n}$$

Target Position	1	2	3	4	5	6	7	8
Source Position	1	3	3	4	5	0	0	2

3. Translate each source word e_{a_j} into the target language

$$\prod_{j=1}^m p(f_j|e_{a_j})$$

English	John	across	across	the	lake	NULL	NULL	swam
Alignment	1	3	3	4	5	0	0	2
Encoded	Jean	a	traversé	le	lac	à	la	nage

Order of action


Receiver

Recap: Bayes decision theory

- If we know $p(y)$ and $p(X|y)$, the Bayes decision rule is

$$\hat{y} = \operatorname{argmax}_y p(y|X) = \operatorname{argmax}_y \frac{p(X|y)p(y)}{p(X)}$$

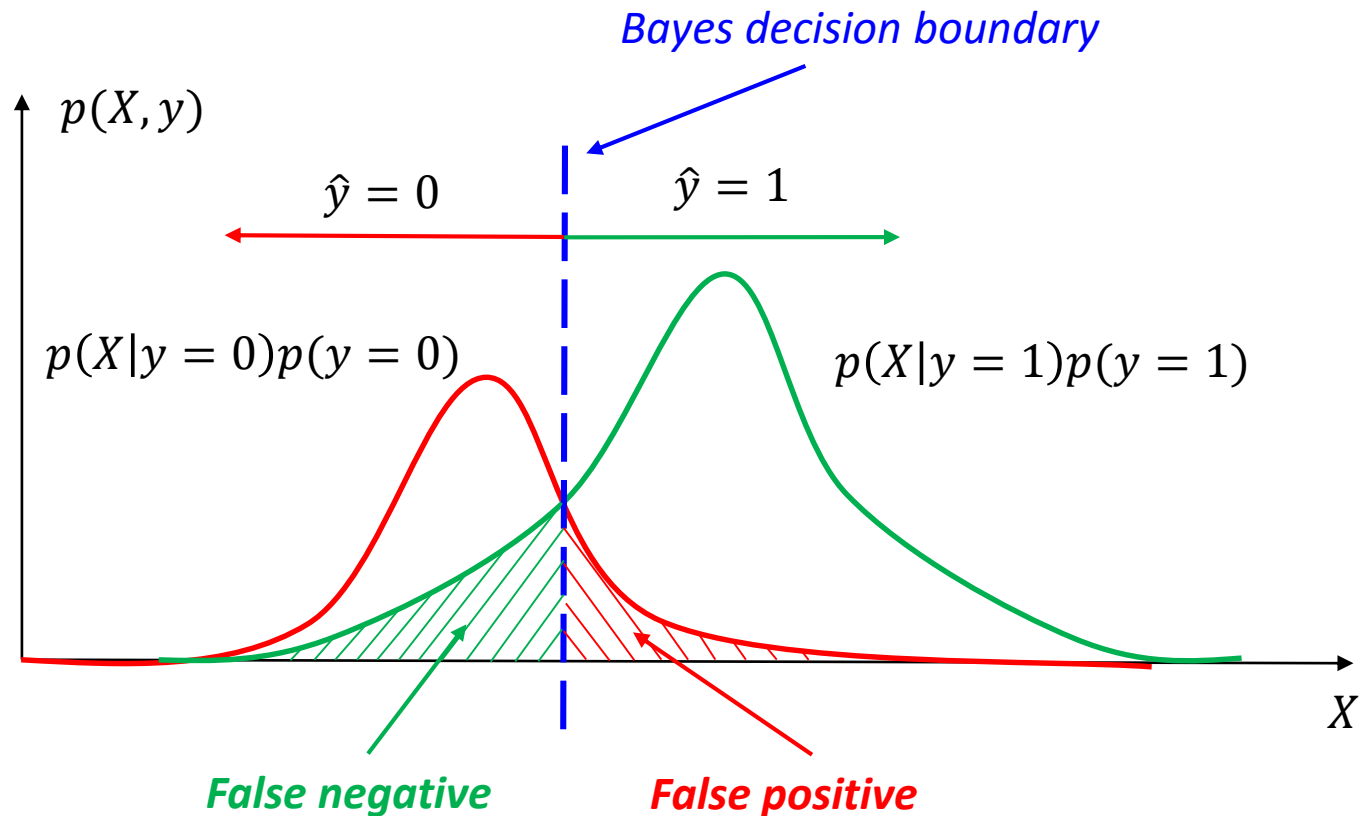
Constant with
respect to y



- Example in binary classification
 - $\hat{y} = 1$, if $p(X|y = 1)p(y = 1) > p(X|y = 0)p(y = 0)$
 - $\hat{y} = 0$, otherwise
- This leads to optimal classification result
 - Optimal in the sense of ‘risk’ minimization

Recap: Bayes risk

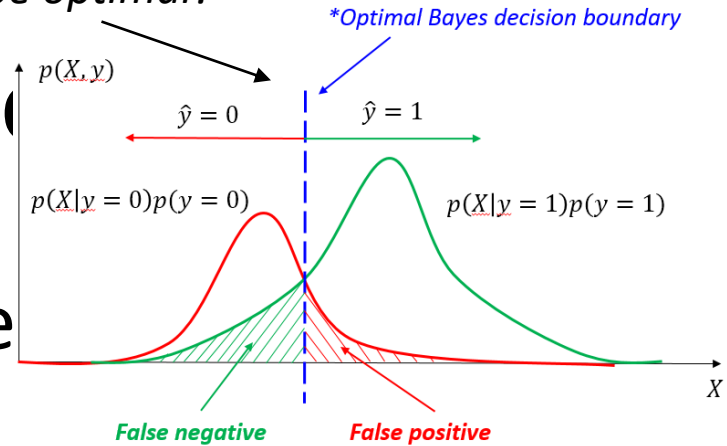
- Risk – assign instance to a wrong class



Will this still be optimal?

Loss function

- The penalty we will pay when misclassifying instances



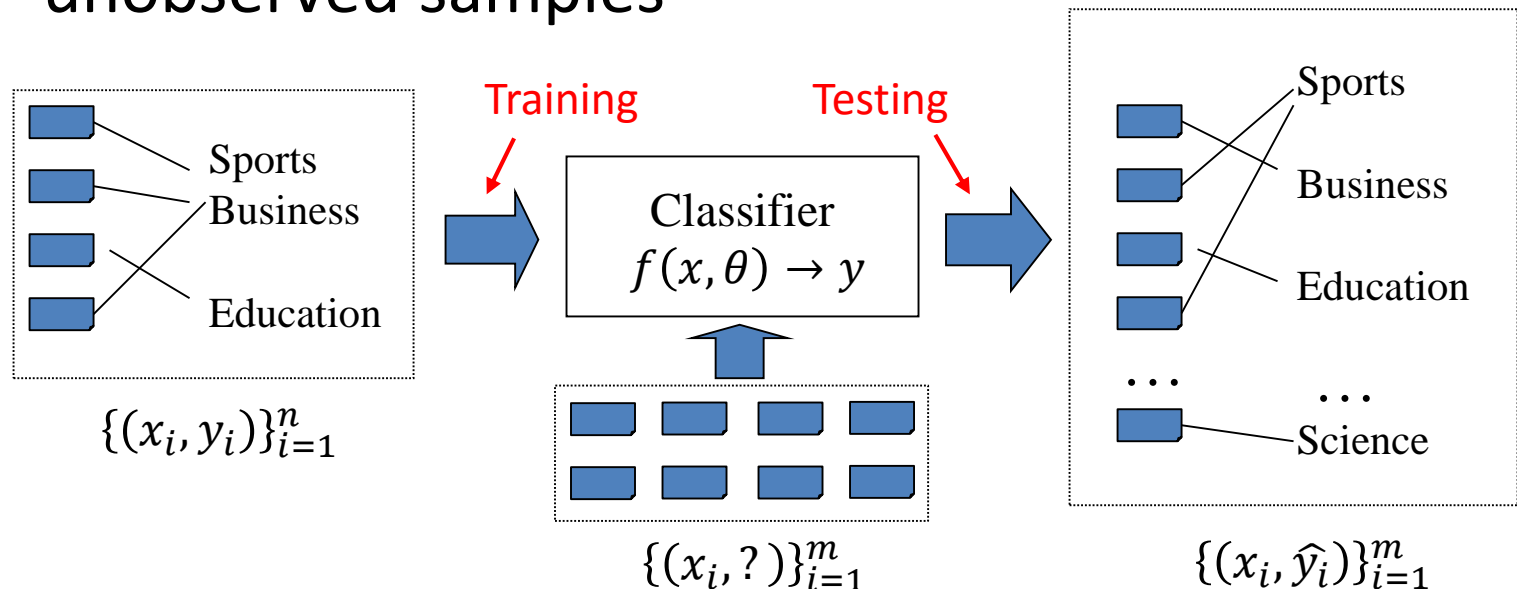
Penalty when misclassifying c_1 to c_0 \rightarrow $E[L] = L_{1,0}p(y=1) \int_{R_0} p(x)p(x|y=1)dx$ Region where we assign x to class 0

Penalty when misclassifying c_0 to c_1 \rightarrow $+L_{0,1}p(y=0) \int_{R_1} p(x)p(x|y=0)dx$ Region where we assign x to class 1

- Goal of classification in general
 - Minimize loss

Supervised text categorization

- Supervised learning
 - Estimate a model/method from labeled data
 - It can then be used to determine the labels of the unobserved samples



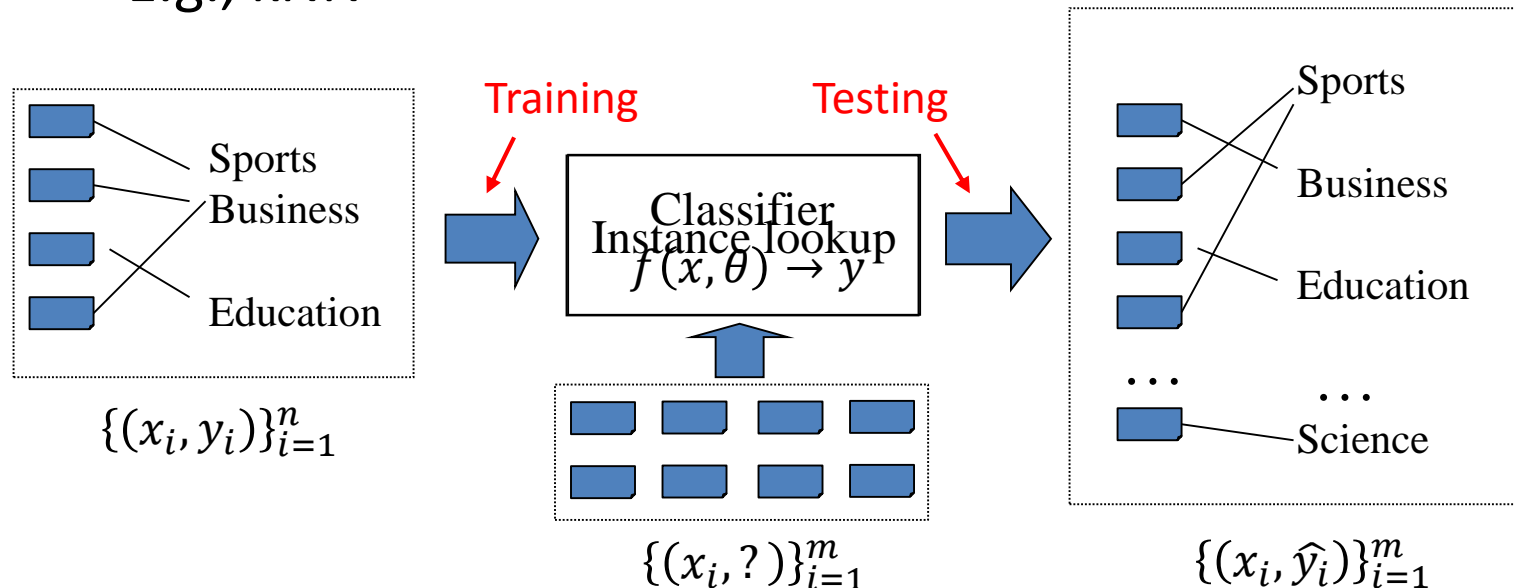
Type of classification methods

- Model-less

- Instance based classifiers

- Use observation directly
 - E.g., kNN

Key: assuming similar items have similar class labels!



Type of classification methods

- Model-based

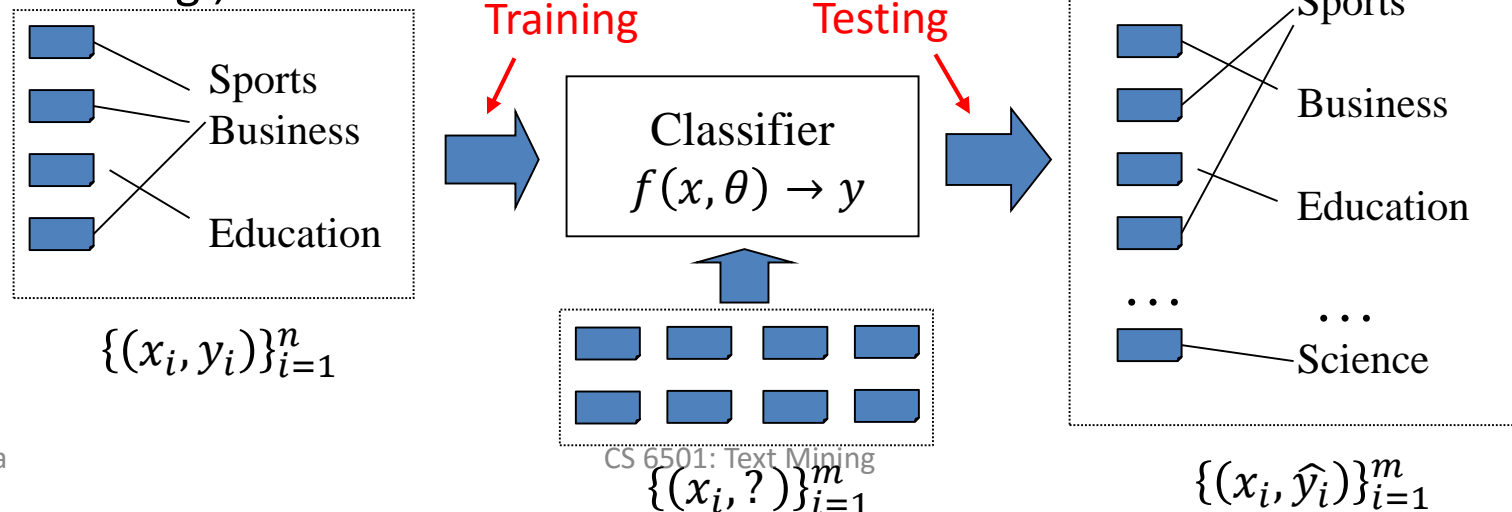
- Generative models

- Modeling joint probability of $p(x, y)$
 - E.g., Naïve Bayes

Key: i.i.d. assumption!

- Discriminative models

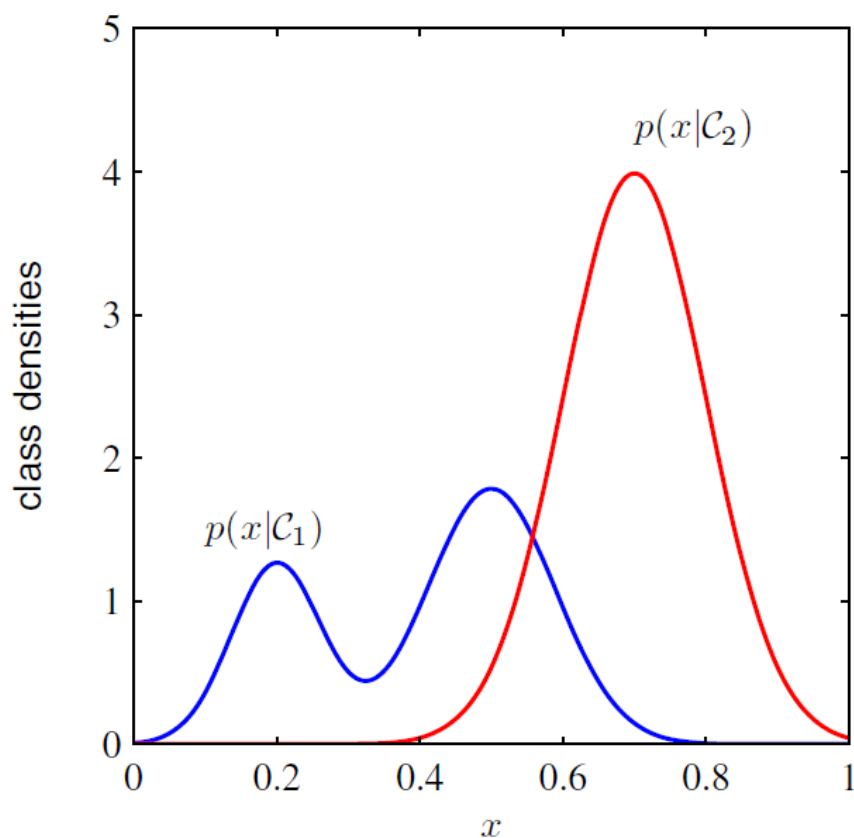
- Directly estimate a decision rule/boundary
 - E.g., SVM



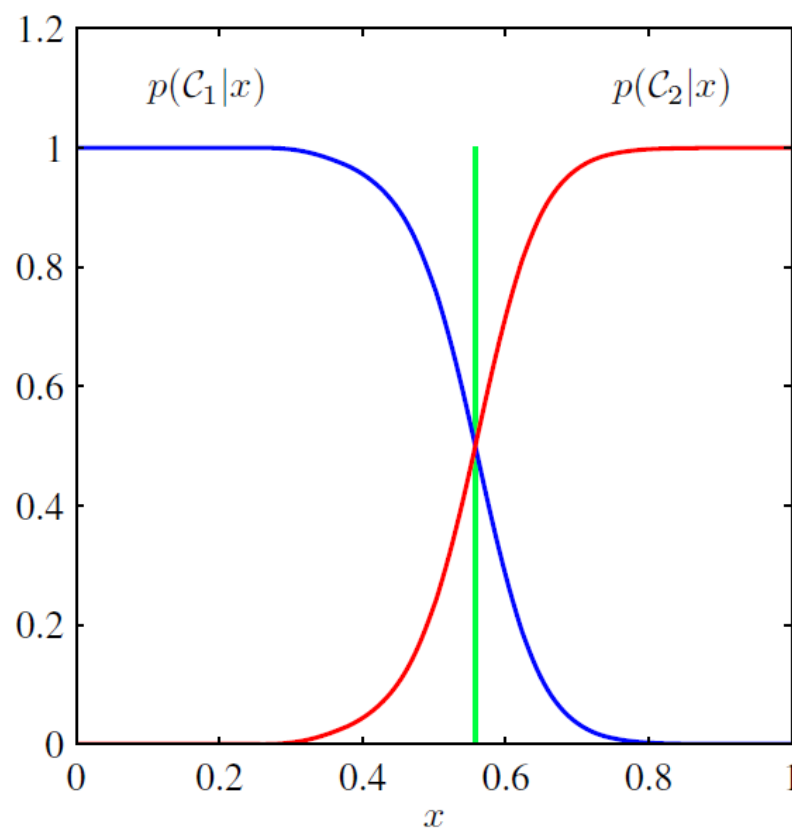
Generative V.S. discriminative models

- Binary classification as an example

Generative Model's view



Discriminative Model's view



Generative V.S. discriminative models

Generative

- Specifying joint distribution
 - Full probabilistic specification for all the random variables
- Dependence assumption has to be specified for $p(x|y)$ and $p(y)$
- Flexible, can be used in unsupervised learning

Discriminative

- Specifying conditional distribution
 - Only explain the target variable
- Arbitrary features can be incorporated for modeling $p(y|x)$
- Need labeled data, only suitable for (semi-) supervised learning

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, [President Obama](#) walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky [State of the Union address](#) after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political
News



Sports
News



Entertainment
News

1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, [President Obama](#) walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky [State of the Union address](#) after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political
News



Sports
News



Entertainment
News



1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

- 1.1 How to represent the text documents?
- 1.2 Do we need all those features?

Feature construction for text categorization

- Vector space representation
 - Standard procedure in document representation
 - Features
 - N-gram, POS tags, named entities, topics
 - Feature value
 - Binary (presence/absence)
 - TF-IDF (many variants)

Recall MP1

- How many unigram+bigram are there in our controlled vocabulary?
 - 130K on Yelp_small
- How many review documents do we have there for training?
 - 629K Yelp_small

Very sparse feature representation!

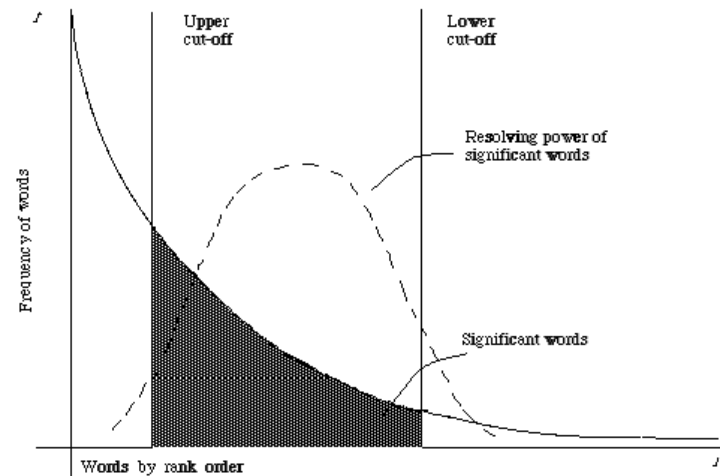


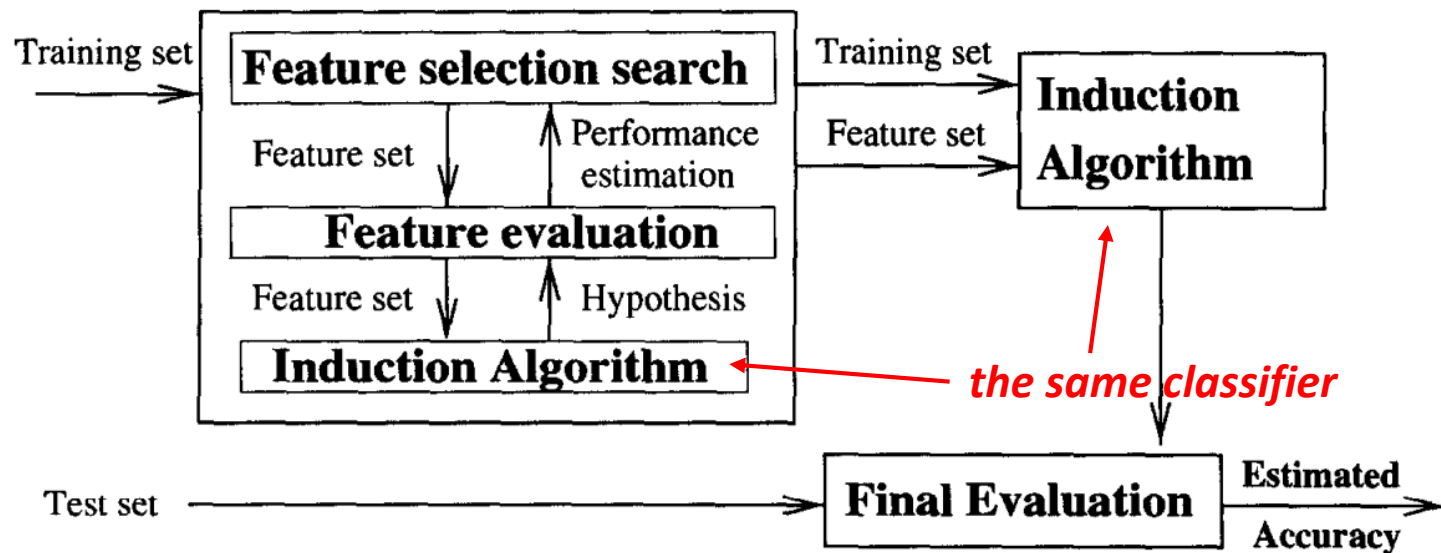
Figure 2.1. A plot of the hyperbolic curve relating f , the frequency of occurrence and r , the rank order (Adapted from Schultz⁴⁴ page 120)

Feature selection for text categorization

- Select the most informative features for model training
 - Reduce noise in feature representation
 - Improve final classification performance
 - Improve training/testing efficiency
 - Less time complexity
 - Fewer training data

Feature selection methods

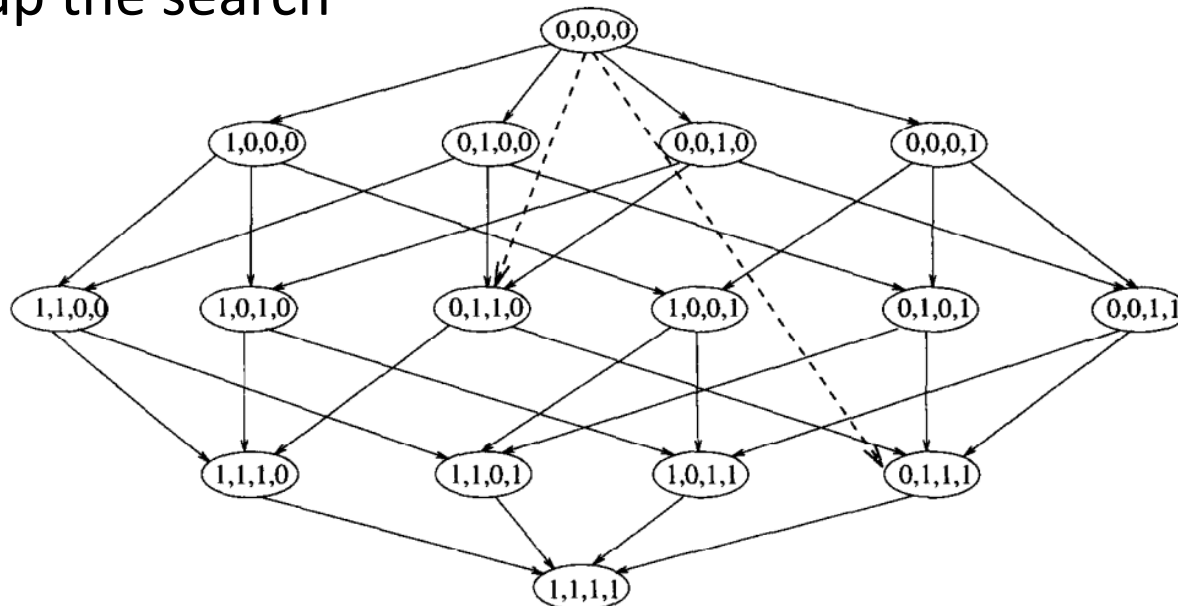
- Wrapper method
 - Find the best subset of features for a particular classification method



R. Kohavi, G.H. John/Artificial Intelligence 97 (1997) 273-324

Feature selection methods

- Wrapper method
 - Search in the whole space of feature groups
 - Sequential forward selection or genetic search to speed up the search



R. Kohavi, G.H. John/Artificial Intelligence 97 (1997) 273-324

CS 6501: Text Mining

Feature selection methods

- Wrapper method
 - Consider all possible dependencies among the features
 - Impractical for text categorization
 - Cannot deal with large feature set
 - A NP-complete problem
 - No direct relation between feature subset selection and evaluation

Feature selection methods

- Filter method
 - Evaluate the features independently from the classifier and other features
 - No indication of a classifier's performance on the selected features
 - No dependency among the features
 - Feasible for very large feature set



R. Kohavi, G.H. John/Artificial Intelligence 97 (1997) 273-324

Feature scoring metrics

- Document frequency
 - Rare words: non-influential for global prediction, reduce vocabulary size

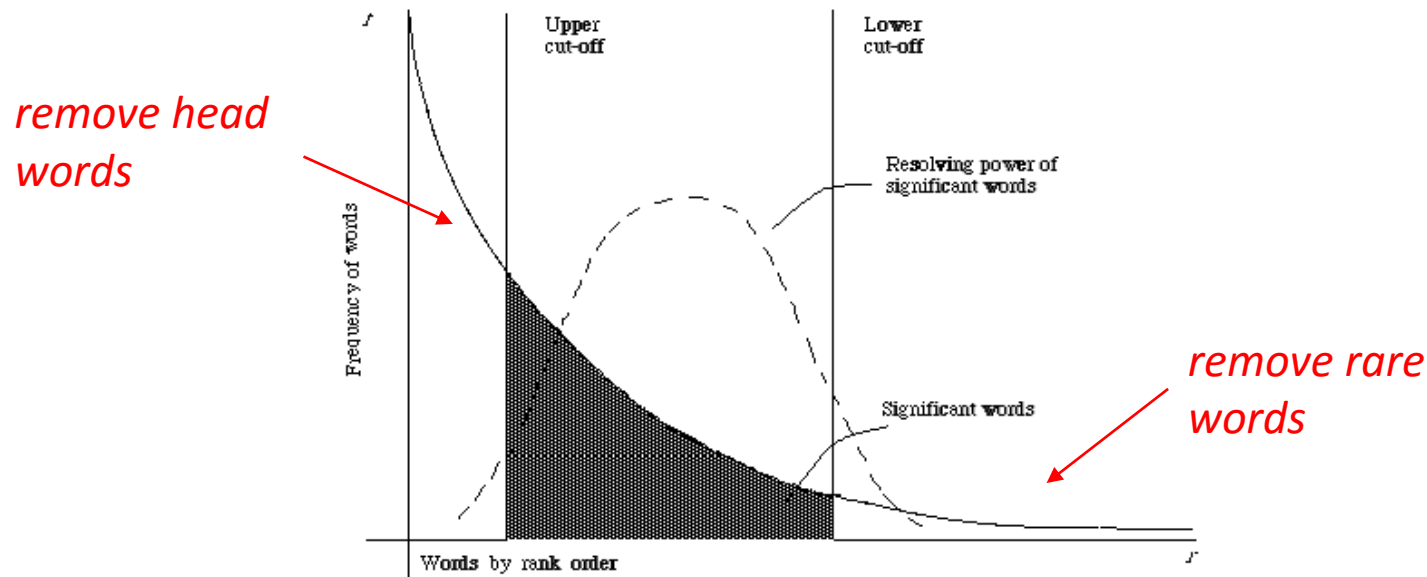
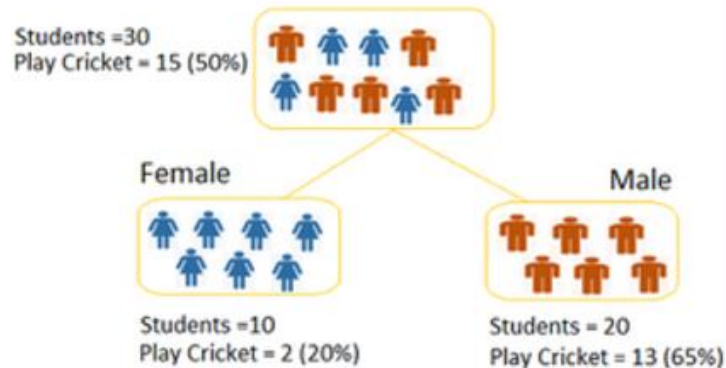


Figure 2.1. A plot of the hyperbolic curve relating f , the frequency of occurrence and r , the rank order (Adapted from Schultz⁴⁴ page 120)

Feature scoring metrics

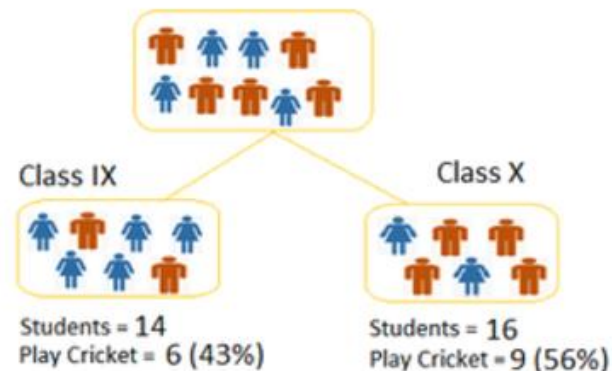
- Information gain
 - Decrease in entropy of categorical prediction when the feature is present v.s. absent

Split on Gender



class uncertainty decreases

Split on Class



class uncertainty intact

Feature scoring metrics

- Information gain
 - Decrease in entropy of categorical prediction when the feature is present or absent

$$IG(t) = - \sum_c p(c) \log p(c) + p(t) \sum_c p(c|t) \log p(c|t) + p(\bar{t}) \sum_c p(c|\bar{t}) \log p(c|\bar{t})$$

Entropy of class label along

Entropy of class label if t is present

Entropy of class label if t is absent

probability of seeing class label c in documents where t occurs

probability of seeing class label c in documents where t does not occur

Feature scoring metrics

- χ^2 statistics
 - Test whether distributions of two categorical variables are independent of one another
 - H_0 : they are independent
 - H_1 : they are dependent

	t	\bar{t}
c	A	B
\bar{c}	C	D

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

$DF(t)$ $N - DF(t)$ $\#Pos\ doc$ $\#Neg\ doc$

Feature scoring metrics

- χ^2 statistics
 - Test whether distributions of two categorical variables are independent of one another

- Degree of freedom = $(\#col-1) \times (\#row-1)$

- Significance level: α , i.e., $p\text{-value} < \alpha$

Look into χ^2 distribution table to find the threshold

	t	\bar{t}
c	36	30
\bar{c}	14	25

DF=1, $\alpha = 0.05 \Rightarrow$
threshold = 3.841



We cannot reject H_0



t is not a good feature to choose

$$\chi^2(t, c) = \frac{105(36 \times 25 - 14 \times 30)^2}{50 \times 55 \times 66 \times 39} = 3.418$$

Feature scoring metrics

- χ^2 statistics
 - Test whether distributions of two categorical variables are independent of one another
 - Degree of freedom = $(\#col-1) \times (\#row-1)$
 - Significance level: α , i.e., $p\text{-value} < \alpha$
 - For the features passing the threshold, rank them by descending order of χ^2 values and choose the top k features

Feature scoring metrics

- χ^2 statistics with multiple categories
 - $\chi^2(t) = \sum_c p(c) \chi^2(c, t)$
 - Expectation of χ^2 over all the categories
 - $\chi^2(t) = \max_c \chi^2(c, t)$
 - Strongest dependency between a category
- Problem with χ^2 statistics
 - Normalization breaks down for the very low frequency terms
 - χ^2 values become incomparable between high frequency terms and very low frequency terms

Distribution assumption becomes inappropriate in this test

Feature scoring metrics

- Many other metrics

- Mutual information

- Relatedness between term t and class c

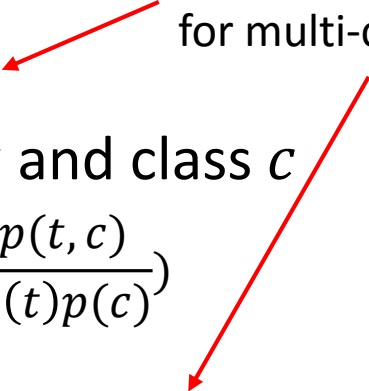
$$PMI(t; c) = p(t, c) \log\left(\frac{p(t, c)}{p(t)p(c)}\right)$$

- Odds ratio

- Odds of term t occurring with class c normalized by that without c

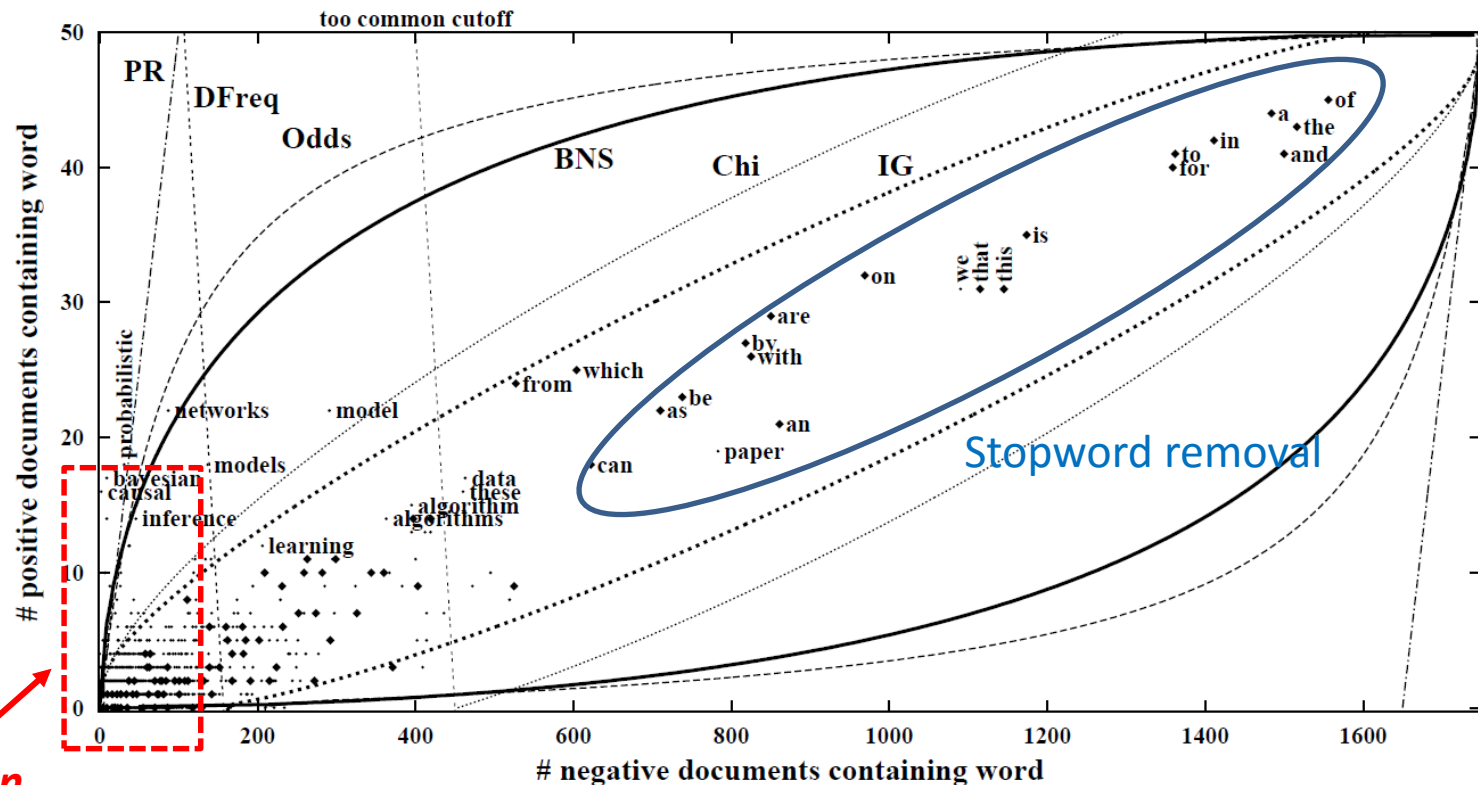
$$Odds(t; c) = \frac{p(t, c)}{1 - p(t, c)} \times \frac{1 - p(t, \bar{c})}{p(t, \bar{c})}$$

Same trick as in χ^2 statistics
for multi-class cases



A graphical analysis of feature selection

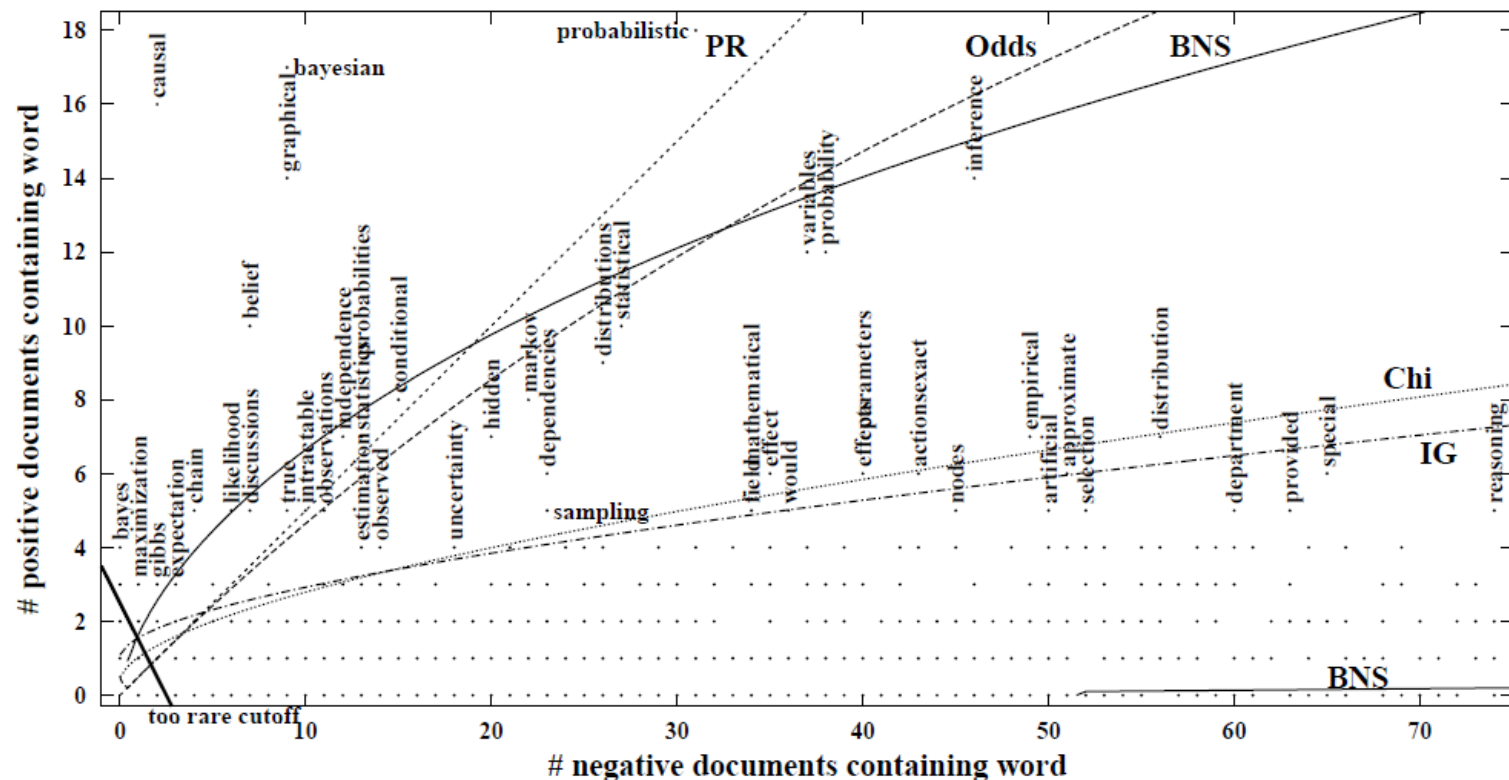
- Isoclines for each feature scoring metric
 - Machine learning papers v.s. other CS papers



Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. JMLR, 3, 1289-1305.

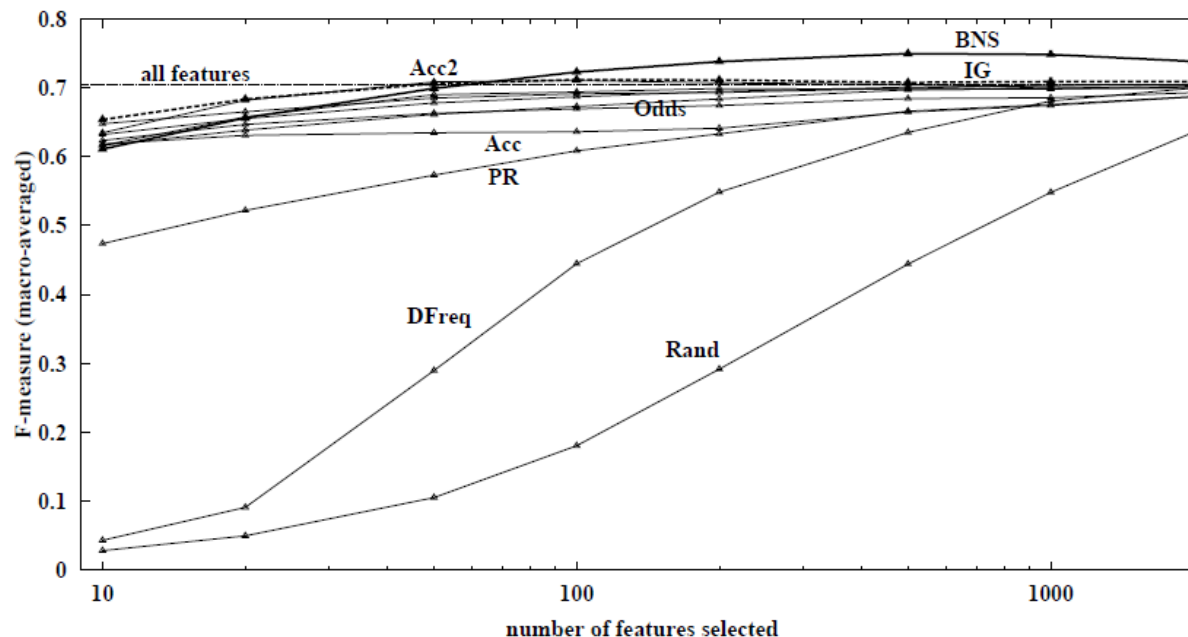
A graphical analysis of feature selection

- Isoclines for each feature scoring metric
 - Machine learning papers v.s. other CS papers



Effectiveness of feature selection methods

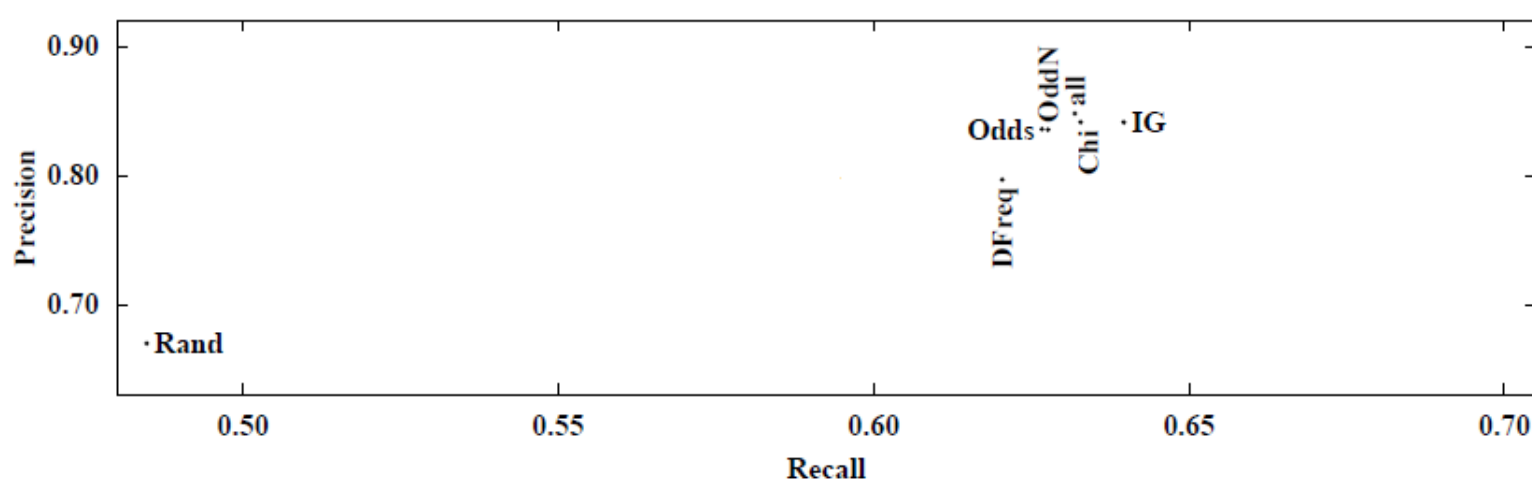
- On a multi-class classification data set
 - 229 documents, 19 classes
 - Binary feature, SVM classifier



Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. JMLR, 3, 1289-1305.

Effectiveness of feature selection methods

- On a multi-class classification data set
 - 229 documents, 19 classes
 - Binary feature, SVM classifier



Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. JMLR, 3, 1289-1305.

Empirical analysis of feature selection methods

- Text corpus
 - Reuters-22173
 - 13272 documents, 92 classes, 16039 unique words
 - OHSUMED
 - 3981 documents, 14321 classes, 72076 unique words
- Classifier: kNN and LLSF

Method	DF	IG	CHI	MI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	Y
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, [President Obama](#) walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky [State of the Union address](#) after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political
News



Sports
News



Entertainment
News



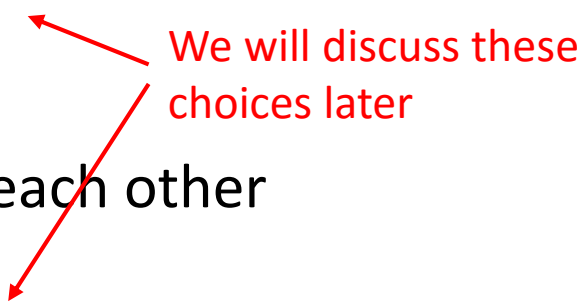
1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

2.1 What is the unique property of this problem?

2.2 What type of classifier we should use?

Model specification

- Specify dependency assumptions
 - Linear relation between x and y
 - $w^T x \rightarrow y$
 - Features are independent among each other
 - Naïve Bayes, linear SVM
 - Non-linear relation between x and y
 - $f(x) \rightarrow y$, where $f(\cdot)$ is a non-linear function of x
 - Features are not independent among each other
 - Decision tree, kernel SVM, mixture model
 - Choose based on our domain knowledge of the problem
- 

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, [President Obama](#) walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky [State of the Union address](#) after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political
News



Sports
News



Entertainment
News



1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

- 3.1 How to estimate the parameters in the selected model?
- 3.2 How to control the complexity of the estimated model?

Model estimation and selection

- General philosophy
 - Loss minimization

$$E[L] = L_{1,0}p(y=1) \int_{R_0} p(x)dx + L_{0,1}p(y=0) \int_{R_1} p(x)dx$$

Penalty when
misclassifying c_1 to c_0

Penalty when
misclassifying c_0 to c_1

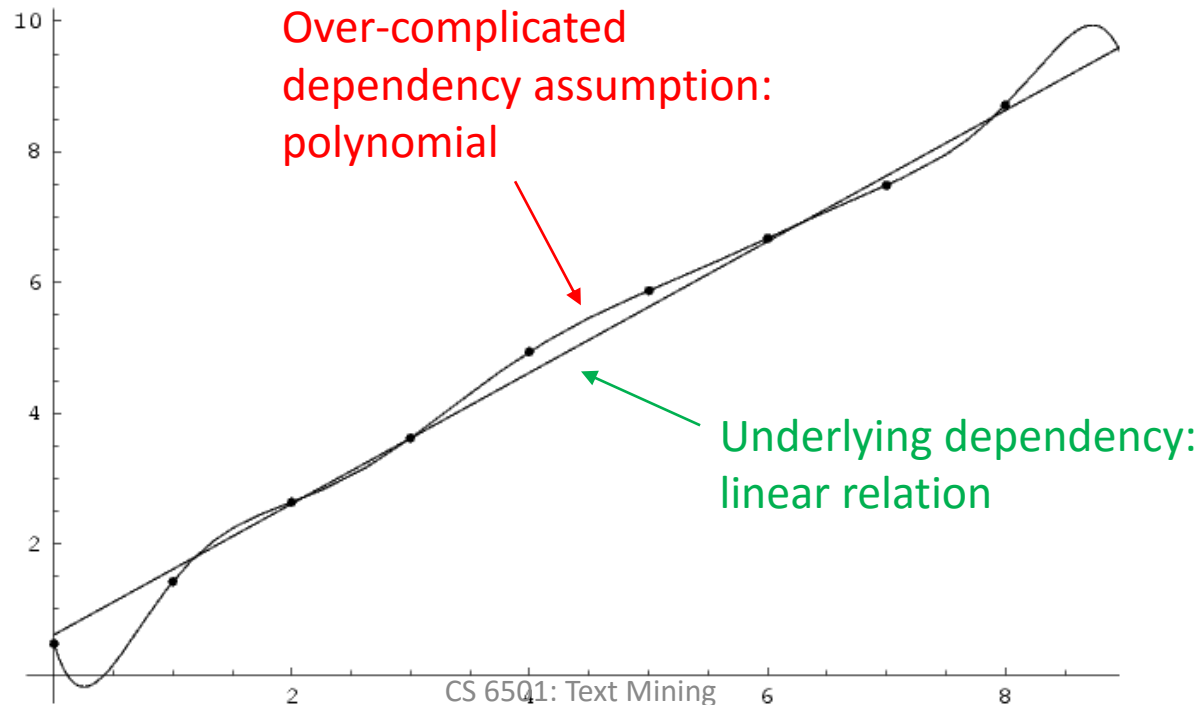
Empirically estimated from training set

Empirical loss!

Key assumption: Independent and Identically Distributed!

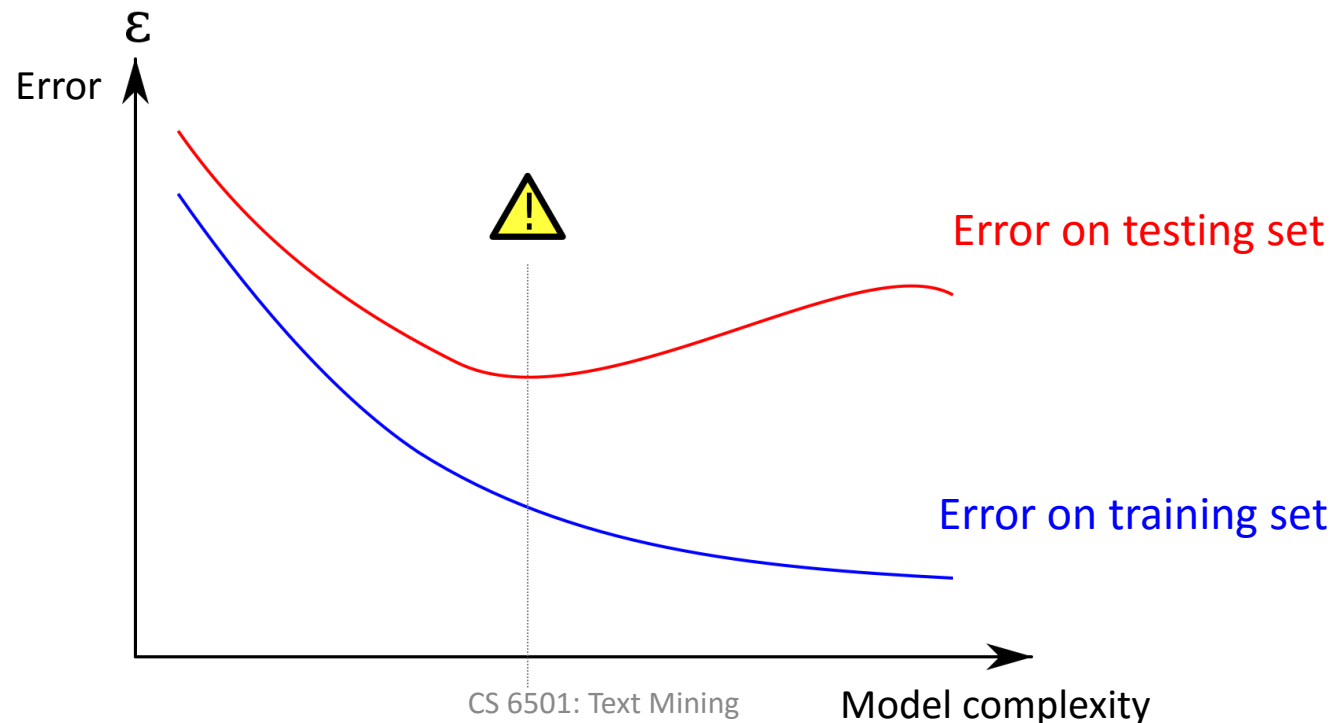
Empirical loss minimization

- Overfitting
 - Good empirical loss, terrible generalization loss
 - High model complexity -> prone to overfit noise



Generalization loss minimization

- Avoid overfitting
 - Measure model complexity as well
 - Model selection and regularization

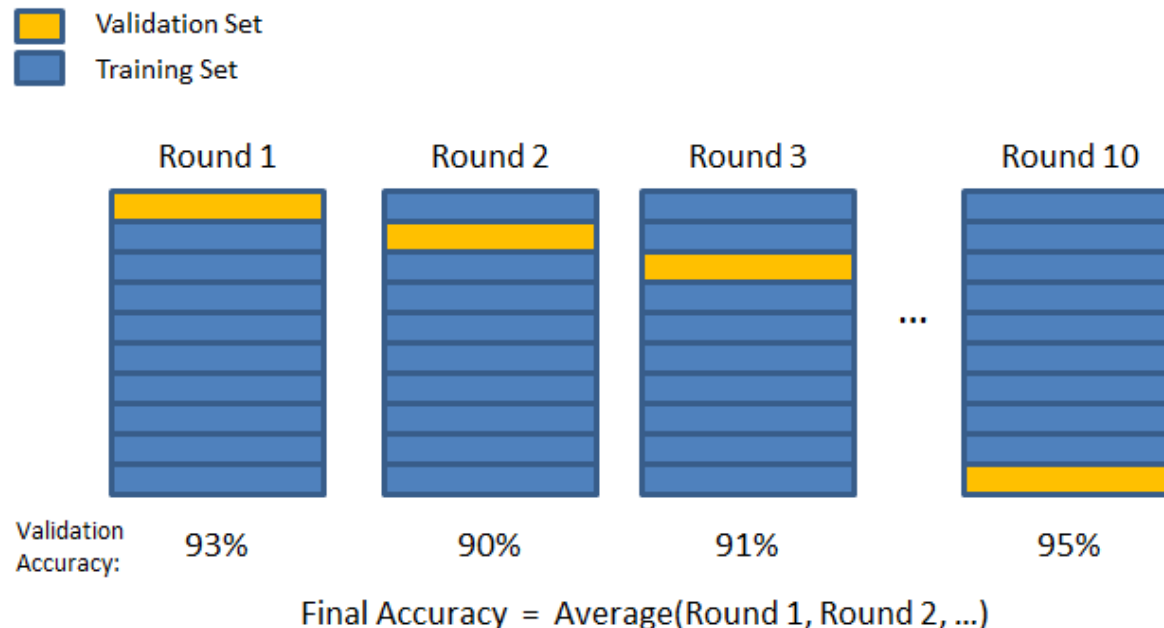


Generalization loss minimization

- Cross validation
 - Avoid noise in train/test separation
 - k -fold cross-validation
 1. Partition all training data into k equal size disjoint subsets;
 2. Leave one subset for validation and the other $k-1$ for training;
 3. Repeat step (2) k times with each of the k subsets used exactly once as the validation data.

Generalization loss minimization

- Cross validation
 - Avoid noise in train/test separation
 - k -fold cross-validation



Generalization loss minimization

- Cross validation
 - Avoid noise in train/test separation
 - k -fold cross-validation
 - Choose the model (among different models or same model with different settings) that has the best average performance on the validation sets
 - Some statistical test is needed to decide if one model is significantly better than another

Will cover it shortly

General steps for text categorization

POLITICS | WHITE HOUSE MEMO

Gloom Lifts, and Obama Goes All Out

By MICHAEL D. SHEAR JAN. 21, 2015

Email

Share

Tweet

Save

More

WASHINGTON — The morning after major Democratic losses in last year's midterm elections, [President Obama](#) walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

On Tuesday, Mr. Obama offered an estimated 30 million viewers a glimpse of that attitude when he delivered a self-assured, almost cocky [State of the Union address](#) after a year in which current and former White House advisers said he was often frustrated and at times discouraged.



Obama's Zinger in State of Union Address
Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.



Political
News



Sports
News



Entertainment
News



1. Feature construction and selection
2. Model specification
3. Model estimation and selection
4. Evaluation

Consider:

4.1 How to judge the quality of learned model?

4.2 How can you further improve the performance?

Classification evaluation

- Accuracy
 - Percentage of correct prediction over all predictions, i.e., $p(y^* = y)$
 - Limitation
 - Highly skewed class distribution
 - $p(y^* = 1) = 0.99$
 - » Trivial solution: all testing cases are positive
 - Classifiers' capability is only differentiated by 1% testing cases

Evaluation of binary classification

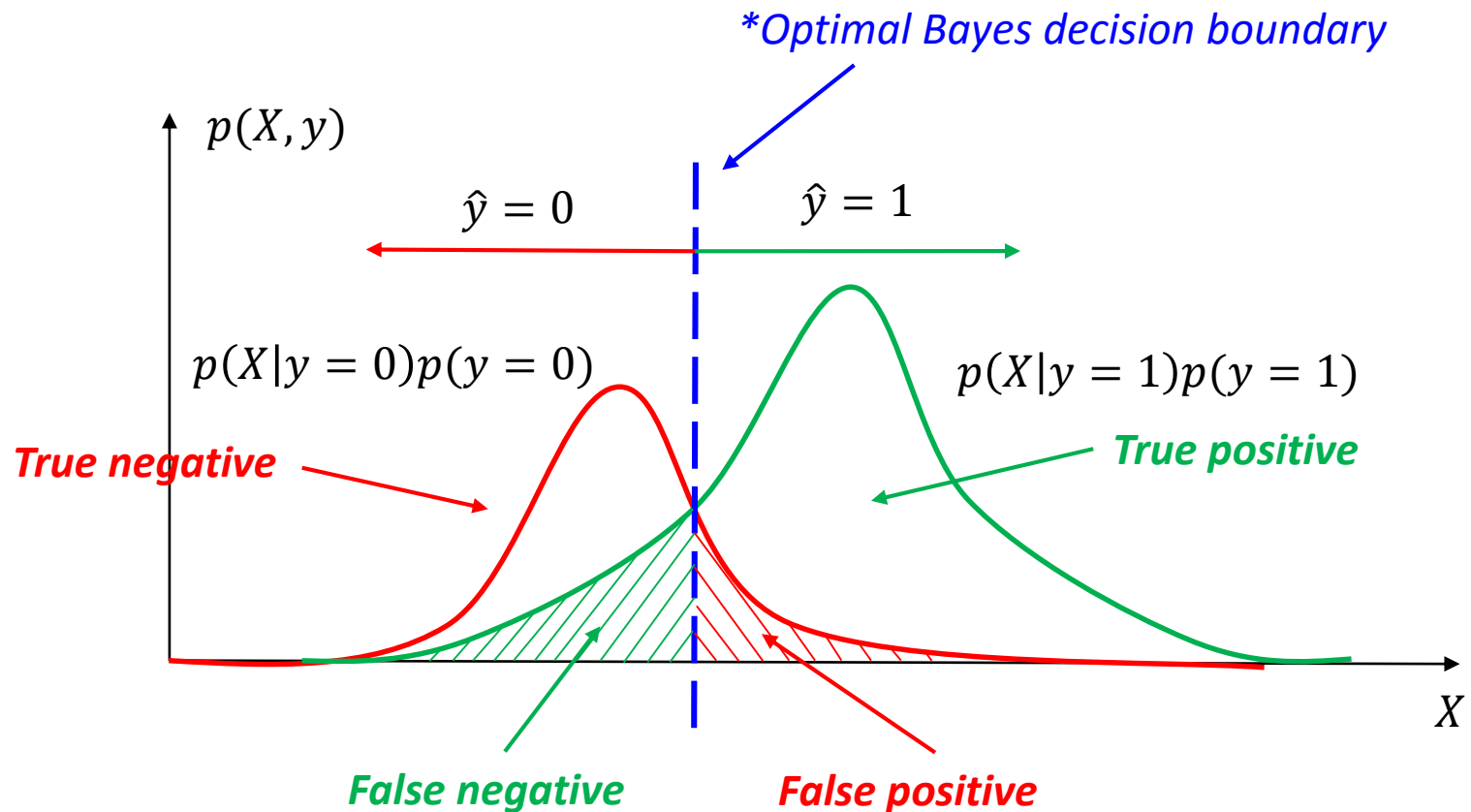
- Precision
 - Fraction of predicted positive documents that are indeed positive, i.e., $p(y^* = 1|y = 1)$
- Recall
 - Fraction of positive documents that are predicted to be positive, i.e., $p(y = 1|y^* = 1)$

	$y^* = 1$	$y^* = 0$
$y = 1$	true positive (TP)	false positive (FP)
$y = 0$	false negative (FN)	true negative (TN)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Evaluation of binary classification



Precision and recall trade off

- Precision decreases as the number of documents predicted to be positive increases

(ur kee	No.	Approach	Precision		Recall		all
			AVG	STD	AVG	STD	
	1	Triple-S	0.31	0.19	0.36	0.26	
• The	2	BP Graph Matching	0.60	0.45	0.19	0.30	
pe	3	RefMod-Mine/NSCM	0.37	0.22	0.39	0.27	
	4	RefMod-Mine/ESGM	0.16	0.26	0.12	0.21	
–	5	Bag-of-Words Similarity	0.56	0.23	0.32	0.28	
	6	PMLM	0.12	0.05	0.58	0.20	er
(7	ICoP	0.36	0.24	0.37	0.26	

- Recall: prefers a classifier to recognize more documents

Summarizing precision and recall

- With a single value
 - In order to compare different classifiers
 - F-measure: weighted harmonic mean of precision and recall, α balances the trade-off

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} \quad \left(F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} \right)$$

- Why harmonic mean?

- Classifier1: P:0.53, R:0.36
- Classifier2: P:0.01, R:0.99

*Equal weight between
precision and recall*

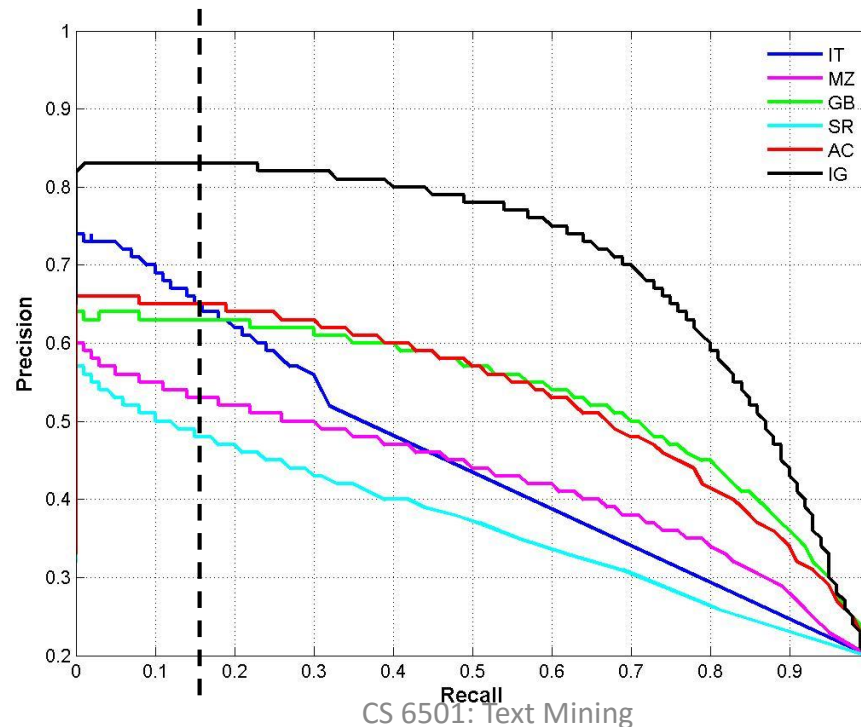
H	A
0.429	0.445
0.019	0.500

Summarizing precision and recall

- With a curve
 1. Order all the testing cases by the classifier's prediction score (assuming the higher the score is, the more likely it is positive);
 2. Scan through each testing case: treat all cases above it as positive (including itself), below it as negative; compute precision and recall;
 3. Plot precision and recall computed for each testing case in step (2).

Summarizing precision and recall

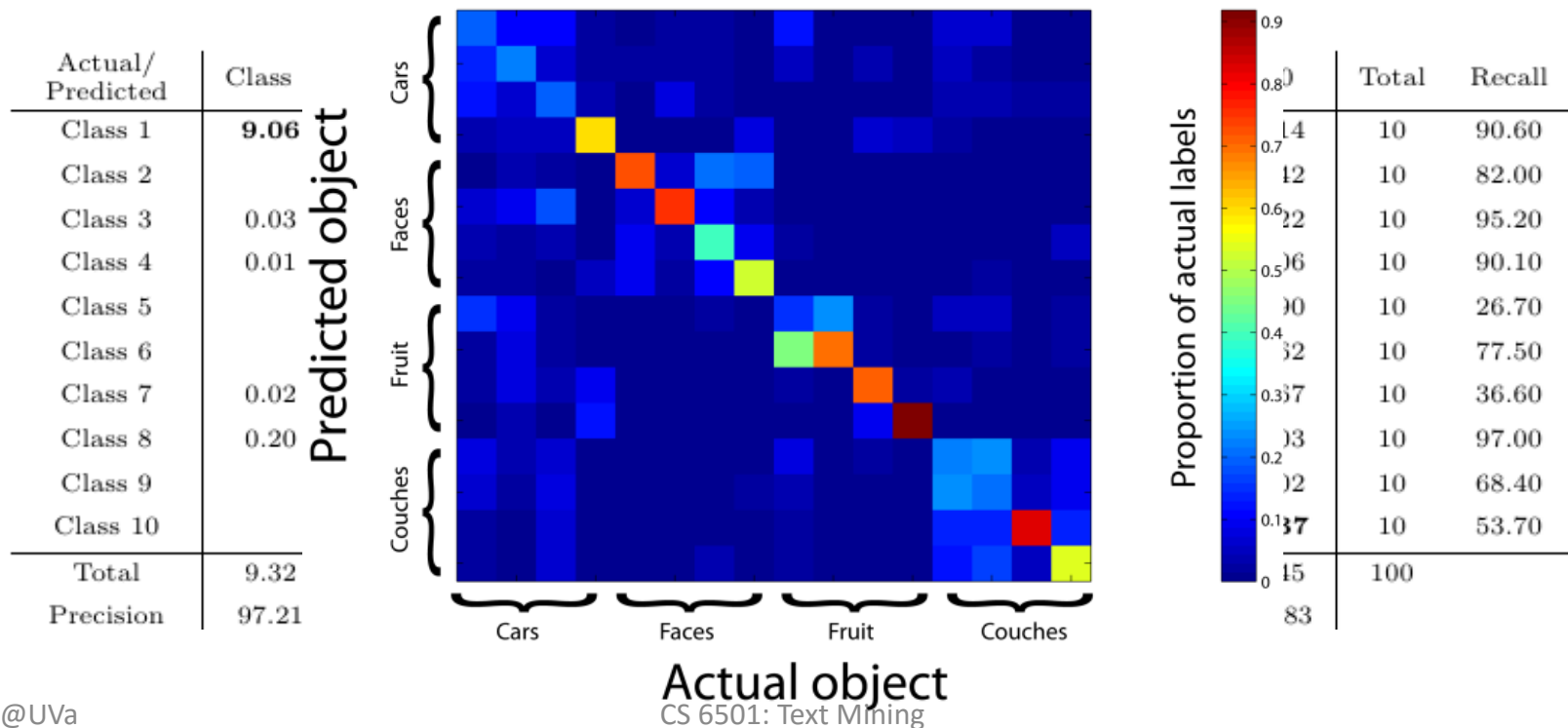
- With a curve
 - A.k.a., precision-recall curve
 - Area Under Curve (AUC)



Under each recall level, we prefer a higher precision

Multi-class categorization

- Confusion matrix
 - A generalized contingency table for precision and recall



Statistical significance tests

- How confident you are that an observed difference doesn't simply result from the train/test separation you chose?

<u>Fold</u>	<u>Classifier 1</u>	<u>Classifier 2</u>
1	0.20	0.18
2	0.21	0.19
3	0.22	0.21
4	0.19	0.20
5	0.18	0.37
Average	0.20	0.23

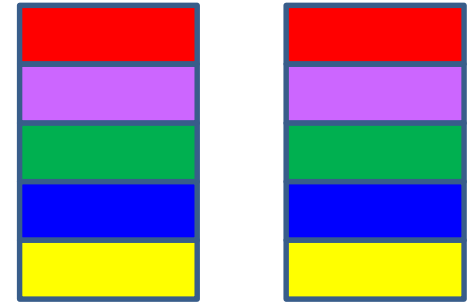
Background knowledge

- p -value in statistic test is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (e.g., if observation is totally random)
- If p -value is smaller than the chosen significance level (α), we reject the null hypothesis (e.g., observation is not random)
- We seek to reject the null hypothesis (we seek to show that the observation is not a random result), and so small p -values are good

Paired t -test

Algorithm 1

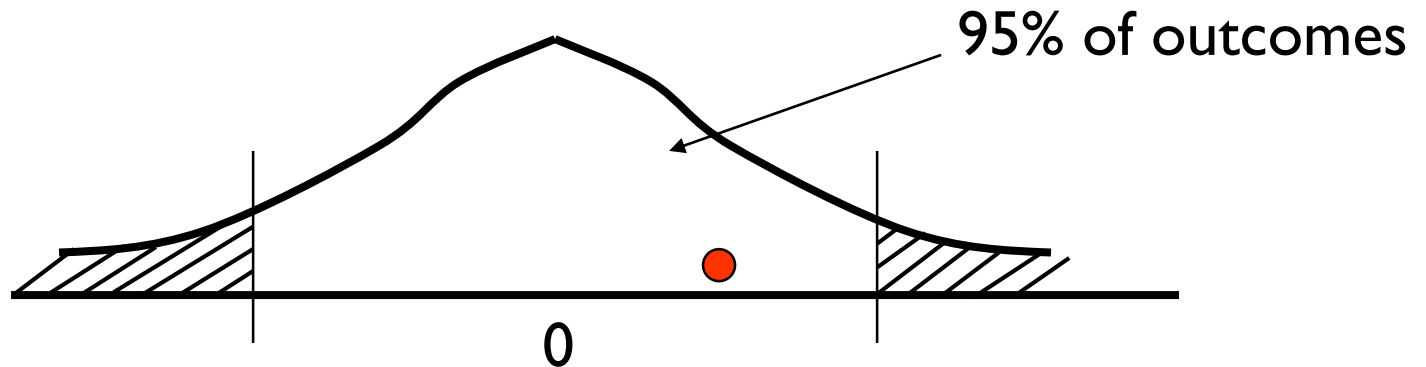
Algorithm 2



- Paired t -test
 - Test if two sets of observations are significantly different from each other
 - On k -fold cross validation, different classifiers are applied onto the same train/test separation
 - Hypothesis: difference between two responses measured on the same statistical unit has a zero mean value
- One-tail v.s. two-tail?
 - If you aren't sure, use two-tail

Statistical significance test

<u>Fold</u>	<u>Classifier A</u>	<u>Classifier B</u>	<u>paired t-test</u>
1	0.20	0.18	+0.02
2	0.21	0.19	+0.02
3	0.22	0.21	+0.01
4	0.19	0.20	-0.01
5	0.18	0.37	-0.19
Average	0.20	0.23	$p=0.4987$



What you should know

- Bayes decision theory
 - Bayes risk minimization
- General steps for text categorization
 - Text feature construction
 - Feature selection methods
 - Model specification and estimation
 - Evaluation metrics

Today's reading

- Introduction to Information Retrieval
 - Chapter 13: Text classification and Naive Bayes
 - 13.1 – Text classification problem
 - 13.5 – Feature selection
 - 13.6 – Evaluation of text classification