# 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

Access

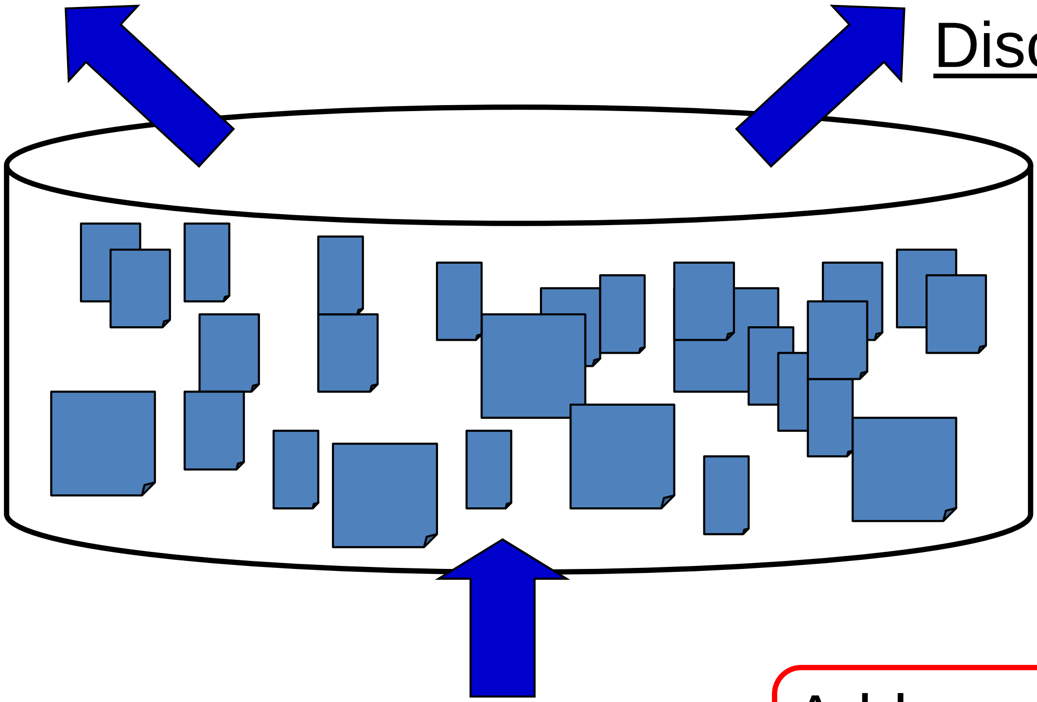
**Serve for IR applications**

<u>Filter</u> information

**Sub-area of DM research**

Mining

<u>Discover</u> knowledge



**Based on NLP/ML techniques**

Organization

<u>Add</u> Structure/Annotations

# 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.


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

"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.

**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.

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

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

# 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>

<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>
```

Spam=**True**/False

# 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 no 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.

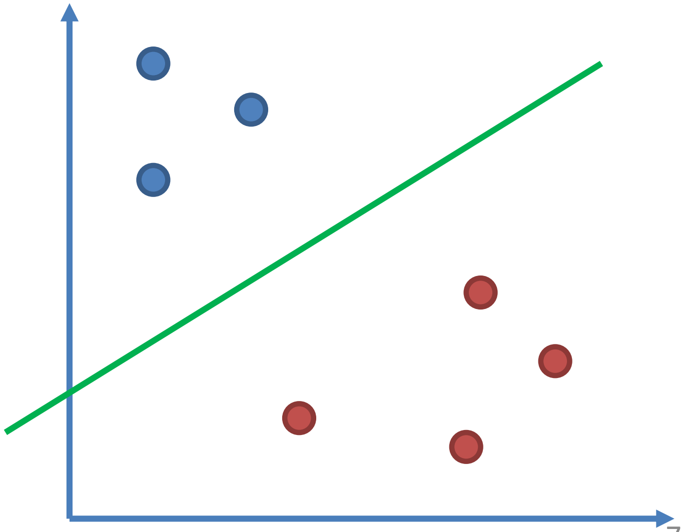
# Basic notions about categorization

- Data points/Instances
  - $X$: a $m$-dimensional feature vector

*vector space representation*

- Labels
  - $y$: a <u>categorical</u> value from $\{0, \dots, k-1\}$
- Classification hyper-plane
  - $f(X) \to y$

***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

$$\hat{y} = argmax_y p(y|X) = argmax_y \frac{p(X|y)p(y)}{p(X)}$$

Constant with respect to $y$

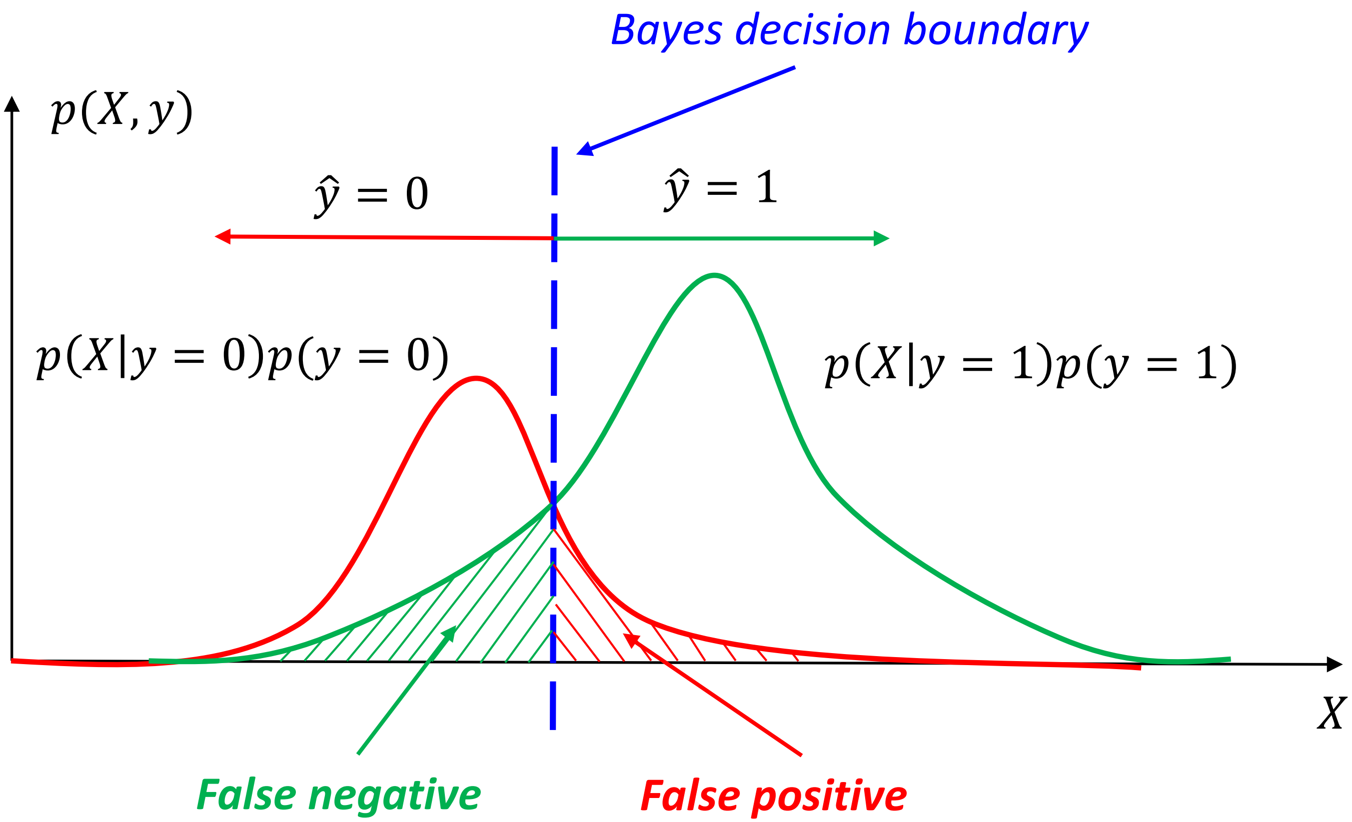
  – Example in binary classification
    - $\hat{y} = 1$, if $p(X|y = 1)p(y = 1) > p(X|y = 0)p(y = 0)$
    - $\hat{y} = 0$, otherwise
- This leads to <u>optimal</u> classification result
  – Optimal in the sense of 'risk' minimization

# 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'

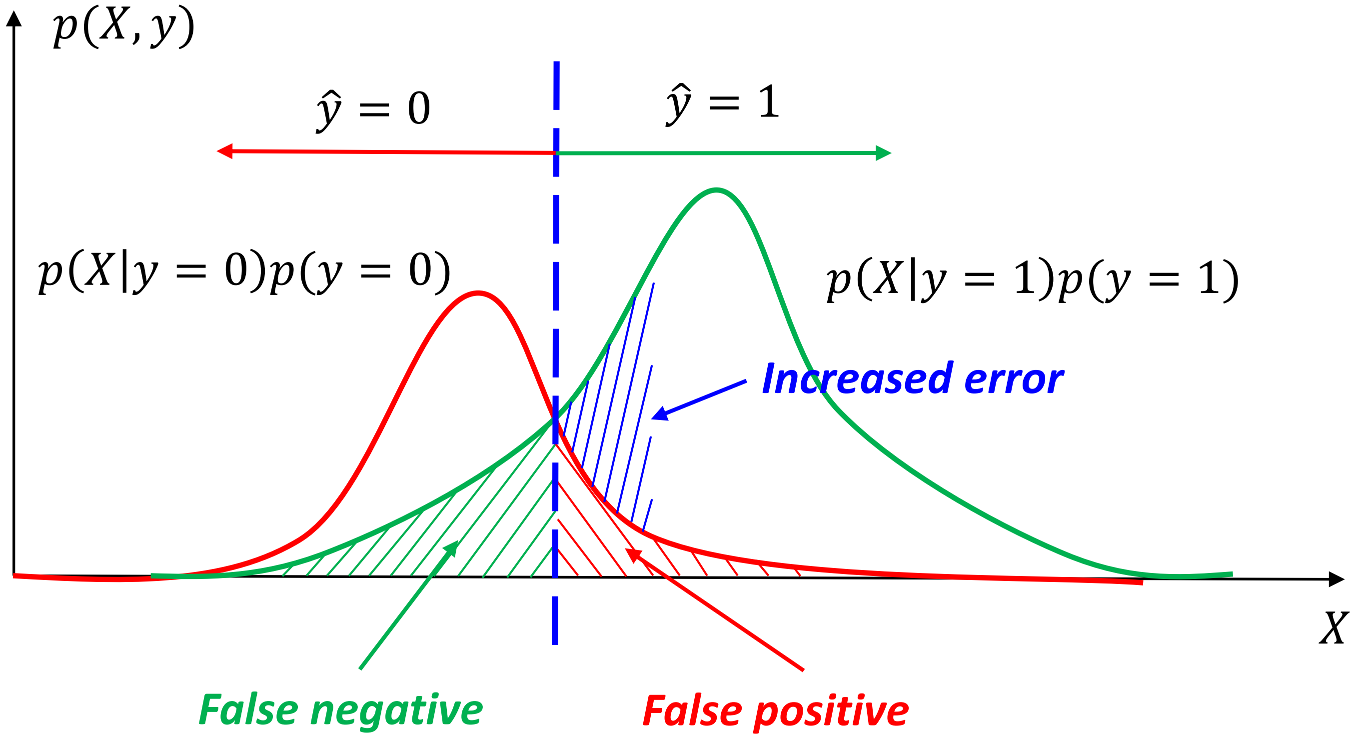
# Bayes risk

- Risk – assign instance to a wrong class



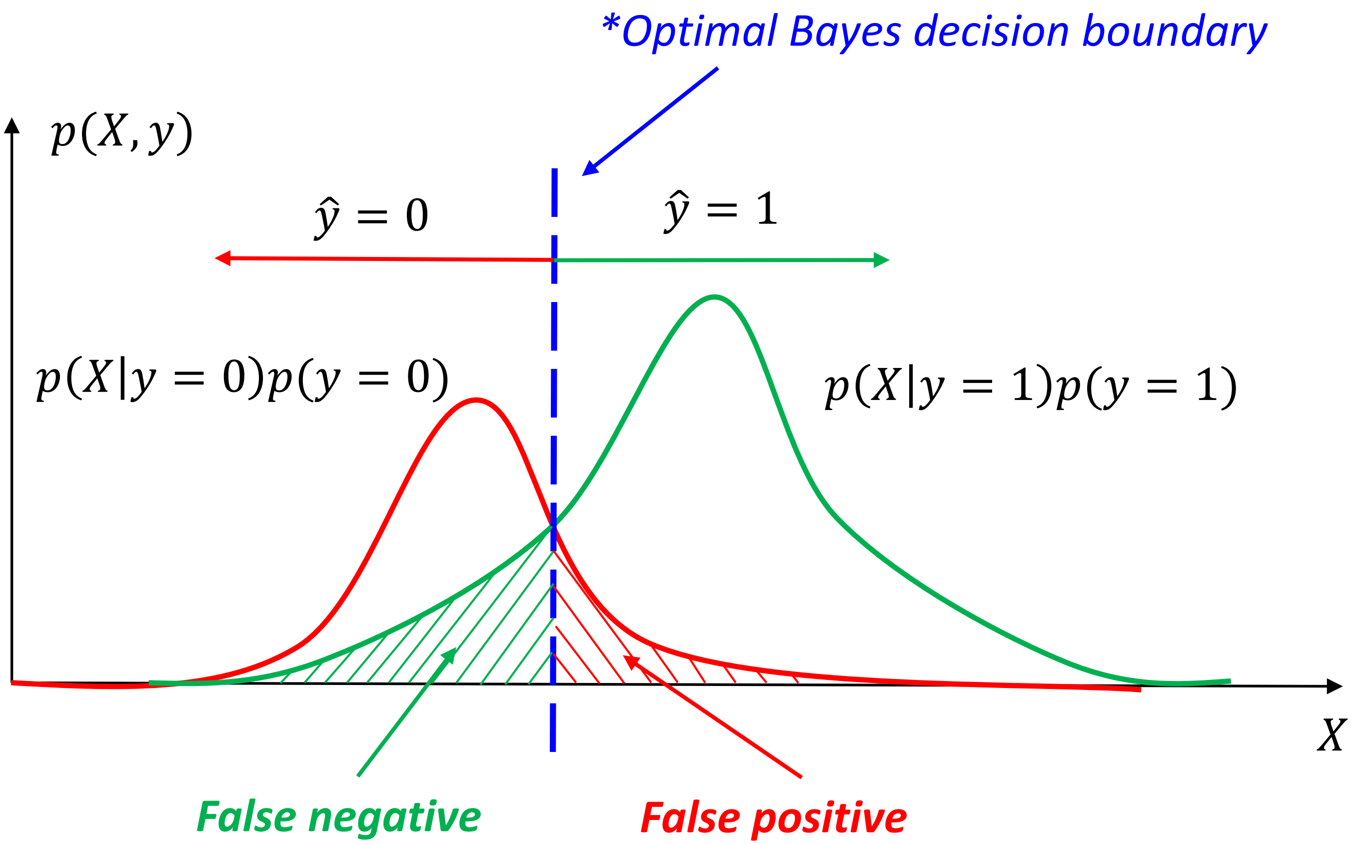
# Bayes risk

- Risk – assign instance to a wrong class

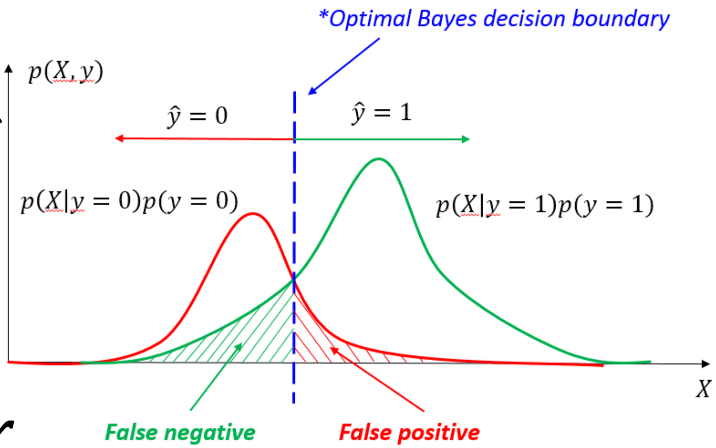


# Bayes risk

- Risk – assign instance to a wrong class



# Bayes risk




- Expected risk

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

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

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

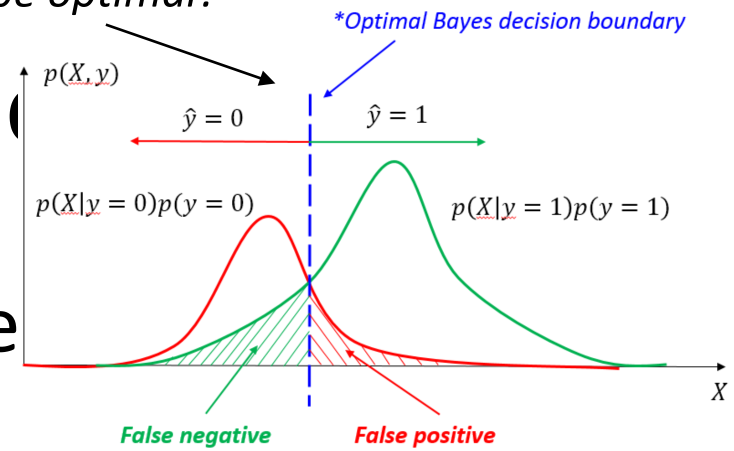
Region where we assign $x$ to class 0

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

Region where we assign $x$ to class 1

*Will the error of assigning $c_1$ to $c_0$ be always <u>equal</u> to the error of assigning $c_0$ to $c_1$?*

Will this still be optimal?




# Loss functi

- The penalty we will pay whe instances

Region where we assign $x$ to class 0

Penalty when misclassifying $c_1$ to $c_0$

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

Region where we assign $x$ to class 1

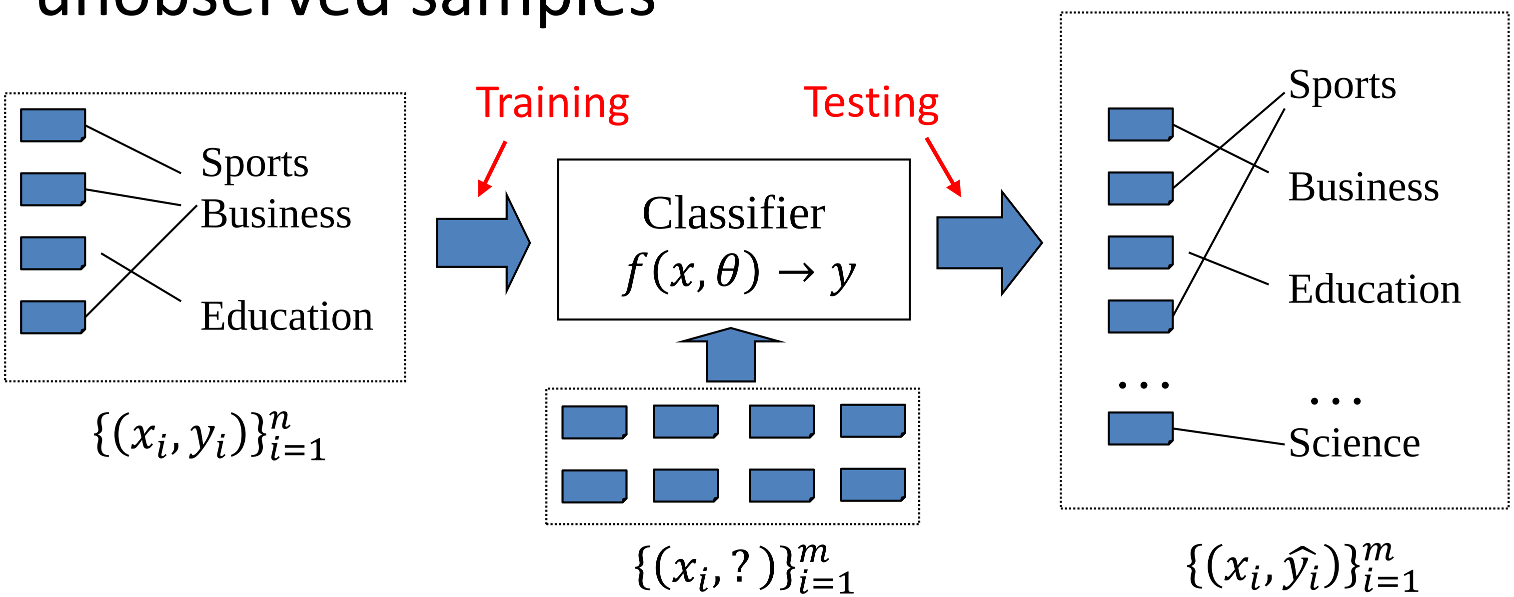
Penalty when misclassifying $c_0$ to $c_1$

$$+L_{0,1}p(y=0)\int_{R_1} p(x)p(x|y=1)dx$$

- Goal of classification in general
  - Minimize loss

# Supervised text categorization

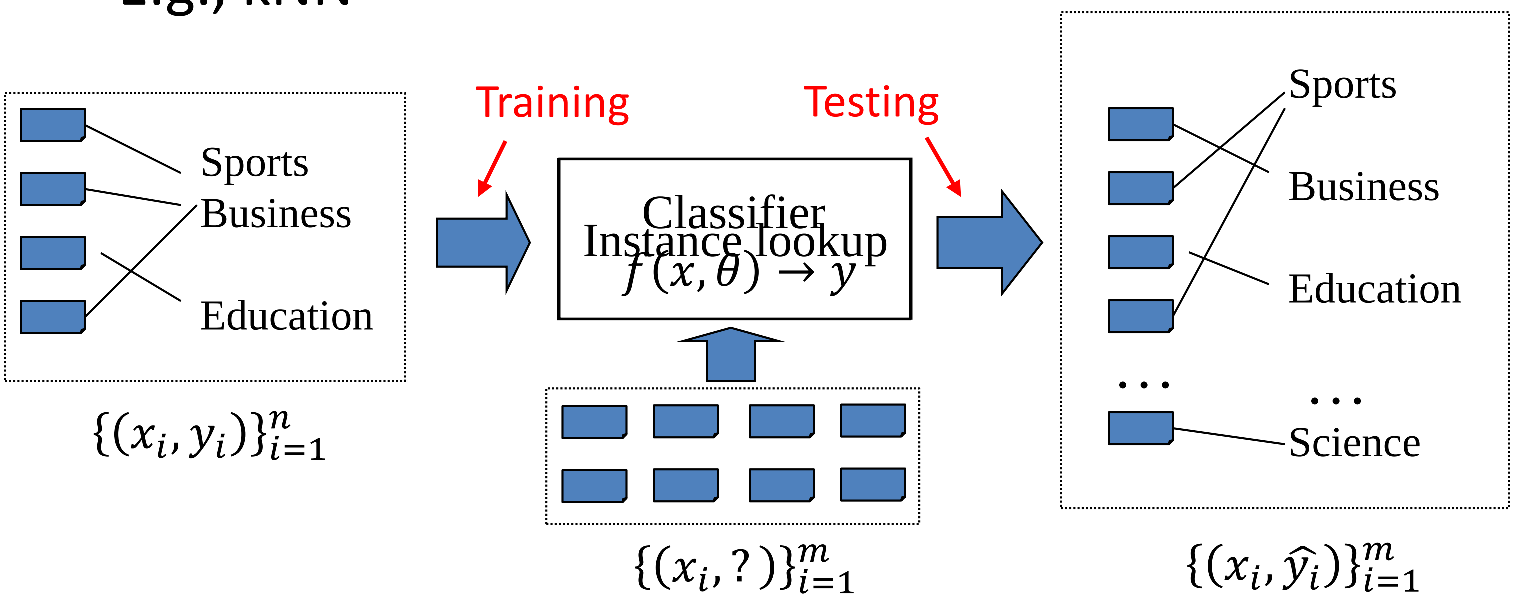
- Supervised learning
  - Estimate a model/method from labeled data
  - It can then be used to determine the labels of the unobserved samples



# Type of classification methods

- Model-less
  - Instance based classifiers
    - Use observation directly
    - E.g., kNN

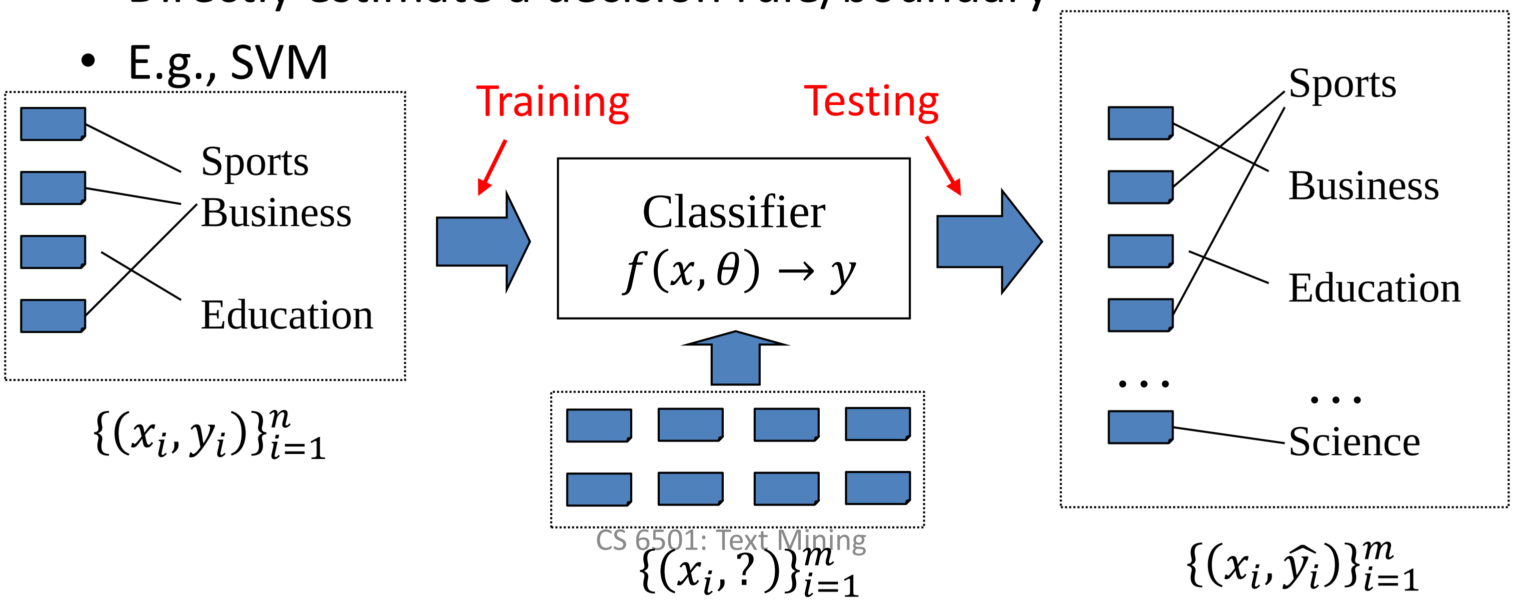
***Key: assuming similar items have similar class labels!***



# Type of classification methods

- Model-based
  - Generative models
    - Modeling joint probability of $p(x, y)$
    - E.g., Naïve Bayes
  - Discriminative models
    - Directly estimate a decision rule/boundary
    - E.g., SVM

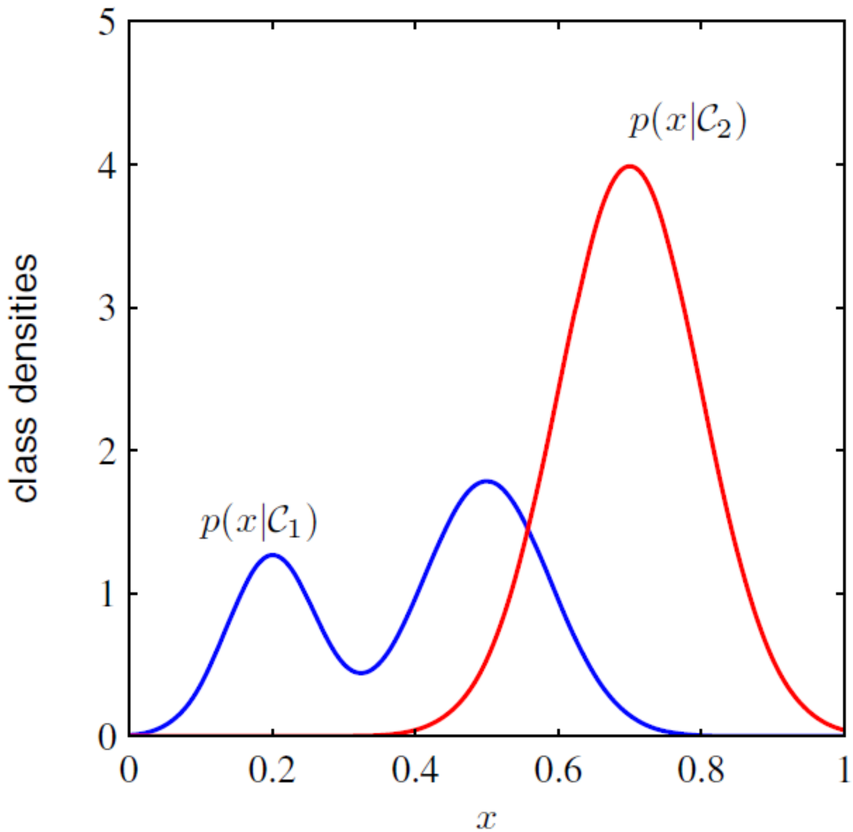
***Key: i.i.d. assumption!***



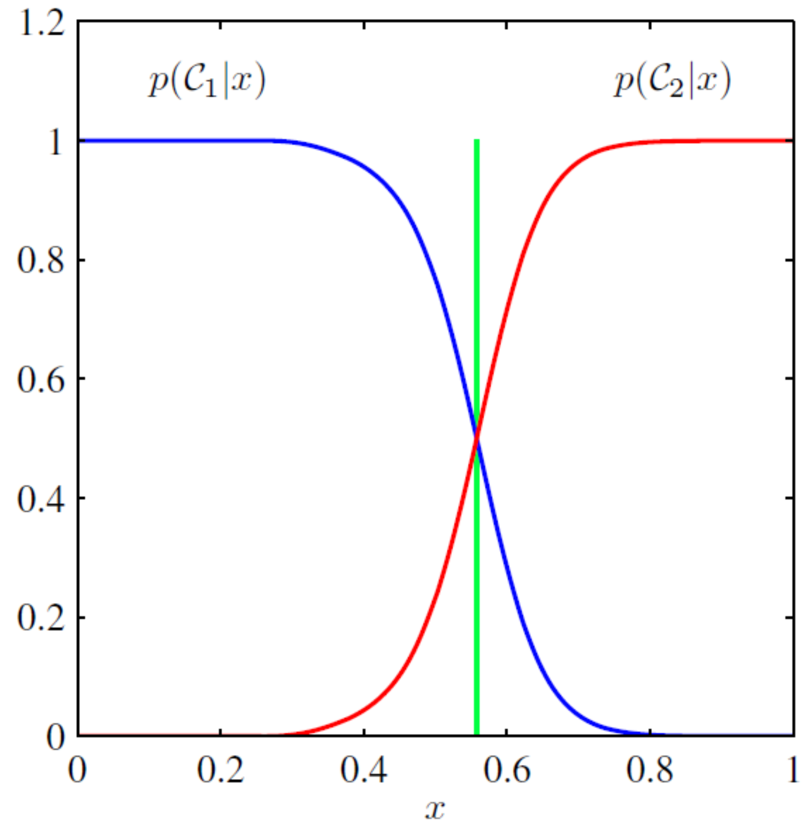
# Generative V.S. discriminative models

- 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

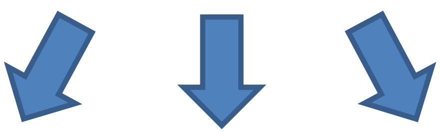
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"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.

PLAY VIDEO

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

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.



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

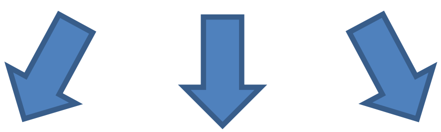
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"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.

PLAY VIDEO

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

*Very sparse feature representation!*

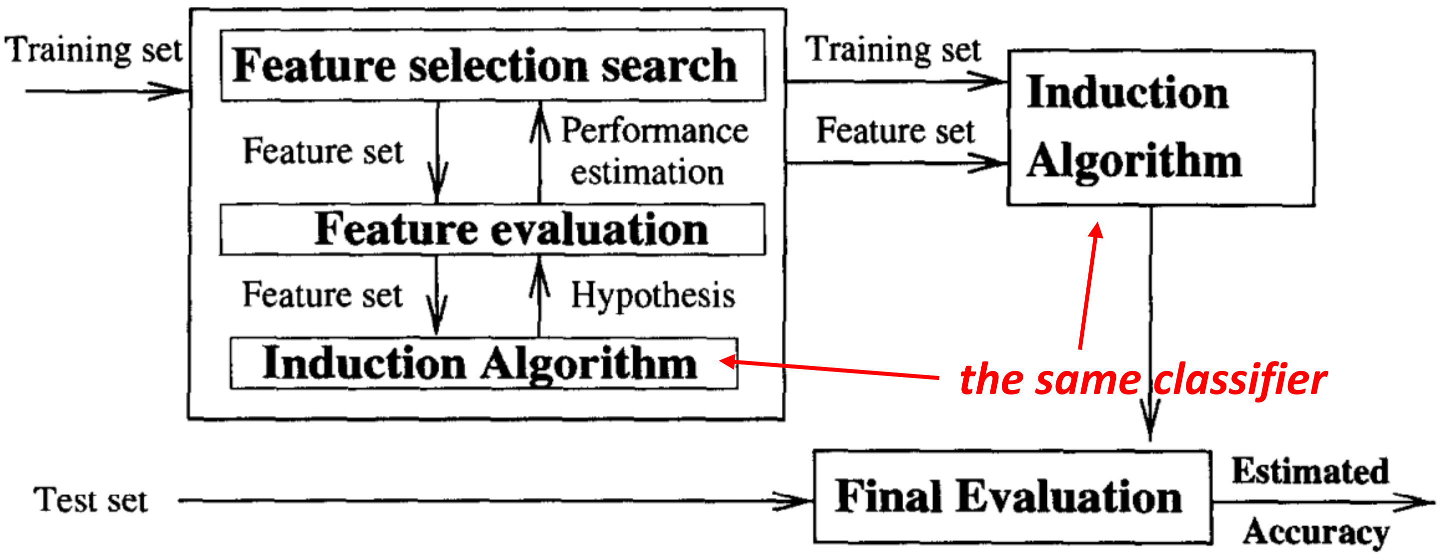
- How many review documents do we have there for training?
  - 629K Yelp_small

# 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

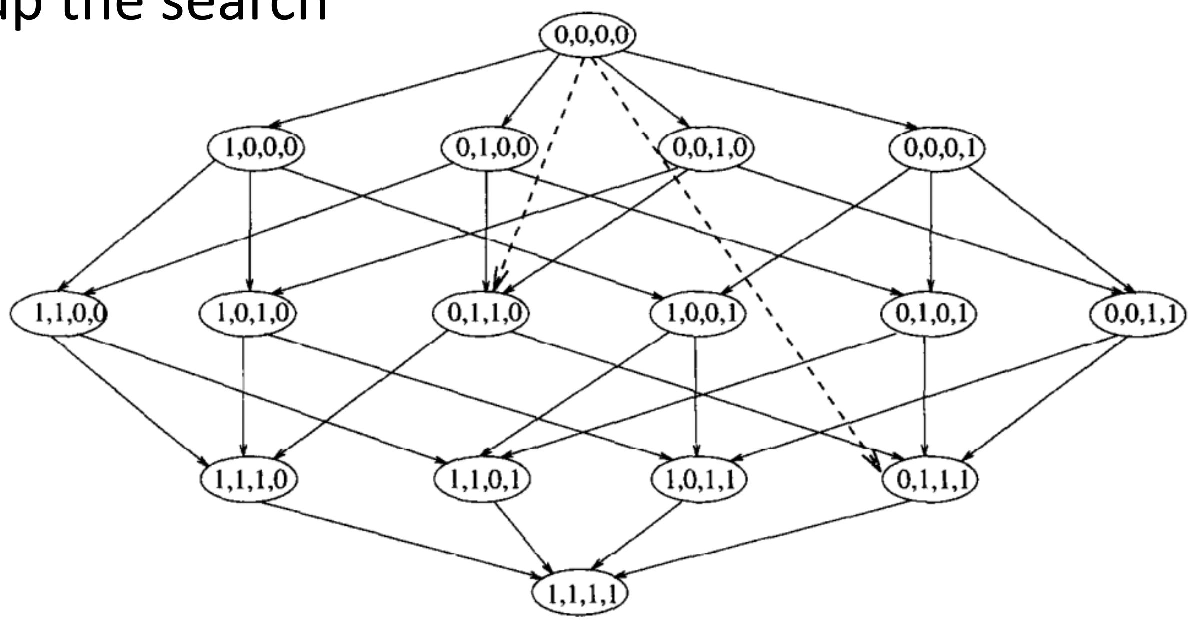
- Wrapper method
  - Find the best subset of features for a particular classification method




*R. Kohavi, G.H. John/Artijicial 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



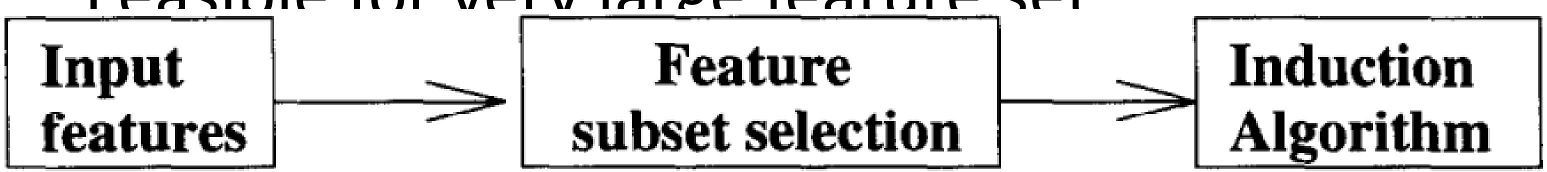

*R. Kohavi, G.H. John/Artijicial Intelligence 97 (1997) 273-324*

# 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/Artijicial Intelligence 97 (1997) 273-324*

# Feature scoring metrics

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

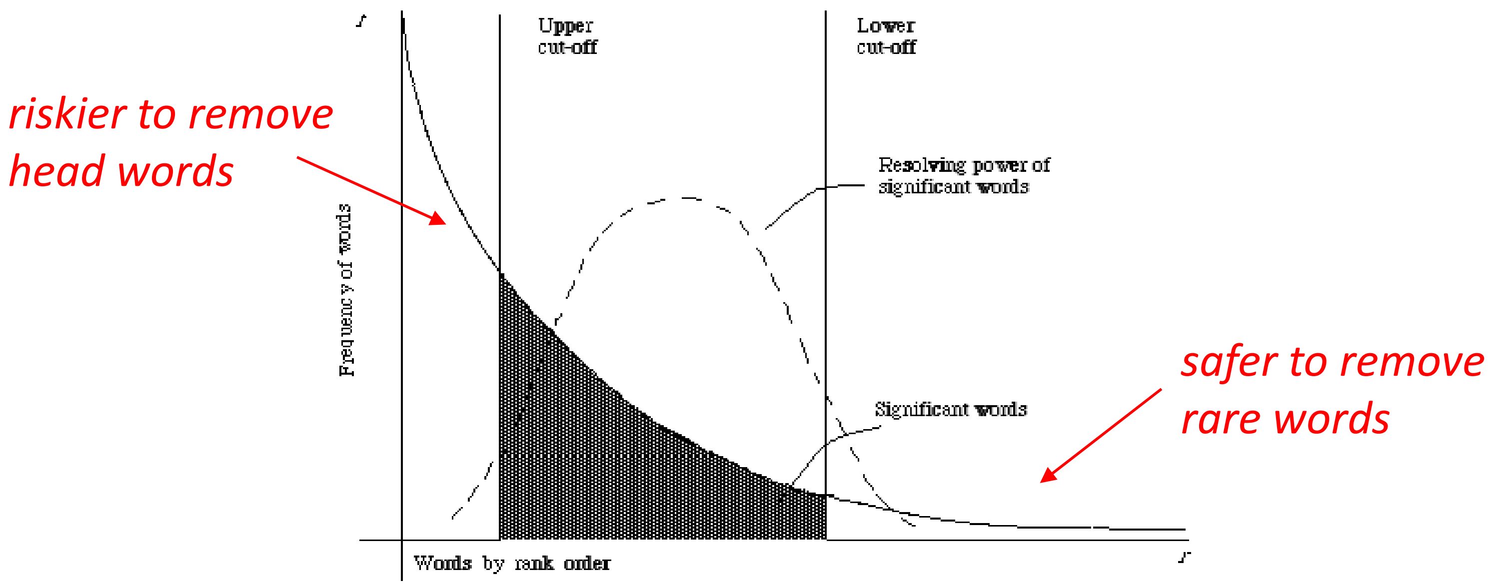
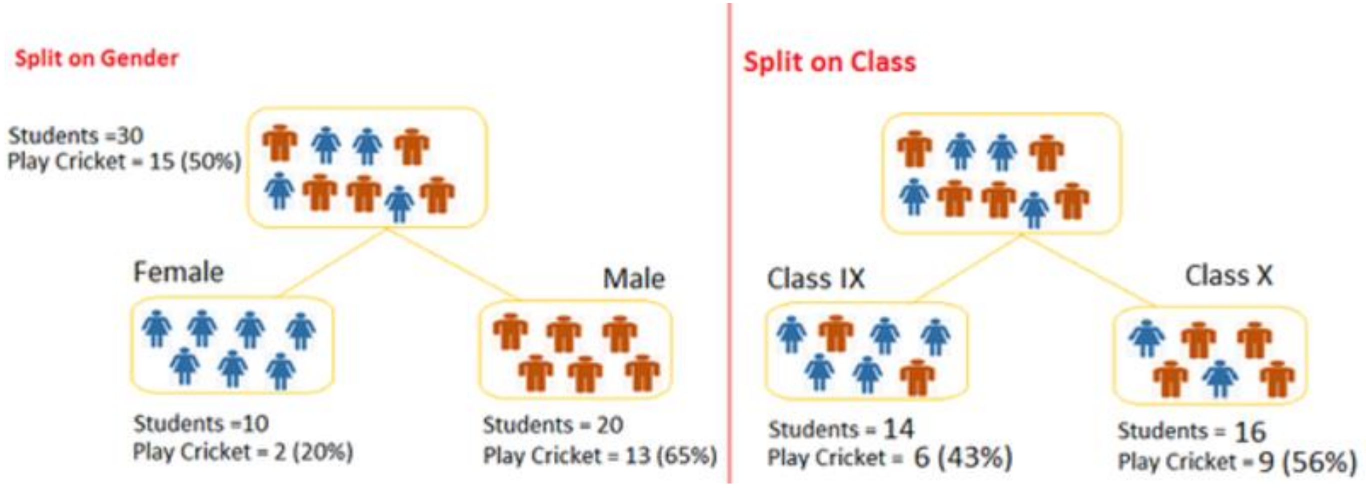



*Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)*

# Feature scoring metrics

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




class uncertainty decreases class uncertainty intact

# Feature scoring metrics

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

$$IG(t) = -\sum_c p(c)\log p(c)$$

Entropy of class label along

$$+p(t)\sum_c p(c|t)\log p(c|t)$$

Entropy of class label if $t$ is <u>present</u>

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

Entropy of class label if $t$ is <u>absent</u>

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

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

| | $t$ | $\bar{t}$ |
|---|---|---|
| $c$ | $A$ | $B$ |
| $\bar{c}$ | $C$ | $D$ |

$$\chi^2(t,c) = \frac{(A+B+C+D)(AD-BC)^2}{\underbrace{(A+C)}_{DF(t)}\underbrace{(B+D)}_{N-DF(t)}\underbrace{(A+B)}_{\#Pos\ doc}\underbrace{(C+D)}_{\#Neg\ doc}}$$

# Feature scoring metrics

- $\chi^2$ statistics
  - Test whether distributions of two categorical variables are independent of one another
    - Degree of freedom = (#col-1)(#row-1)
    - Significance level: $\alpha$, i.e., $p$-value<$\alpha$

Look into $\chi^2$ distribution table to find the threshold

| | $t$ | $\bar{t}$ |
|---|---|---|
| $c$ | 36 | 30 |
| $\bar{c}$ | 14 | 25 |

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

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

We cannot reject $H_0$

$t$ is not a good feature to choose

# Feature scoring metrics

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

# Feature scoring metrics

- $\chi^2$ statistics with multiple categories
  - $\chi^2(t) = \sum_c p(c)\chi^2(c,t)$
    - Expectation of $\chi^2$ over all the categories
  - $\chi^2(t) = \max_c \chi^2(c,t)$
    - Strongest dependency between a category
- Problem with $\chi^2$ statistics
  - Normalization $\boxed{\text{breaks down}}$ for the very low frequency terms
    - $\chi^2$ 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 $c$

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

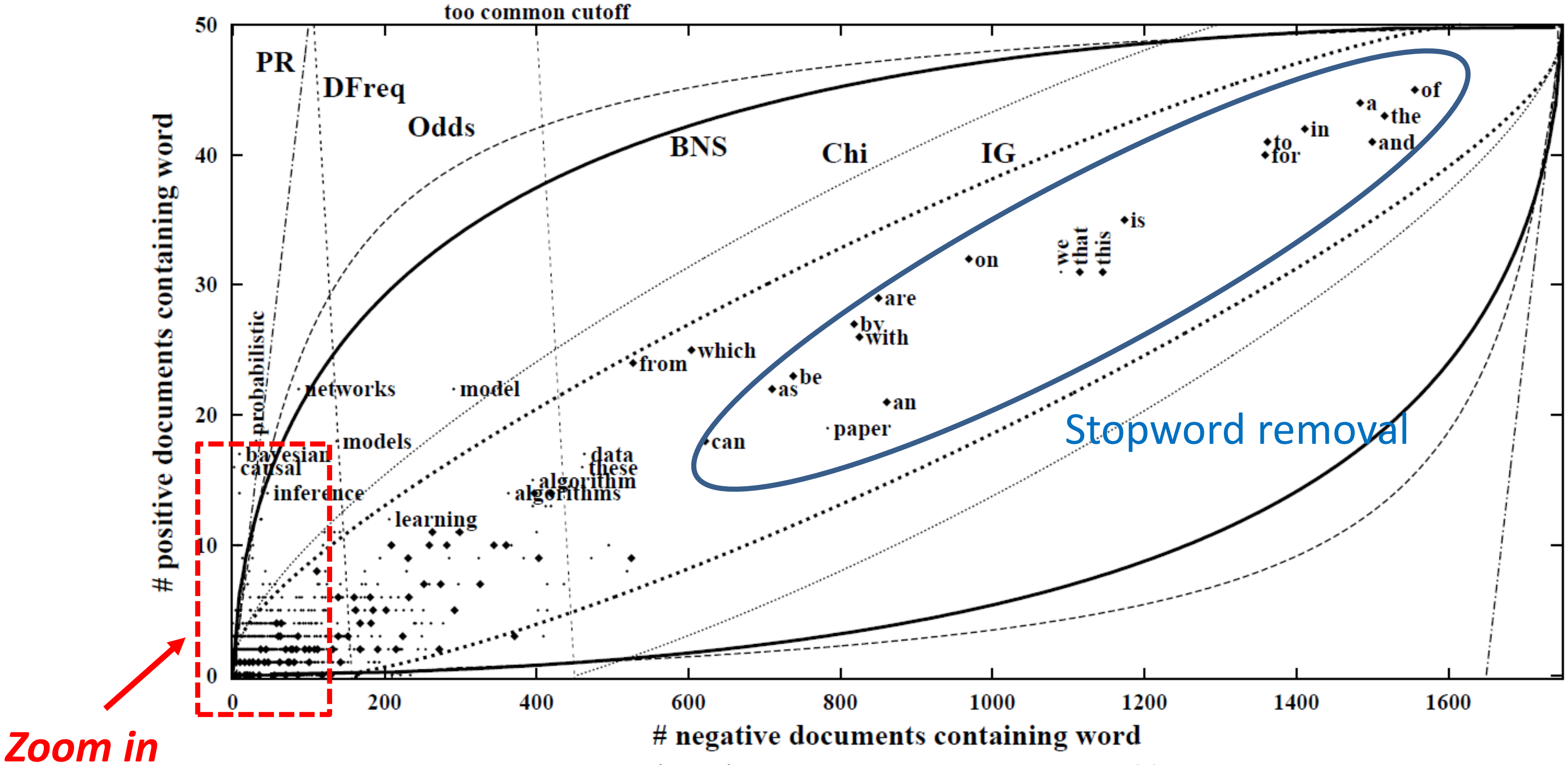
  - Odds ratio
    - Odds of term $t$ occurring with class $c$ normalized by that without $c$

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

Same trick as in $\chi^2$ statistics for multi-class cases

# A graphical analysis of feature selection

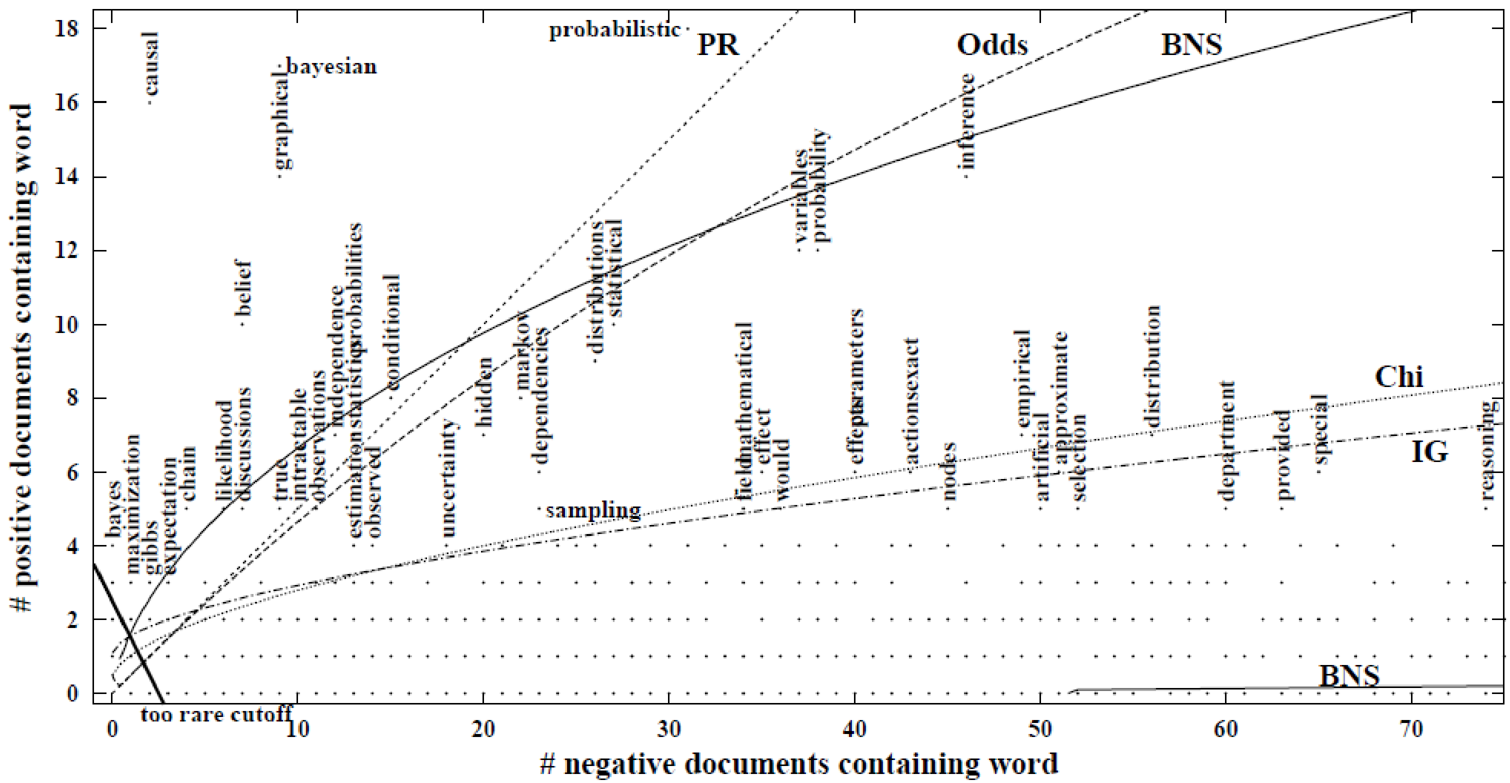
- Isoclines for each feature scoring metric
  - Machine learning papers v.s. other CS papers




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

# A graphical analysis of feature selection

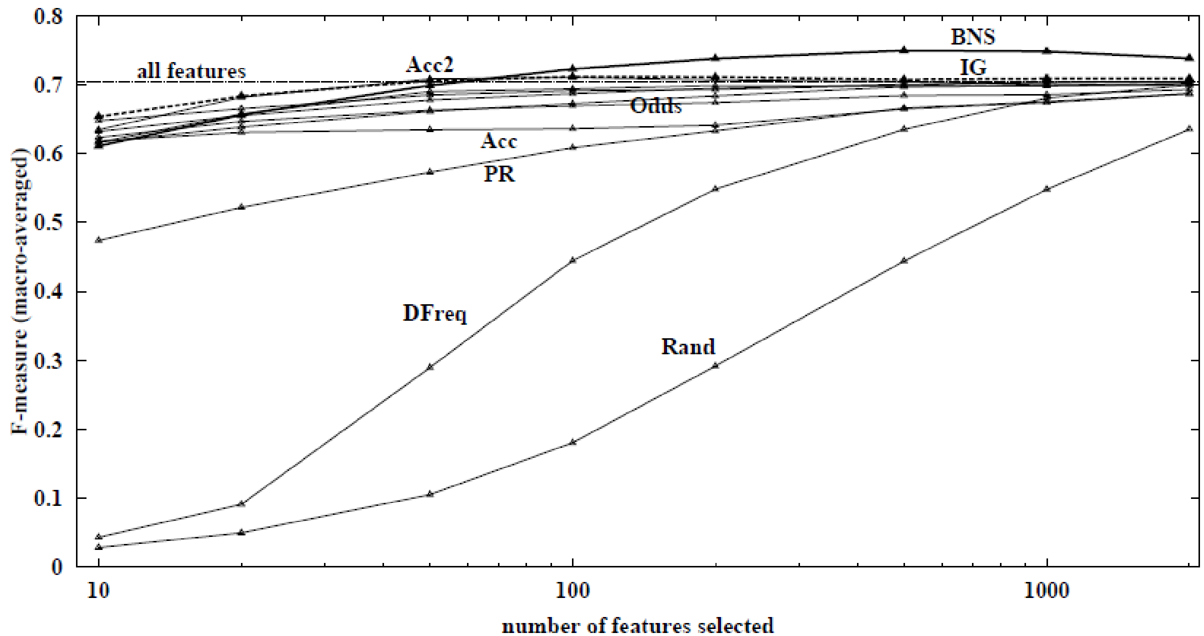
- Isoclines for each feature scoring metric
  – Machine learning papers v.s. other CS papers




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

# Effectiveness of feature selection methods

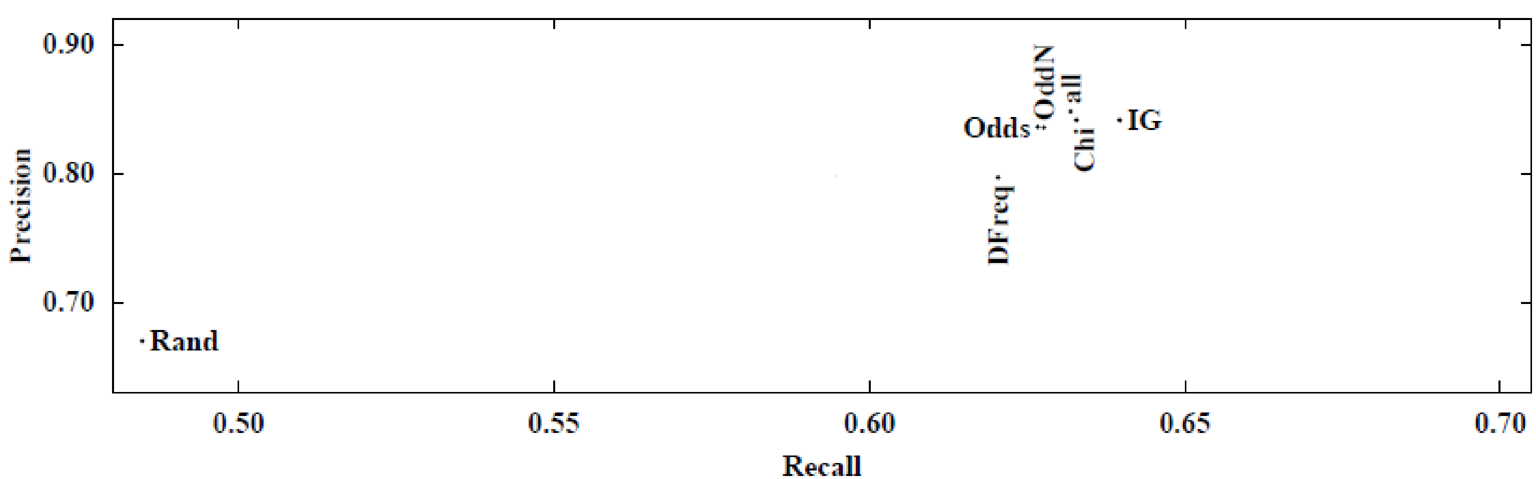
- On a multi-class classification data set
  - 229 documents, 19 classes
  - Binary feature, SVM classifier




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

# 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~~ | ~~N~~ |
| 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

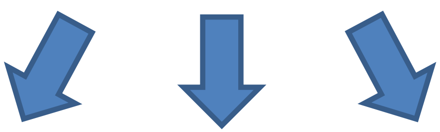
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"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.

PLAY VIDEO

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

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.



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 <u>independent</u> 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 <u>not independent</u> among each other
      - Decision tree, kernel SVM, mixture model
- Choose based on our domain knowledge of the problem

We will discuss these choices later

# 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.

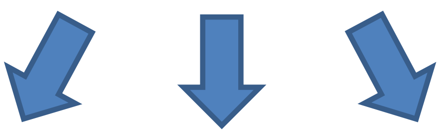
"These next two years are going to be the most interesting time in our lives," he told them, according to a person in the meeting that day.

PLAY VIDEO

Obama's Zinger in State of Union Address

Video by Associated Press on January 20, 2015. Photo by Doug Mills/The New York Times.

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.



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 $c_0$

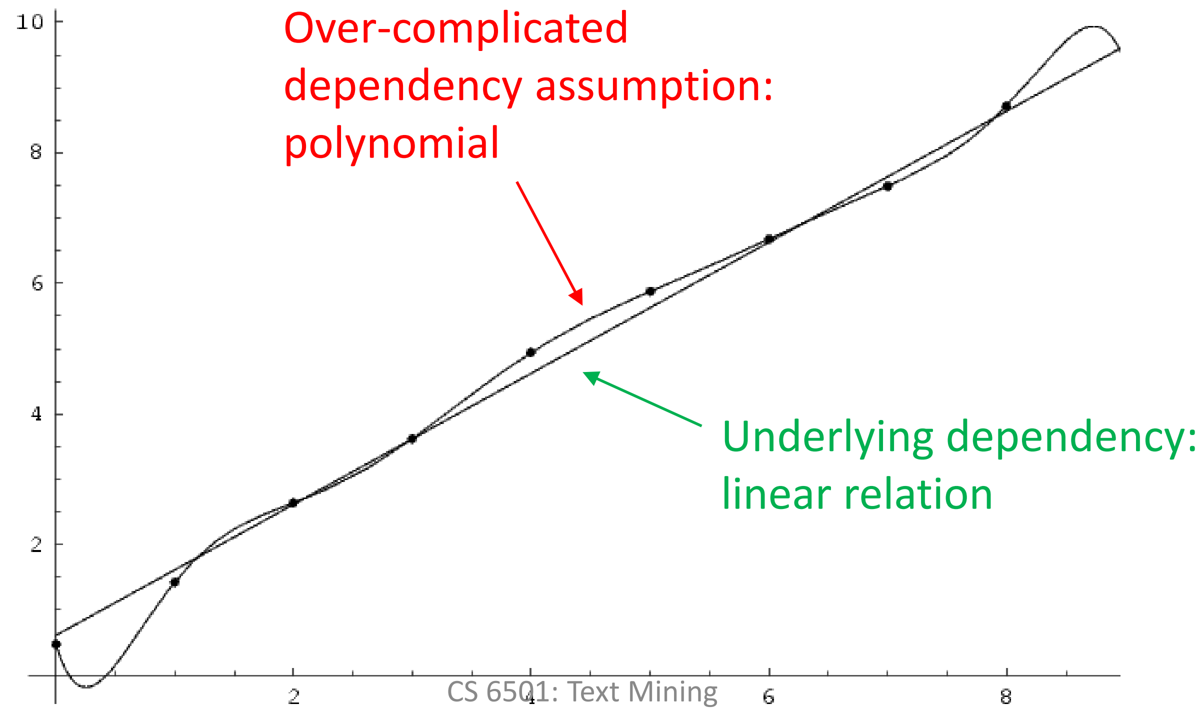
Penalty when misclassifying $c_0$ to $c_1$

Empirically estimated from training set

**Empirical loss!**

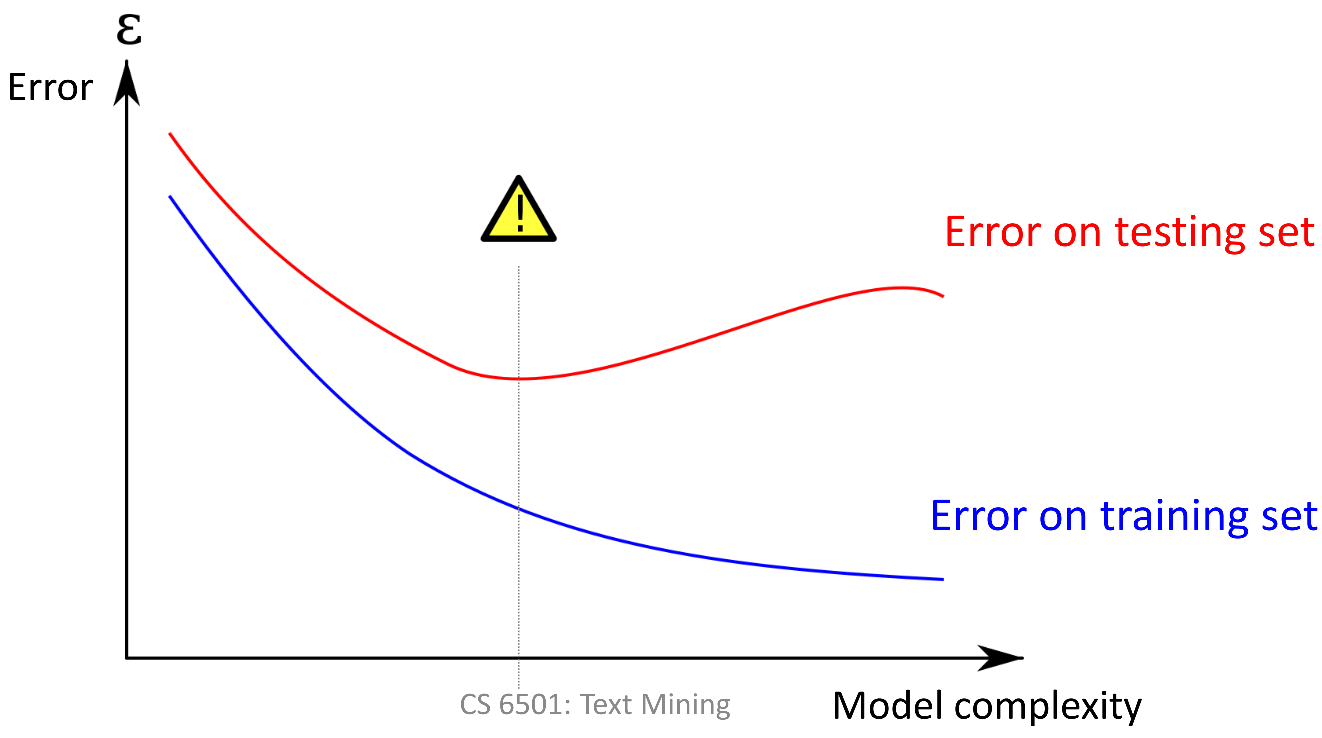
# Empirical loss minimization

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



# Generalization loss minimization

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

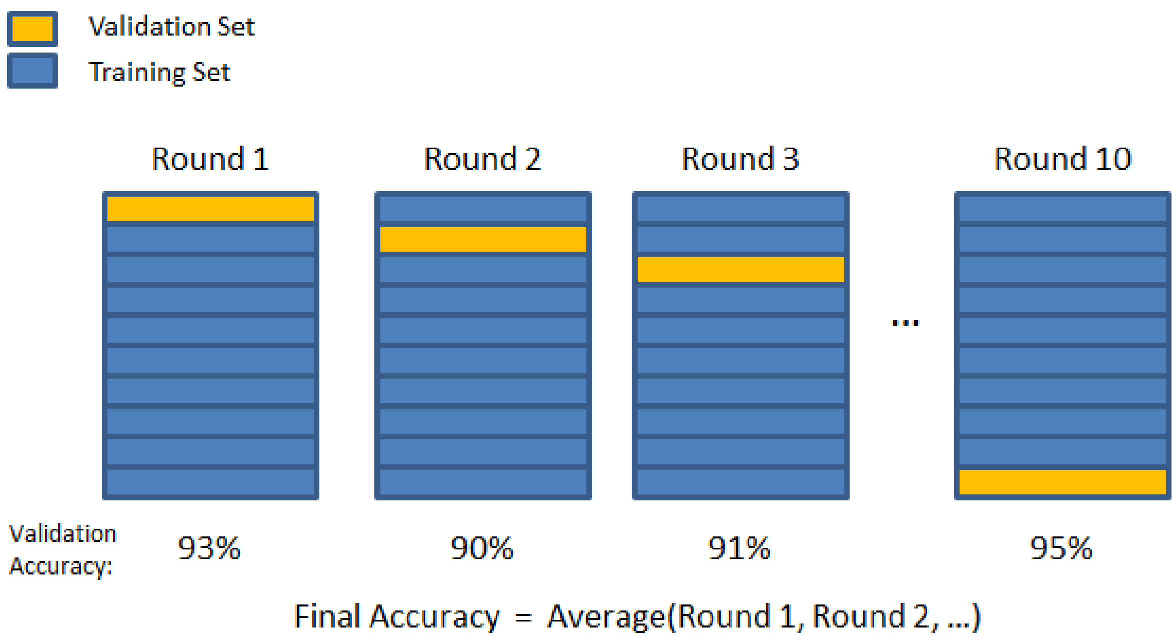


# Generalization loss minimization

- 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

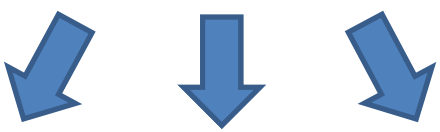
WASHINGTON — The morning after major Democratic losses in last year's midterm elections, President Obama walked into the Roosevelt Room with a message for his despondent staff: I'm not done yet.

"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.



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

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.



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

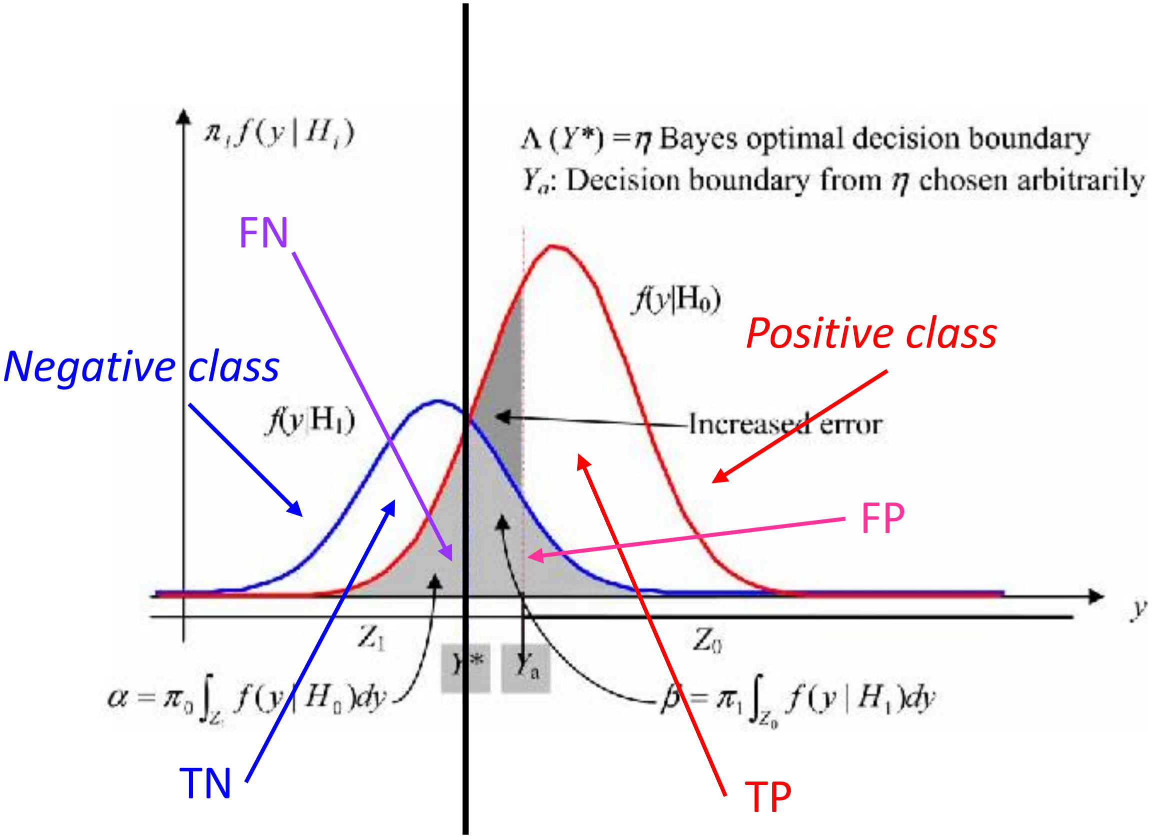
- Precision
  - Fraction of predicted positive documents that are indeed positive, i.e., $p(y^* = 1|y = 1)$
- Recall
  - Fraction of positive documents that are predicted to be positive, i.e., $p(y = 1|y^* = 1)$

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

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

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 (u … all ke…
- Th… pe…
  - – … ver

| No. | Approach | Precision AVG | Precision STD | Recall AVG | Recall STD |
|---|---|---|---|---|---|
| 1 | Triple-S | 0.31 | 0.19 | 0.36 | 0.26 |
| 2 | BP Graph Matching | **0.60** | 0.45 | 0.19 | 0.30 |
| 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 |
| 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, $\alpha$ balances the trade-off

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

*Equal weight between precision and recall*

  - Why harmonic mean?
    - System1: P:0.53, R:0.36
    - System2: P:0.01, R:0.99

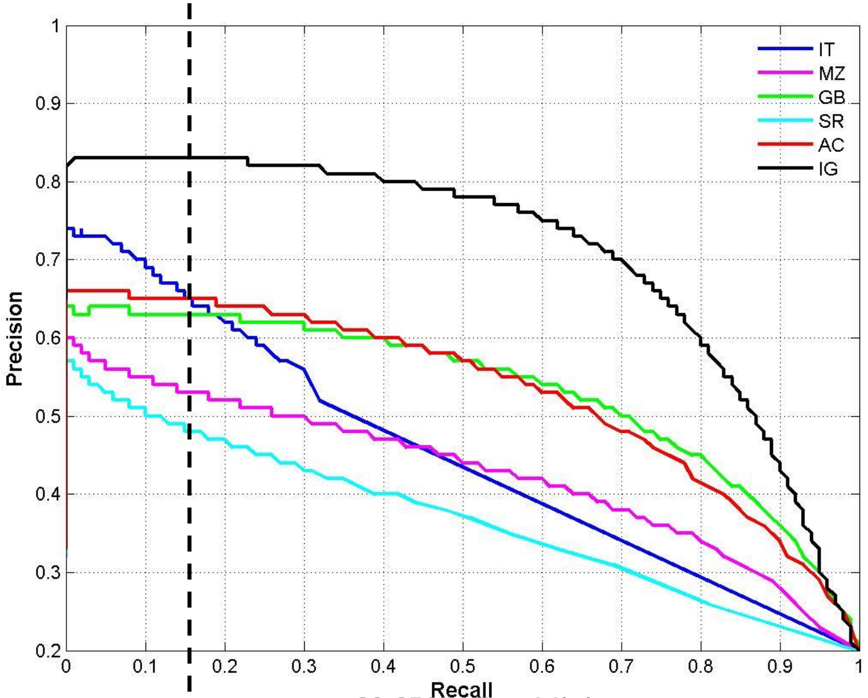
| H | A |
|---|---|
| 0.429 | 0.445 |
| 0.019 | 0.500 |

# 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;
  3. Plot precision and recall computed for each testing case in step (2).

# Summarizing precision and recall

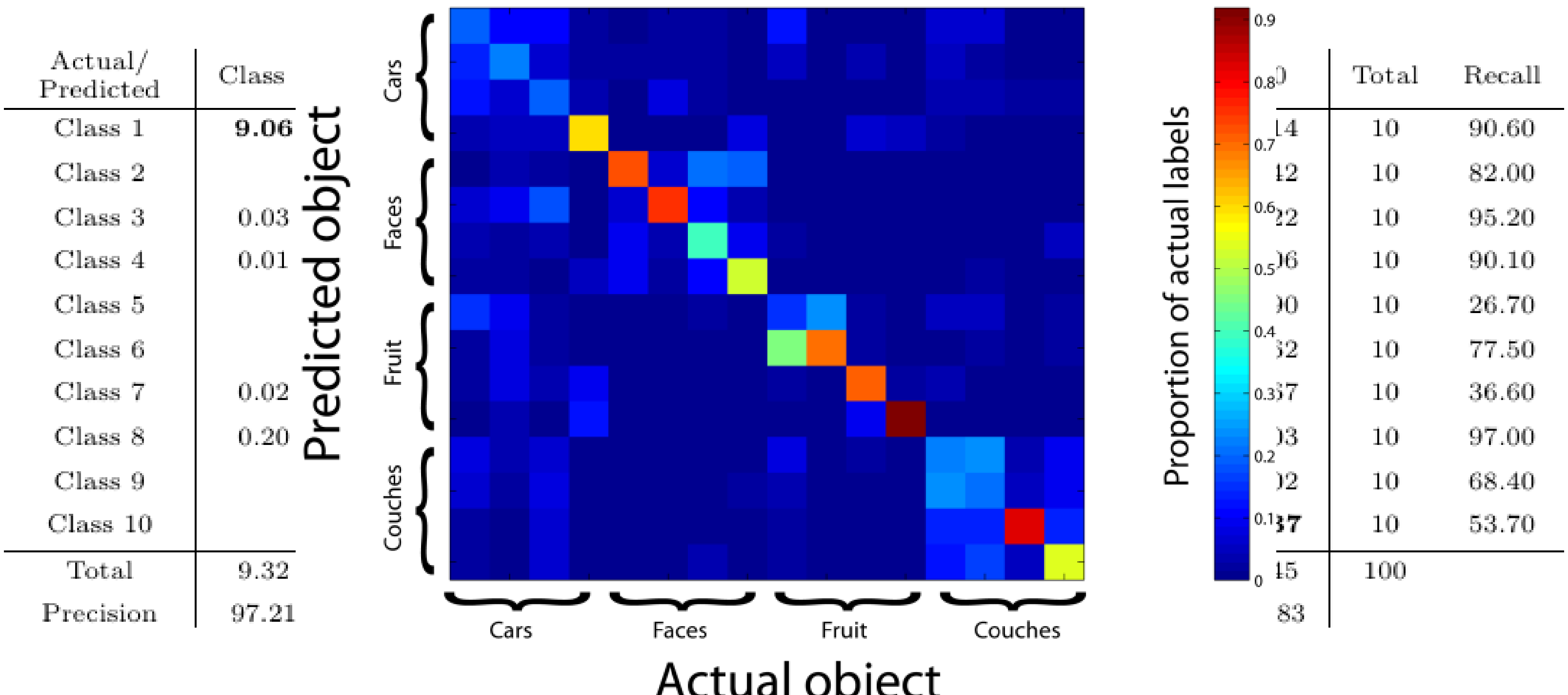
- With a curve
  - A.k.a., precision-recall curve
  - Area Under Curve (AUC)




*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

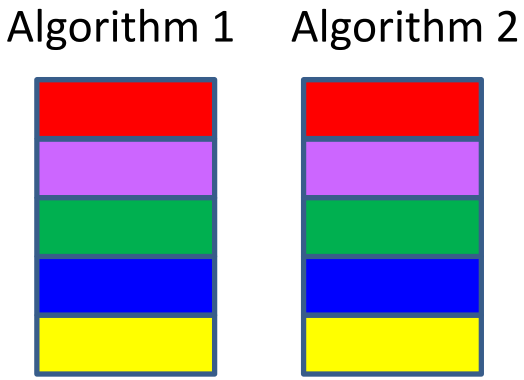
- How confident you are that an observed difference doesn't simply result from the train/test separation you chose?

| Fold | Algorithm 1 | Algorithm 2 |
|---|---|---|
| 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 ($\alpha$), 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 a random result), and so small *p*-values are good

# Paired *t*-test




- 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 testing

| Fold | System A | System B | paired t-test |
|---|---|---|---|
| 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 |