Feature Selection

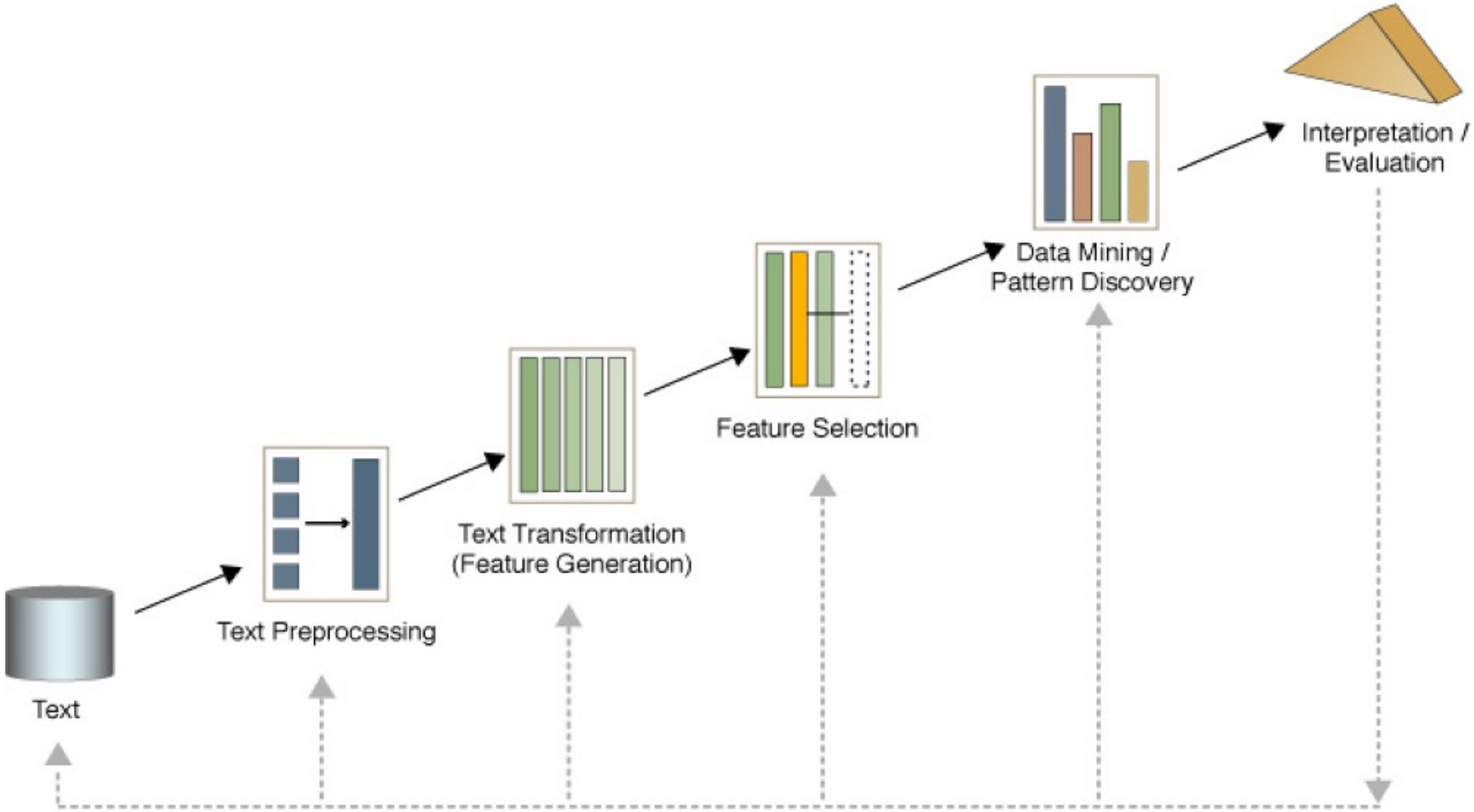
Applied Text Mining, from Foundations to Advanced

Ayoub Bagheri

This lecture

- Feature selection for text data
- Feature reduction vs feature selection
- Other methods

Text mining process



Feature Selection

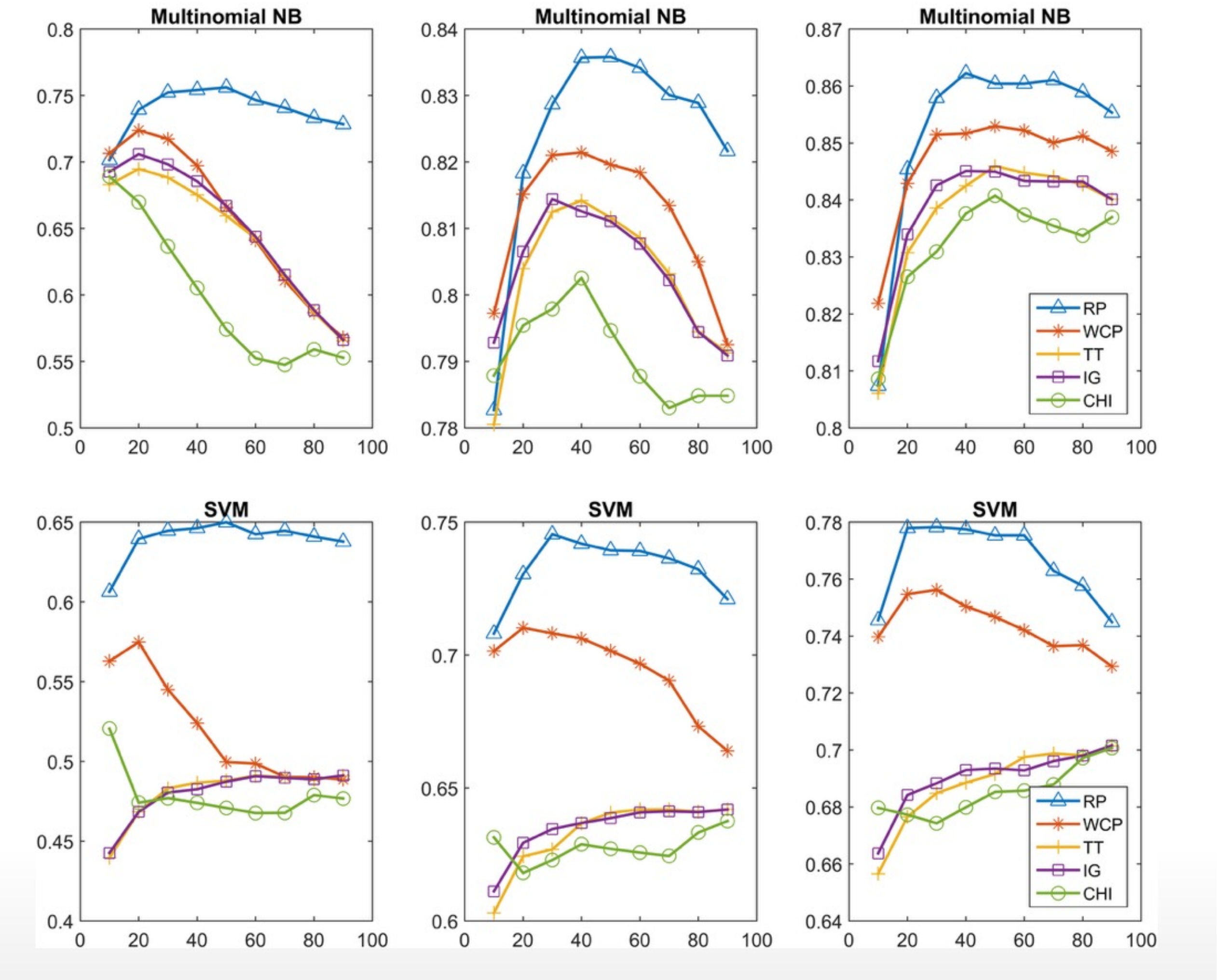
What is feature selection?

- Assume we have some data, and we want to use it to build a classifier, so that to predict something (e.g. email spam classification)
- The data has 10,000 features (e.g., token frequencies)
- We need to cut it down to 1,000 features before we try classification (data mining/pattern discovery). Which 1,000?
- The process of choosing the 1,000 features to use is called Feature Selection

Feature selection vs accuracy

- Relevance popularity: A term event model based feature selection scheme for text classification:

https://doi.org/10.1371/journal.pone.0174341



Why accuracy is reduced?

- Suppose *the best feature* set has 20 features.
- If you *add* another 5 features, typically the accuracy of machine learning may reduce.
- But you still have the original 20 features!
- Why does this happen?

Noise

- The additional features typically add *noise*. Machine learning will pick up on spurious correlations, that might be true in the training set, but not in the test set.
- For some ML methods, more features means more *parameters* to learn (more NN weights, more decision tree nodes, etc…)
- The increased space of possibilities is more difficult to search.

Why do we need feature selection?

1. To improve performance (predictive power).
2. Simplicity of the model (improves efficiency)
3. To visualize the data for model selection.
4. To reduce dimensionality and remove noise.

Feature selection in text

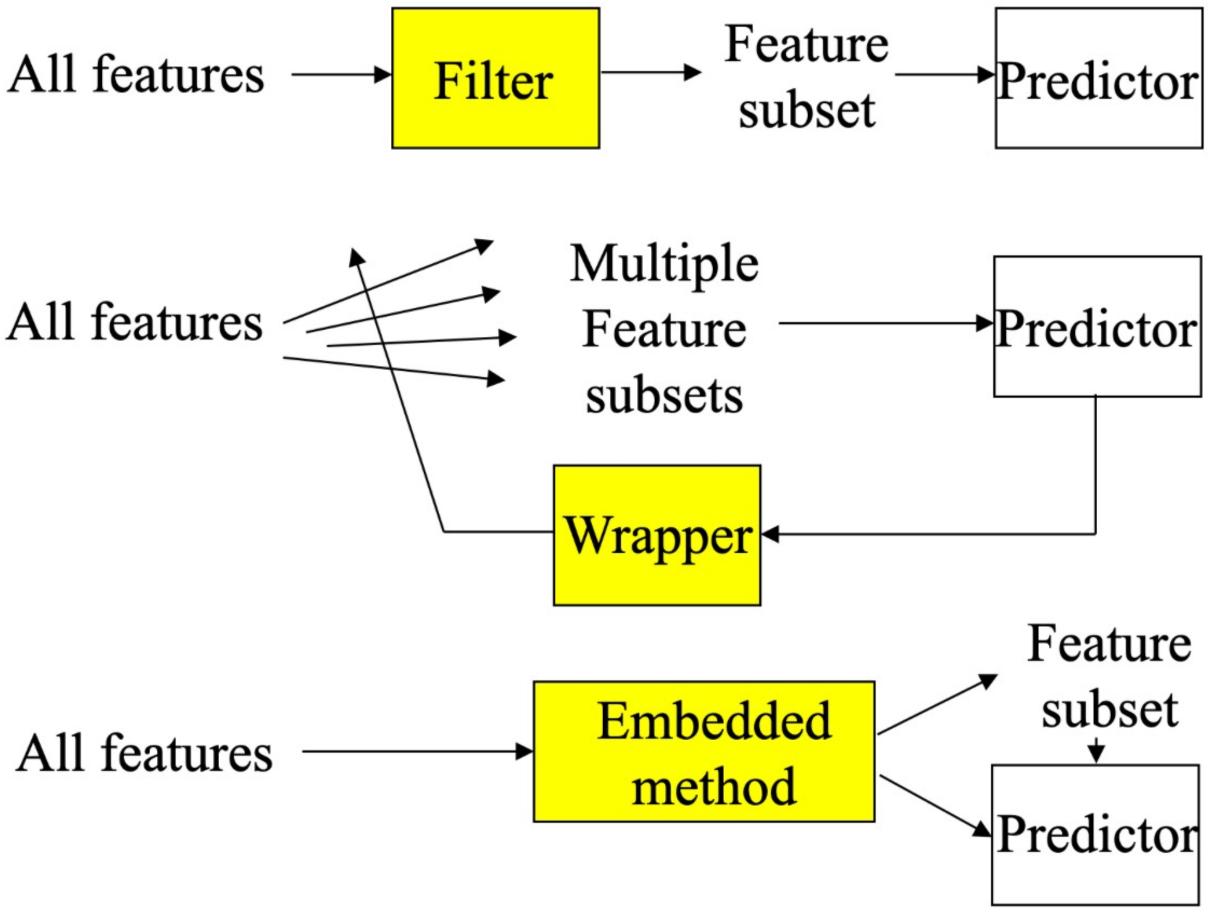
- *Feature Selection* is a process that chooses an optimal subset of features according to a certain criterion.
- *Feature selection* is the process of selecting a specific subset of the terms of the training set and using only them in the classification algorithm.

Feature selection in text

- 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

Feature selection methods



Filter 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

Filter methods

- Document frequency
- Gini index
- Information gain
- Chi-Square (χ^2)
- PMI (Mutual information)
- and more

Document frequency

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

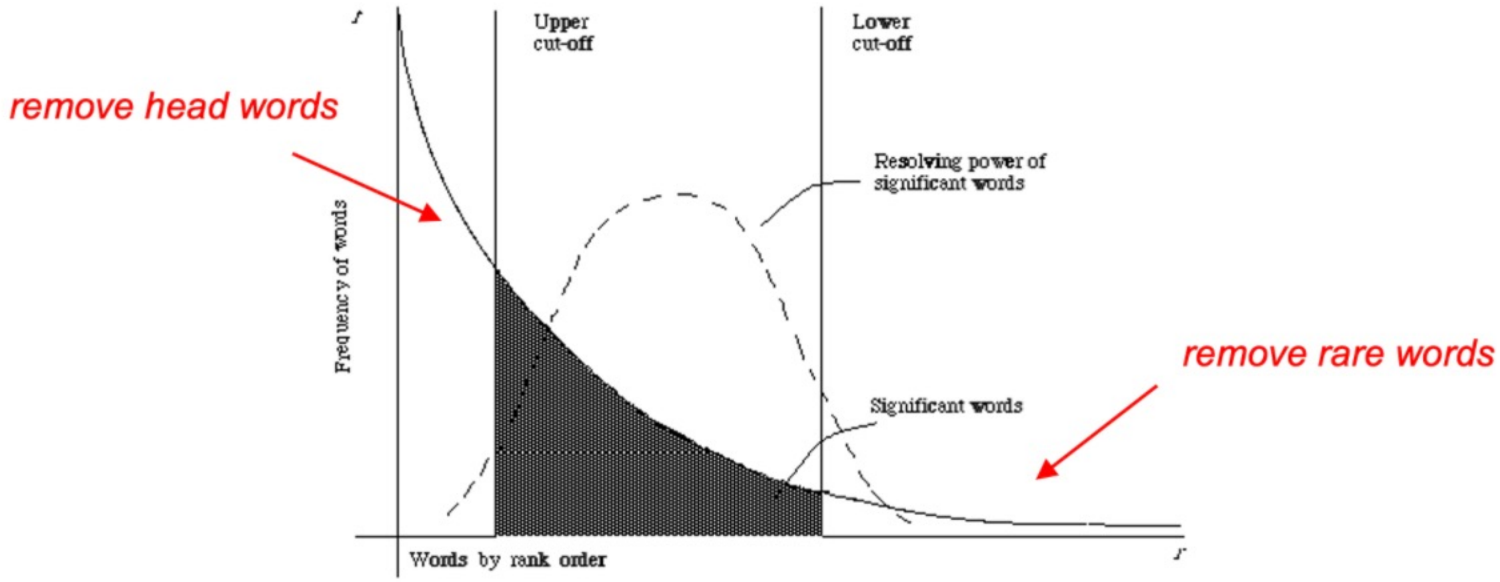


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

Gini index

Let $p(c|t)$ be the conditional probability that a document belongs to class c, given the fact that it contains the term t. Therefore, we have:

$$\sum_{c=1}^{k} p(c|t) = 1$$

Then, the gini-index for the term t, denoted by $G(t)$ is defined as:

$$G(t) = \sum_{c=1}^{k} p(c|t)^2$$

Gini index

- The value of the gini-index lies in the range $(1/k, 1)$.
- Higher values of the gini-index indicate a greater discriminative power of the term t.

Example

- Imagine we have the following emails:

Class	
Spam	Congratulations! You've been selected to receive a free gift.
Spam	Get a free trial of our exclusive protein supplemment
Spam	We invite you to join the VIP club and enjoy free gifts and discounts weekly
Not Spam	Join us for a free community yoga class this Saturday. All levels welcome
Not Spam	I want to invite you to the book club meeting next Thursday.
Not Spam	Join us to our annual family BBQ. Don't miss out on the fun

G(free) = ?

p(spam|free) =?
p(not spam|free) = ?

$$G(t) = \sum_{c=1}^{k} p(c|t)^2$$

Example

- Imagine we have the following emails:

Class	
Spam	Congratulations! You've been selected to receive a free gift.
Spam	Get a free trial of our exclusive protein supplemment
Spam	We invite you to join the VIP club and enjoy free gifts and discounts weekly
Not Spam	Join us for a free community yoga class this Saturday. All levels welcome
Not Spam	I want to invite you to the book club meeting next Thursday.
Not Spam	Join us to our annual family BBQ. Don't miss out on the fun

p(spam|free) =?
p(not spam|free) = ?

p(spam|free) = ¾ = 0.75
p(not spam|free) = ¼ = 0.25

$$G(t) = \sum_{c=1}^{k} p(c|t)^2$$

G(free) = $0.75^2 + 0.25^2 = 0.5625 + 0.0625 = 0.625$

Example

- Imagine we have the following emails:

Class	
Spam	Congratulations! You've been selected to receive a free gift. Click the link to claim your prize.
Spam	Get a free trial of our exclusive protein supplemment
Spam	We invite you to join the VIP club and enjoy free gifts and discounts weekly
Not Spam	Join us for a free community yoga class this Saturday. All levels welcome
Not Spam	I want to invite you to the book club meeting next Thursday. Looking forward to it
Not Spam	Join us to our annual family BBQ. Don't miss out on the fun

G(invite) = ?

G(join) = ?

Example

- Imagine we have the following emails:

Class	
Spam	Congratulations! You've been selected to receive a free gift. Click the link to claim your prize.
Spam	Get a free trial of our exclusive protein supplement
Spam	We invite you to join the VIP club and enjoy free gifts and discounts weekly
Not Spam	Join us for a free community yoga class this Saturday. All levels welcome
Not Spam	I want to invite you to the book club meeting next Thursday. Looking forward to it
Not Spam	Join us to our annual family BBQ. Don't miss out on the fun

G(invite) = $0.5^2 + 0.5^2 = 0.5$

G(join) = ?

Example

- Imagine we have the following emails:

Class	
Spam	Congratulations! You've been selected to receive a free gift. Click the link to claim your prize.
Spam	Get a free trial of our exclusive protein supplemment
Spam	We invite you to join the VIP club and enjoy free gifts and discounts weekly
Not Spam	Join us for a free community yoga class this Saturday. All levels welcome
Not Spam	I want to invite you to the book club meeting next Thursday. Looking forward to it
Not Spam	Join us to our annual family BBQ. Don't miss out on the fun

$G(\text{invite}) = 0.5^2 + 0.5^2 = 0.5$

$G(\text{join}) = 0^2 + 1^2 = 1$

G(join) > G(invite) means that *join* has a greater discriminative power than *invite*.

Information gain

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

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

Entropy of class label along

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

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

Chi-square

- χ^2 statistics with multiple categories
  - $\chi^2 = \sum_c p(c)\chi^2(c,t)$
    - Expectation of χ^2 over all the categories
  - $\chi^2(t) = \max_c \chi^2(c,t)$
    - Strongest dependency between a category and a term

This term calculates how strongly a feature is associated with a particular class

PMI

- Mutual information
    - Relatedness between term t and class c

High PMI indicates a strong association between the term and the class.

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

PMI example

	Spam	Not Spam	Total
Offer is present	120	10	130
Offer is absent	180	190	370
Total	300	200	500

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

PMI(offer, spam)= ?

- p(offer, spam)= 120/500 = 0.24
- p(offer)= 130/500 = 0.26
- p(spam)= 300/500 = 0.60

PMI(offer, spam)= 0.24 * log($\frac{0.24}{0.26 * 0.6}$) = 0.24 * 0.431 = 0.1

PMI

	Spam	Not Spam	Total
Offer is present	120	10	130
Offer is absent	180	190	370
Total	300	200	500

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

PMI(offer, not spam)= ?

- p(offer, not spam)= ?
- p(offer)= ?
- p(not spam)= ?

PMI

	Spam	Not Spam	Total
Offer is present	120	10	130
Offer is absent	180	190	370
Total	300	200	500

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

PMI(offer, not spam)= ?

- p(offer, not spam)= 10/500 = 0.02
- p(offer)= 130/500 = 0.26
- p(not spam)= 200/500 = 0.4

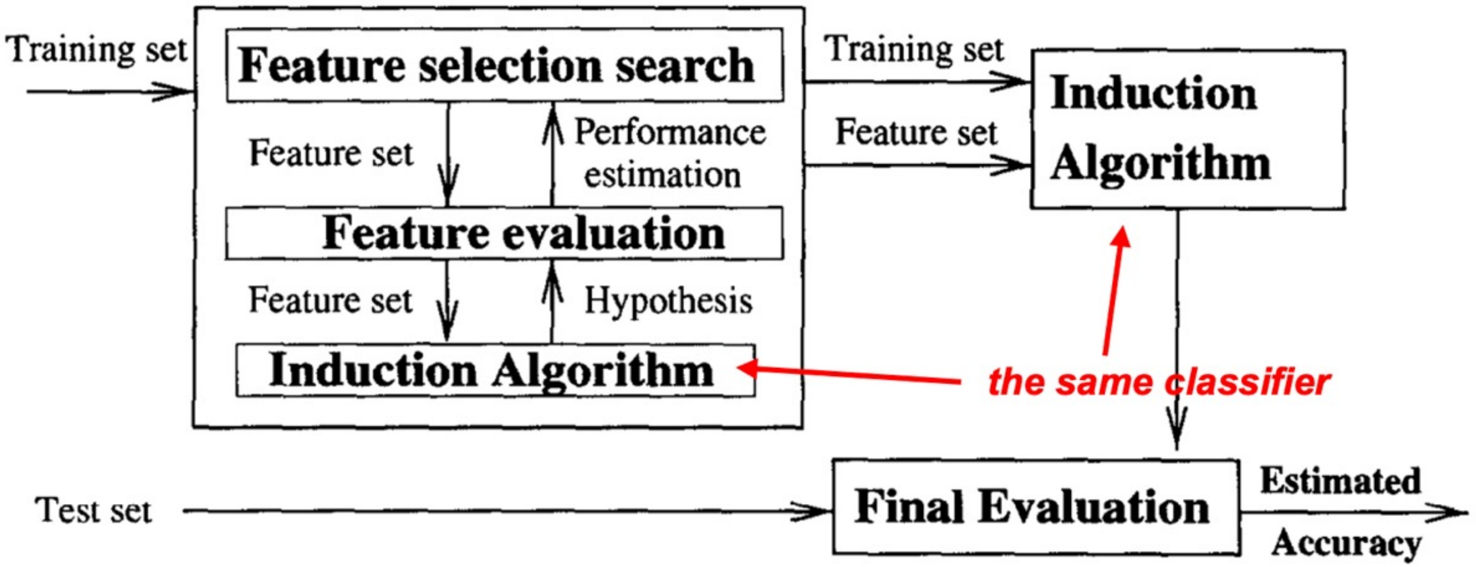
PMI(offer, Not Spam) = 0.02 * log($\frac{0.02}{0.26 * 0.4}$) = 0.02 * (-0.678) = - 0.013

Break

Wrapper 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

Wrapper method

- Optimizes feature set for a specific learning algorithm
- The feature subset selection algorithm is a “wrapper” around the learning algorithm
  - Pick a feature subset and pass it in to a learning algorithm
  - Create training/test set based on the feature subset
  - Train the learning algorithm with the training set
  - Find accuracy (objective) with validation set
  - Repeat for all feature subsets and pick the feature subset which led to the highest predictive accuracy (or other objective)

Wrapper method

- Basic approach is simple
- Variations are based on how to select the feature subsets, since there are an exponential number of subsets

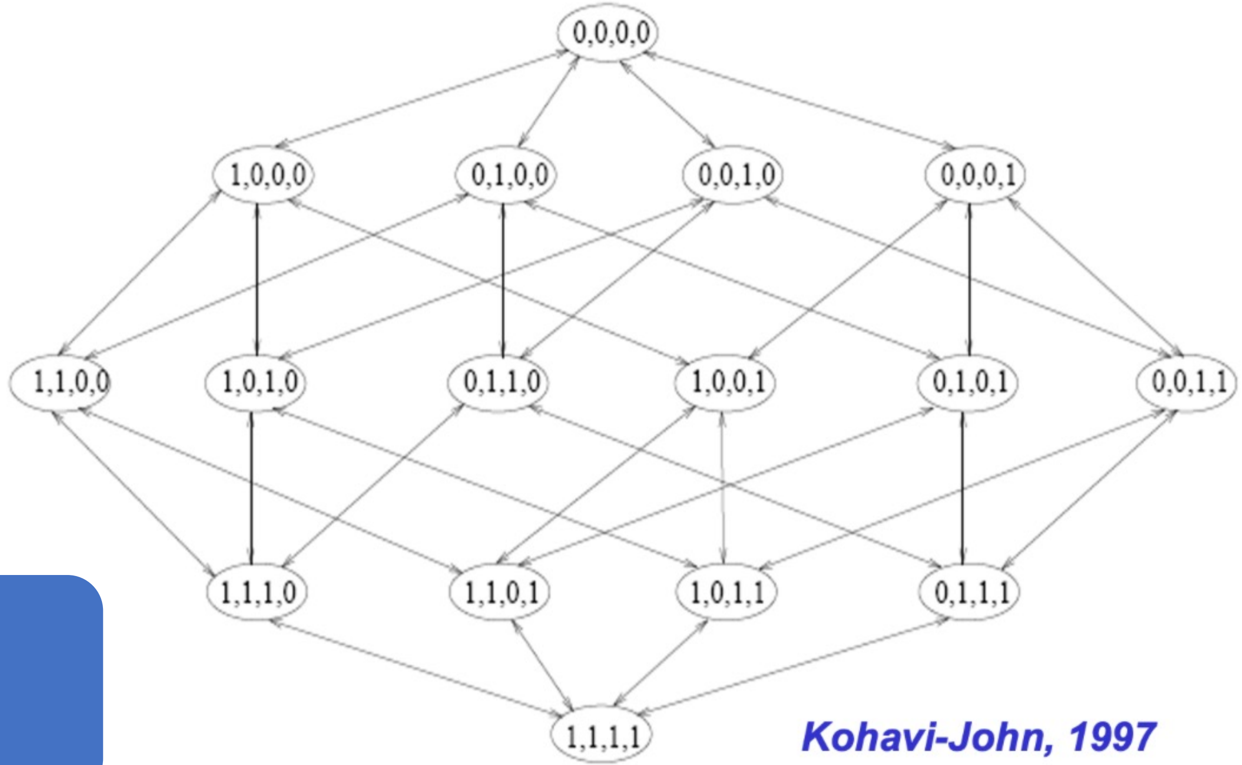
Wrapper method

- Wrapper methods evaluate feature subsets using a specific learning algorithm, which often results in optimal performance for that particular model
- Consider all possible dependencies among the features
- Flexible: Can be used with any model and is adaptable to the problem at hand

Wrapper method

- Impractical for text classification
  - Cannot deal with large feature set
  - A NP-complete problem
  - Given a dataset with N features, the number of possible feature subsets is 2^N.
  - For large n, exhaustively evaluating all subsets becomes computationally infeasible.

Wrapper method



Kohavi-John, 1997

n	2^n
48	281 474 976 710 656
49	562 949 953 421 312
50	1 125 899 906 842 624
51	2 251 799 813 685 248
52	4 503 599 627 370 496
53	9 007 199 254 740 992
54	18 014 398 509 481 984
55	36 028 797 018 963 968
56	72 057 594 037 927 936
57	144 115 188 075 855 872
58	288 230 376 151 711 744
59	576 460 752 303 423 488
60	1 152 921 504 606 846 976
61	2 305 843 009 213 693 952
62	4 611 686 018 427 387 904
63	9 223 372 036 854 775 808

2^{10} is 1024
2^{20} is 1.048.576

N features, 2^N possible feature subsets!

Wrapper method

- Evaluating all possible subsets can be computationally expensive, especially with a large feature set, as each evaluation requires training a model.
- Risk of overfitting: Tuning the model too closely to the training data can lead to overfitting, as the method seeks feature sets that maximize training performance rather than generalization.
- Changing the selection of the algorithm require re-evaluation of the feature selection process.

Search strategies

- Exhaustive search
- Greedy search: forward selection or backward elimination
- Simulated annealing
- Genetic algorithms

- *These will not be discussed in this course.*

Embedded Methods

Embedded Methods

- The process of selecting features is integrated with the model training process itself.
- Embedded methods include feature selection as part of the model optimization process.
- The model focuses only on the most relevant features and often results in better performance with reduced computational cost compared to wrapper methods.

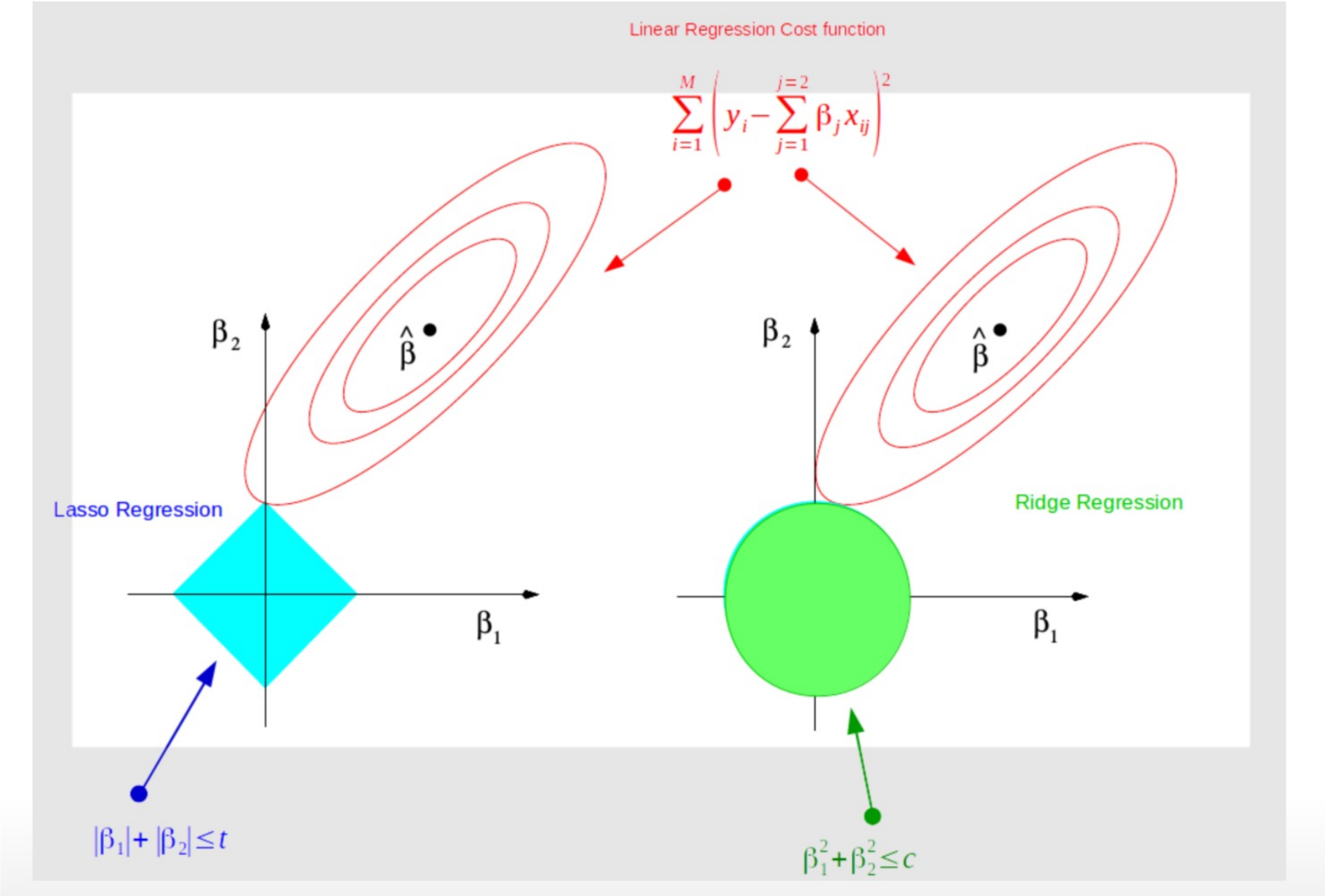
Formalism

- Many learning algorithms are cast into a minimization of some regularized functional:

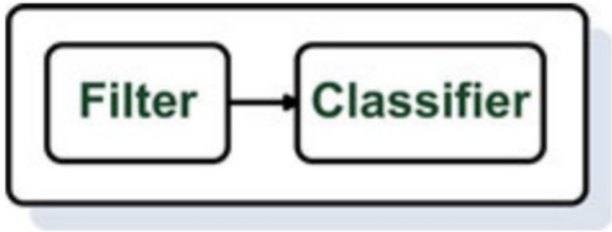
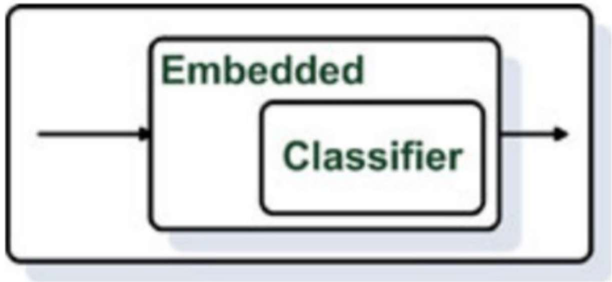
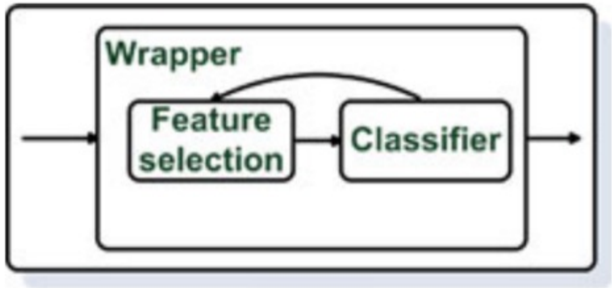
$$\min_{\alpha} \hat{R}(\alpha, \sigma) = \min_{\alpha} \underbrace{\boxed{\sum_{k=1}^{m} L(f(\alpha, \sigma \circ x_k), y_k)}}_{\text{Loss function}} + \underbrace{\boxed{\Omega(\alpha)}}_{\text{Regularization Term}}$$

- The loss function measures how far off the model's predictions are from the actual values, aiming to minimize this difference.
- The regularization term adds a penalty to prevent the model from becoming overly complex, promoting simplicity and better generalization.

Lasso & Ridge regression



Comparing methods

Method	Advantages	Disadvantages
Filter 	Independence of the classifier Lower computational cost than wrappers Fast Good generalization ability	No interaction with the classifier
Embedded 	Interaction with the classifier Lower computational cost than wrappers Captures feature dependencies	Classifier-dependent selection
Wrapper 	Interaction with the classifier Captures feature dependencies	Computationally expensive Risk of overfitting Classifier-dependent selection

Feature Reduction

Feature selection vs feature reduction

- *Feature Selection* seeks a *subset* of the n original features which retains most of the relevant information
  - Wrappers (e.g. forward selection), Filters (e.g. PMI), Embedded (e.g. Lasso, Regularized SVM)
- *Feature Reduction* *combines/fuses* the n original features into a smaller set of *newly* created features which hopefully retains most of the relevant information from *all* the original features (e.g. LDA, PCA, etc.)

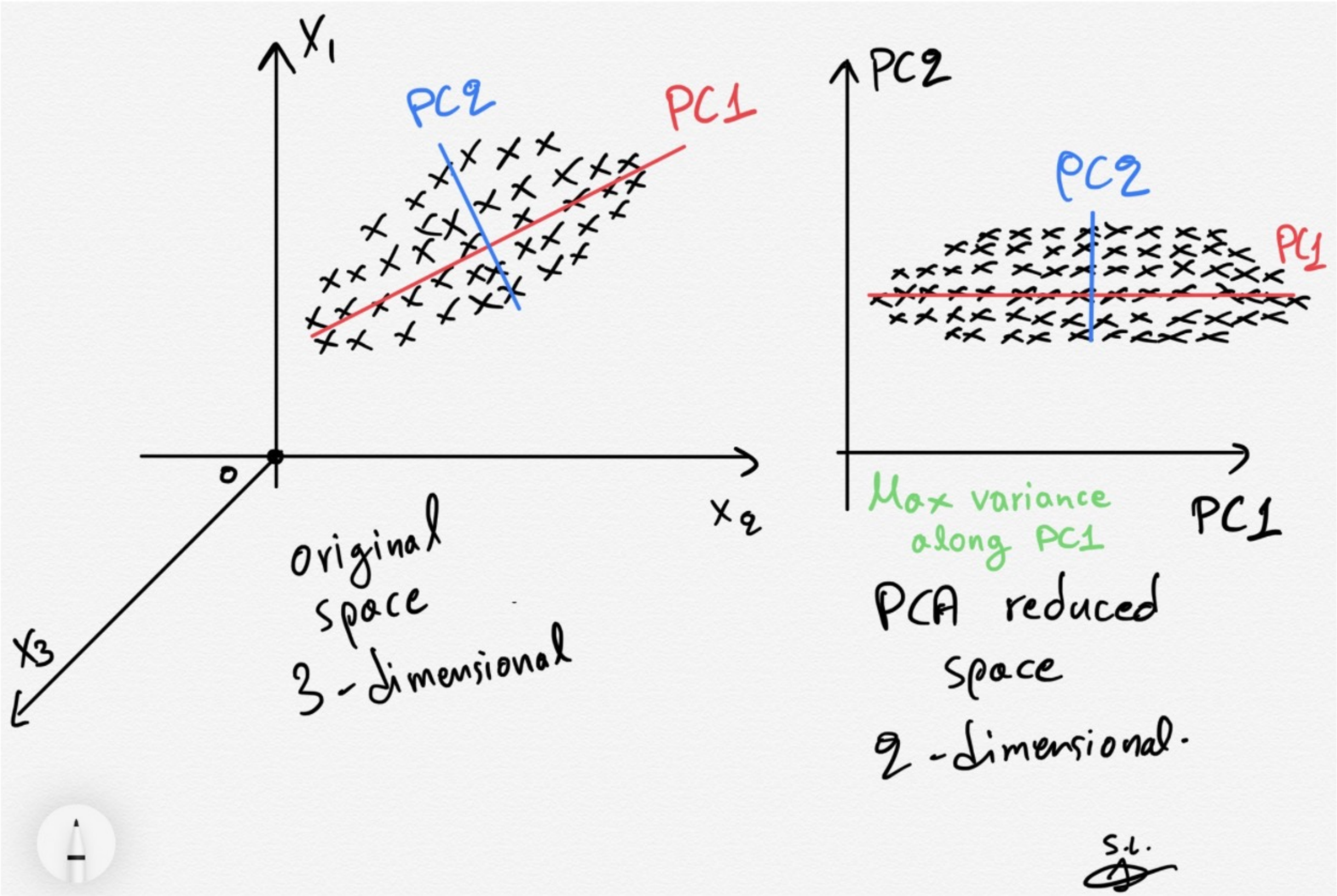
PCA: Principal Component Analysis

- PCA is a dimensionality / feature reduction technique that transforms high-dimensional text features (e.g., term-frequency matrices) into a smaller set of uncorrelated components, preserving as much variance as possible.
- It helps in reducing redundancy and noise, making models more efficient while retaining key information for analysis.

PCA: Principal Component Analysis

- PCA is a linear method for dimensionality reduction, finds a sequence of linear combinations of features that have maximum variance and are uncorrelated
- Allows to combine much of the information contained in n features into p features where $p<n$
- PCA is *unsupervised* in that it does not consider the output class / value of an instance – There are other algorithms which do (e.g. LDA: Linear Discriminant Analysis)
- PCA works well in many cases where data have mostly linear correlations.

PCA overview



https://towardsdatascience.com/

PCA Advantages & Disadvantages

- Advantages:
  - Reduces the number of features while preserving as much variance as possible, simplifying data and computational requirements.
  - Eliminates noise and redundant features
  - Better visualization of high-dimensional data by projecting it into 2D or 3D plots.
- Disadvantages:
  - Interpretability: Principal components are linear combinations of original features, making them hard to interpret in terms of the original feature set.
  - Reducing dimensions might lead to the loss of some information

Summary

- Feature selection is the process of identifying and choosing the most relevant features in a dataset to improve model performance and efficiency
- Feature selection methods:
  - Filter methods (e.g., Mutual Information, Chi-Square)
  - Wrapper methods (e.g., Forward/Backward Selection)
  - Embedded methods (e.g., LASSO, RIDGE)

Summary

- Reducing the dimensionality of text features helps improve:
  - Computational efficiency
  - Making models faster
  - Improve performance
  - And more scalable.

Practical 3

Questions?