Feature Selection in Text

Applied Text Mining

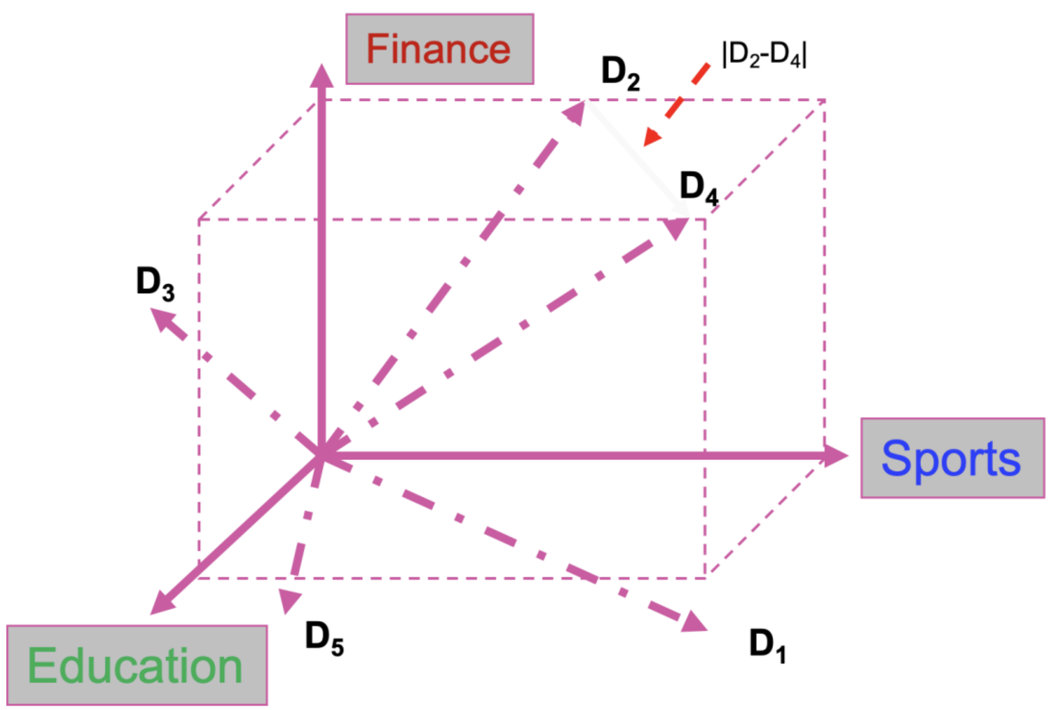
Ayoub Bagheri

Lecture plan

1. How to do feature selection (FS) for text data?
2. Is PCA a FS method for text?
3. Other methods?

An illustration of VS model

- All documents are projected into this concept space



Feature selection: What

You have some data, and you want to use it to build a classifier, so that you can predict something (e.g. email spam classification)

Feature selection: What

You have some data, and you want to use it to build a classifier, so that you can predict something (e.g. email spam classification)

The data has 10,000 fields (features)

Feature selection: What

You have some data, and you want to use it to build a classifier, so that you can predict something (e.g. email spam classification)

The data has 10,000 fields (features)

you need to cut it down to 1,000 fields before you try machine learning. Which 1,000?

Feature selection: What

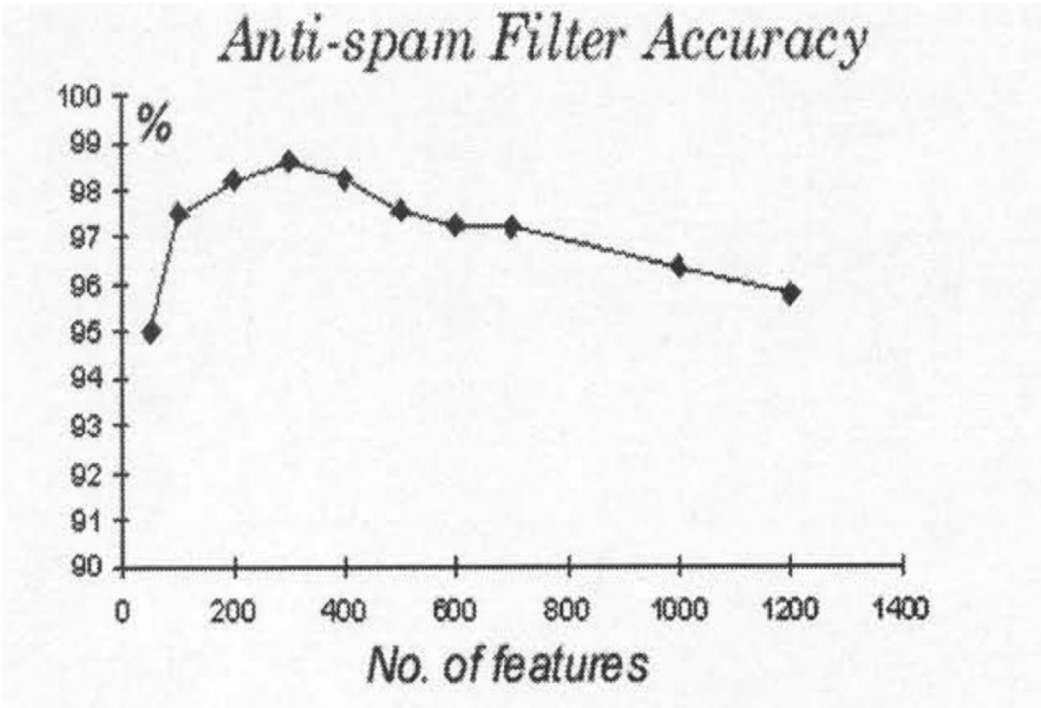
You have some data, and you want to use it to build a classifier, so that you can predict something (e.g. email spam classification)

The data has 10,000 fields (features)

you need to cut it down to 1,000 fields before you try machine learning. Which 1,000?

The process of choosing the 1,000 fields to use is called Feature Selection

Feature selection: Why



From http://elpub.scix.net/data/works/att/02-28.content.pdf

Why accuracy reduces

- 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 / Explosion

- 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.

Feature selection

Why we need FS:

```
1. To improve performance (in terms of speed, predictive power, simplicity of
the model).
```

```
2. To visualize the data for model selection.

3. To reduce dimensionality and remove noise.
```

Feature selection for text

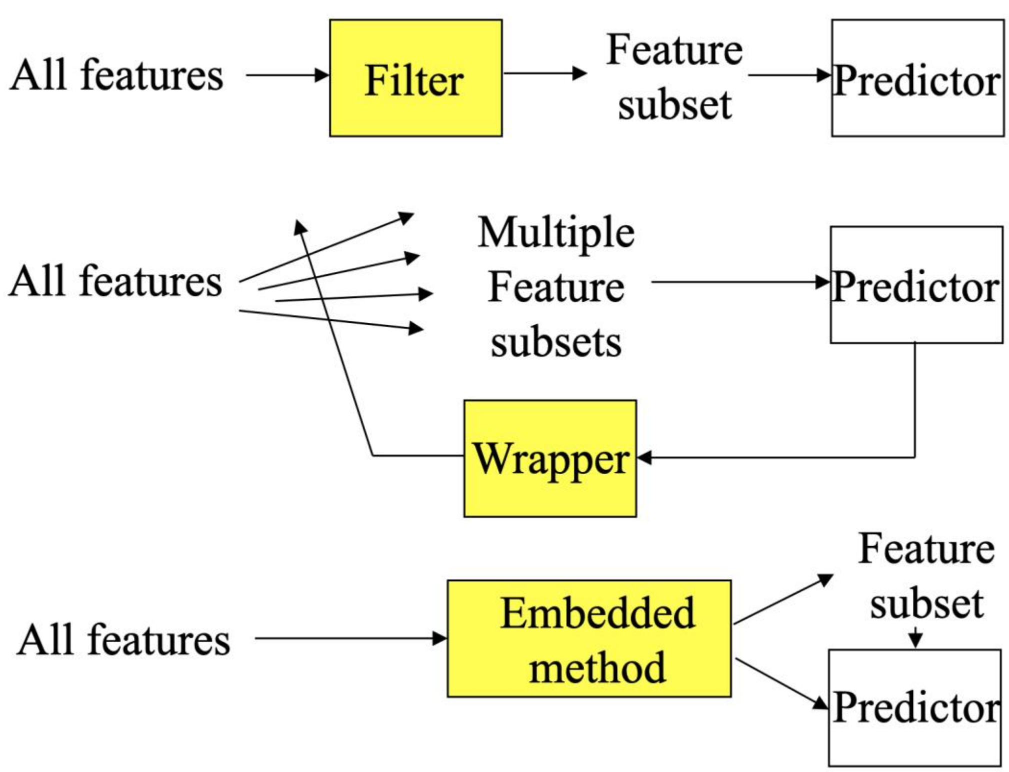
Feature Selection is a process that chooses an optimal subset of features according to a certain criterion.

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.

- 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

Methods

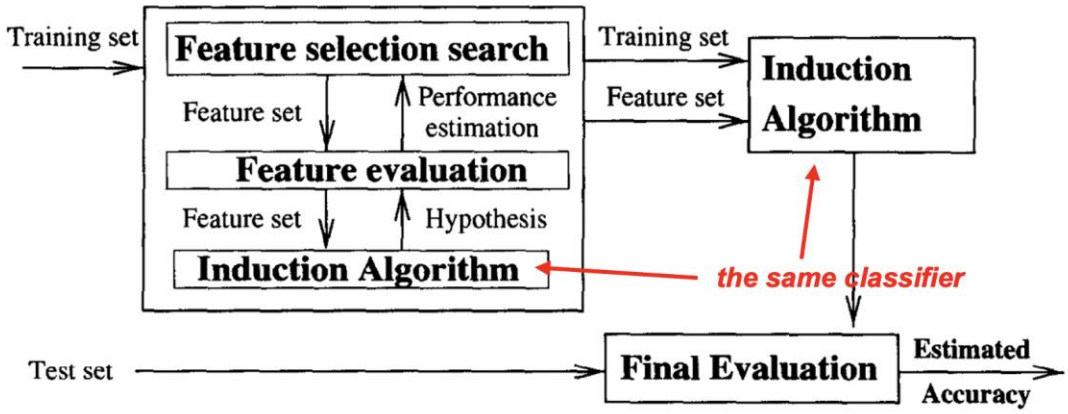
Filters, Wrappers, Embedded, and Hybrid



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

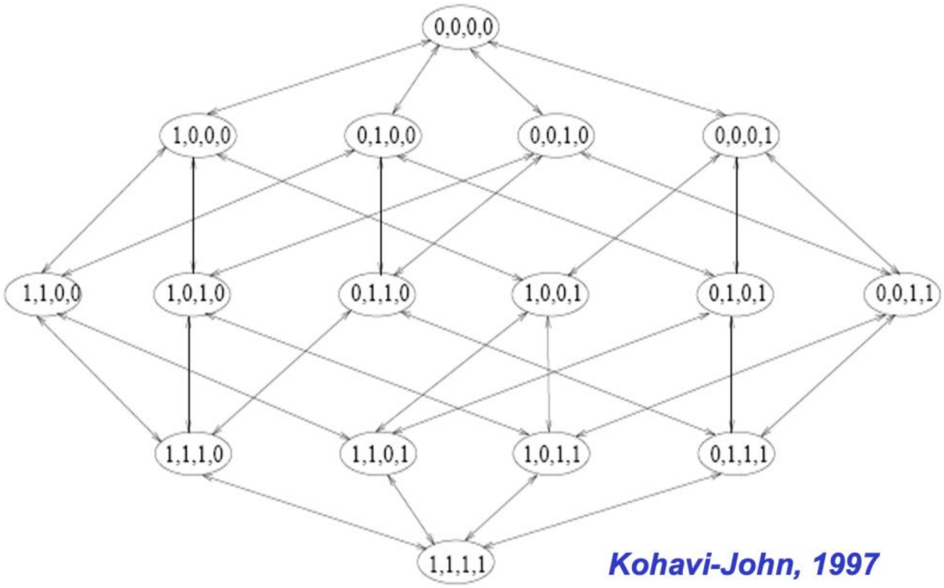
Wrapper method

- Optimizes for a specific learning algorithm
- The feature subset selection algorithm is a “wrapper” around the learning algorithm
    1. Pick a feature subset and pass it in to a learning algorithm
    2. Create training / test set based on the feature subset
    3. Train the learning algorithm with the training set
    4. Find accuracy (objective) with validation set
    5. Repeat for all feature subsets and pick the feature subset which led to the highest predictive accuracy (or other objective)
- 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 method
    - Consider all possible dependencies among the features
    - Impractical for text classification
        - Cannot deal with large feature set
        - A NP-complete problem
            - No direct relation between feature subset selection and evaluation

Wrappers for feature selection



Kohavi-John, 1997

N features, 2^N possible feature subsets!

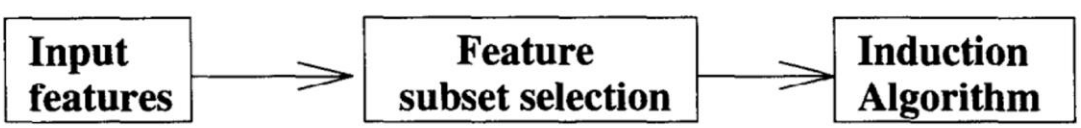
Search strategies

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

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

Document frequency

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

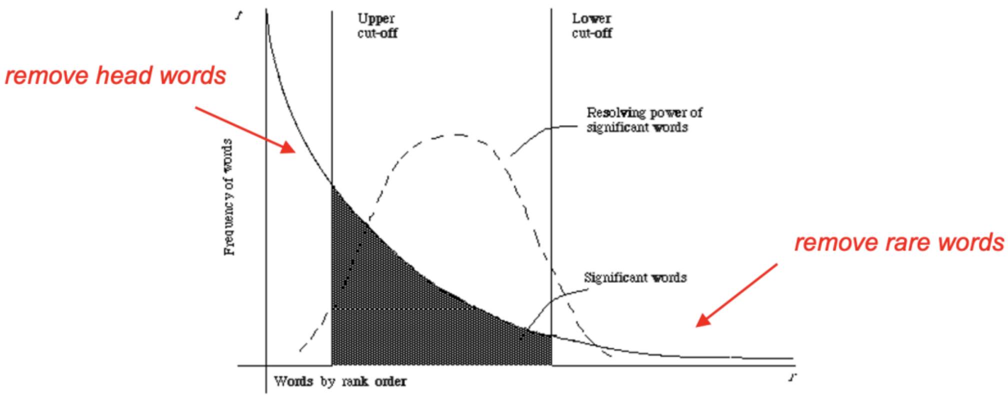


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz[44] page 120)

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.

Information gain

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

Other metrics

- 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

Other metrics

- Many other metrics (Same trick as in χ^2 statistics for multi-class cases)
  - 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

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

Embedded Methods

Formalism

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

$$\min_{\alpha} \hat{R}(\alpha, \sigma) = \min_{\alpha} \sum_{k=1}^{m} L(f(\alpha, \sigma \circ x_k), y_k) + \Omega(\alpha)$$

Formalism

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

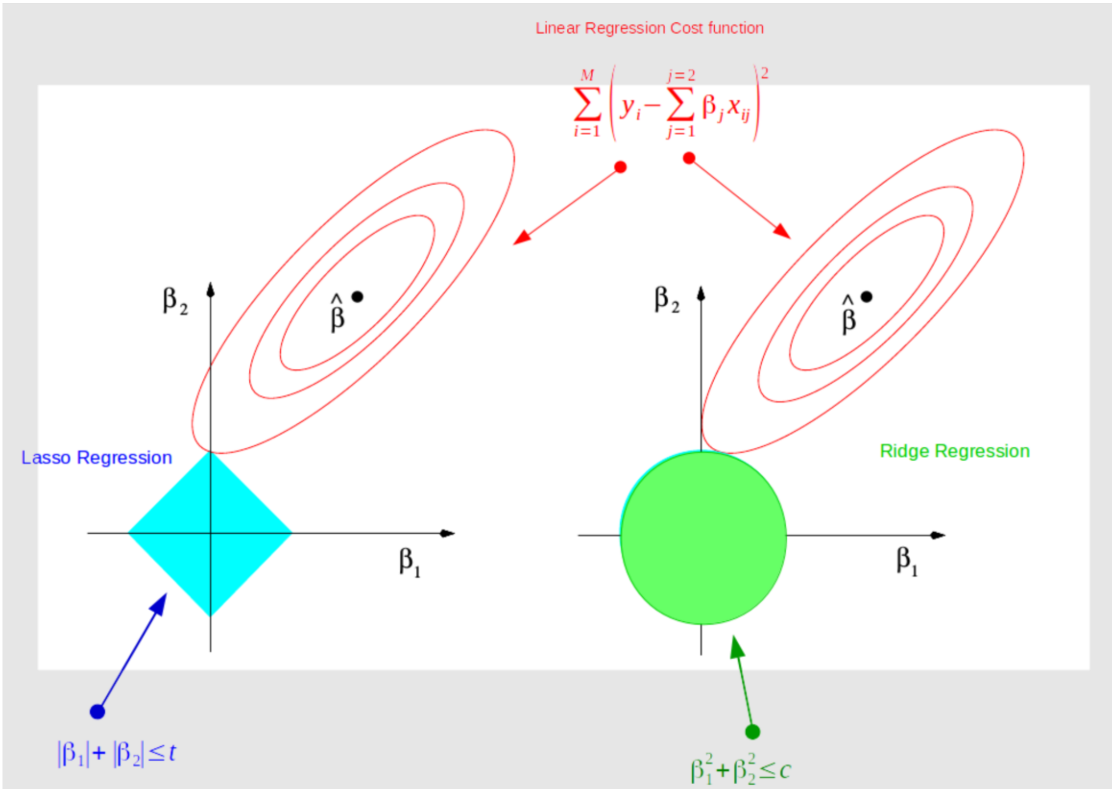
$$\underbrace{\min_{\alpha} \hat{R}(\alpha, \sigma)}_{G(\sigma)} = \min_{\alpha} \sum_{k=1}^{m} L(f(\alpha, \sigma \circ x_k), y_k) + \Omega(\alpha)$$

Empirical error

Regularization capacity control

Justification of RFE and many other embedded methods.

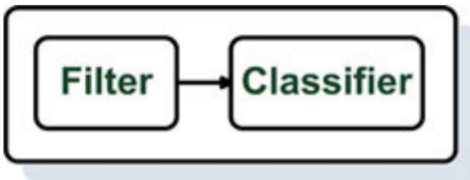
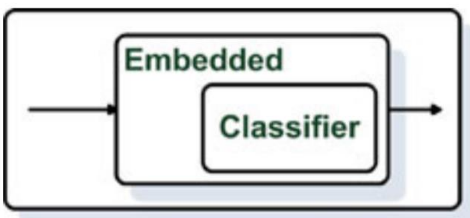
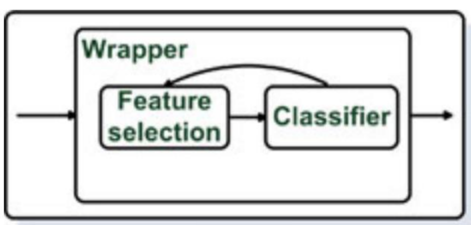
Lasso vs Ridge



The l_1 SVM

- A version of SVM where $\Omega(w) = ||w||^2$ is replaced by the l_1 norm $\Omega(w) = \sum_i |w_i|$
- Can be considered an embedded feature selection method:
  - Some weights will be drawn to zero (tend to remove redundant features)
  - Difference from the regular SVM where redundant features are included

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

PCA

Feature selection vs feature reduction

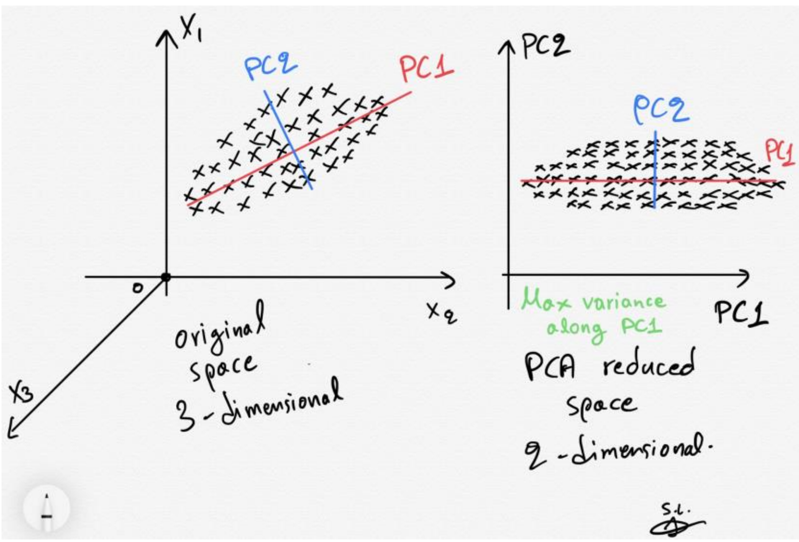
- *Feature Selection* seeks a *subset* of the n original features which retains most of the relevant information
  - Wrappers (e.g. forward selection), Filters (e.g. PMI), Embedded (e.g. Lasso, Regularized SVM)
- *Feature Reduction combines/fuses* the n original features into a smaller set of newly created 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 one of the most common feature reduction techniques

- A linear method for dimensionality reduction
- Allows us 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



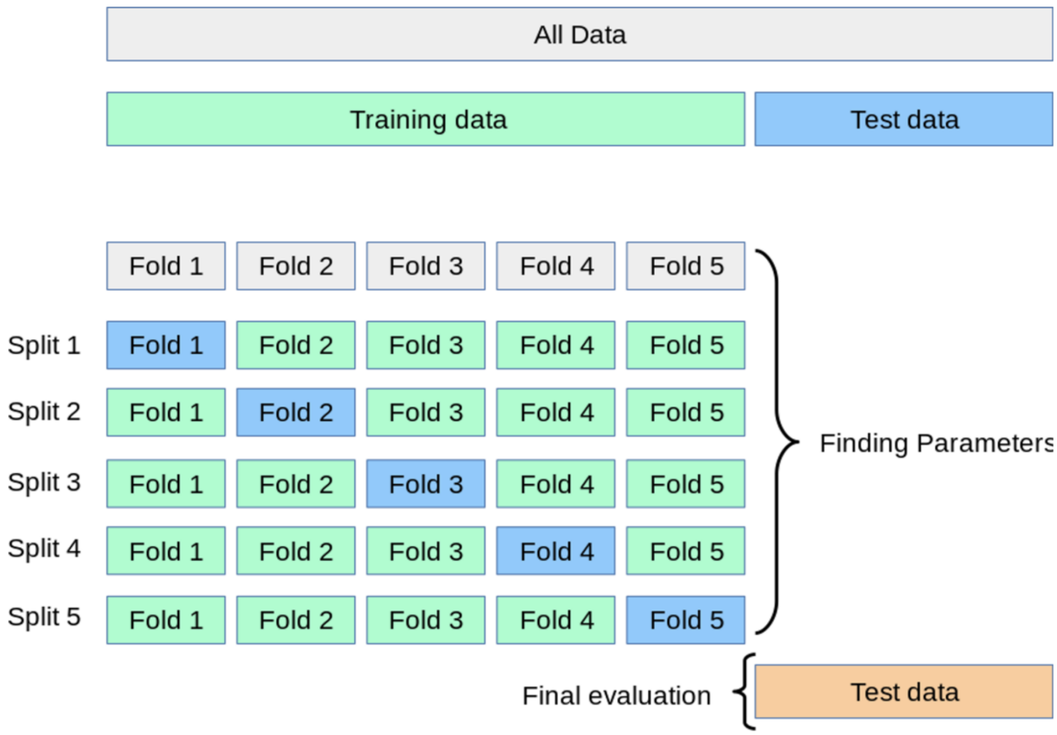
https://towardsdatascience.com/

Evaluation| Supervised learning | Which method to use?

Data Splitting

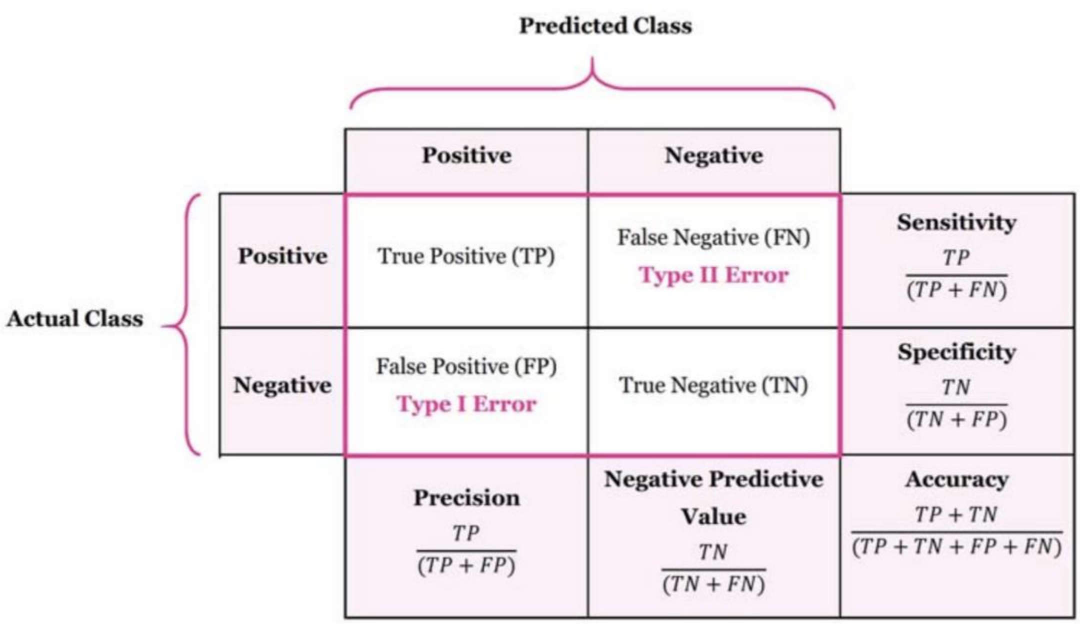
- Training set
  - Validation set (dev set)
    - A validation dataset is a dataset of examples used to tune the hyperparameters (i.e. the architecture) of a classifier. It is sometimes also called the development set or the "dev set".
- Test set

Cross Validation



https://scikit-learn.org/stable/modules/cross_validation.html

Confusion matrix



Accuracy

- What proportion of instances is correctly classified?

  (TP + TN) / (TP + FP + FN + TN)

- Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed.

- Let us say that our target class is very sparse. Do we want accuracy as a metric of our model performance? What if we are predicting if an asteroid will hit the earth? Just say “No” all the time. And you will be 99% accurate. The model can be reasonably accurate, but not at all valuable.

Precision and recall

- Precision: % of selected items that are correct Recall: % of correct items that are selected
- Precision is a valid choice of evaluation metric when we want to be very sure of our prediction.
- Recall is a valid choice of evaluation metric when we want to capture as many positives as possible.

A combined measure: F

A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The harmonic mean is a very conservative average;

Balanced F1 measure - i.e., with $\beta = 1$ (that is, $\alpha = 1/2$): $F = 2PR/(P + R)$

Summary

Summary

- Feature selection for text
- Different methods
- Can be quite effective!

Practical 3