Lecture 9-2:
Feature Selection

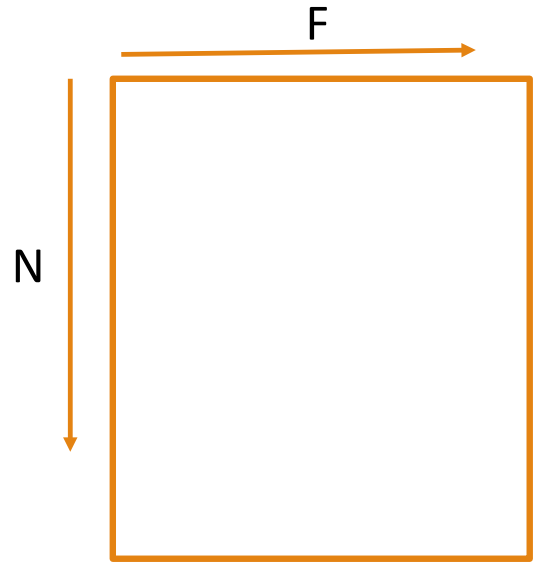
CMPT 733, SPRING 2017

JIANNAN WANG

What? and Why?

Data are often in the form of a table

- N: # of training examples (e.g., tweets, images)
- F: # of features (e.g., bag of words, color histogram)



Feature Selection

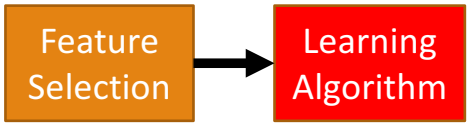
- Selecting a subset of features for use in model construction.

What's bad about "Big F"?

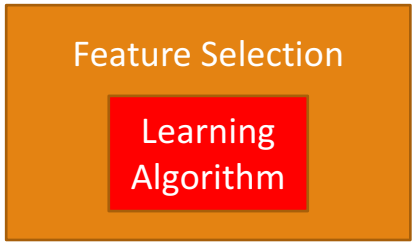
- Slow (training/testing time)
- Inaccurate (due to overfitting)
- Hard to interpret models

How?

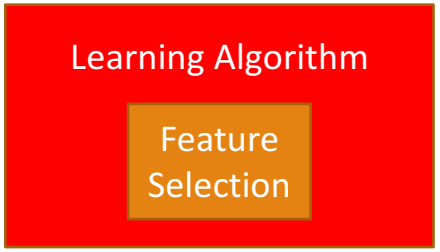
Filter Method



Wrapper Method



Embedded Method



Filter Method

Basic Idea

- Assign a score to each feature
- Filter out useless features based on the scores

Many popular scores [see Yang and Pederson ’97]

- Classification: Chi-squared, information gain, document frequency
- Regression: correlation, mutual information

Wrapper Method

Basic Idea

- Evaluate subsets of features
- Select the best subset

How to evaluate a subset of features?

- Test Error (estimated by cross validation)

How to find the best subset?

- Greedy Algorithms (e.g., forward selection, backward elimination)

Embedded Method

Basic Idea

- Modify a learning algorithm such that it can automatically penalize useless features

Lasso Regression

$$\underset{\beta}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Penalize useless features

Comparisons

Filter Method

- 😀 Good for preprocessing
- ☹ Fails to capture relationships between features

Wrapper Method

- 😀 Capture relationships between features
- ☹ Highly inefficient

Embedded Method

- 😀 Combine the advantages of the above methods
- ☹ Specific to a learning algorithm

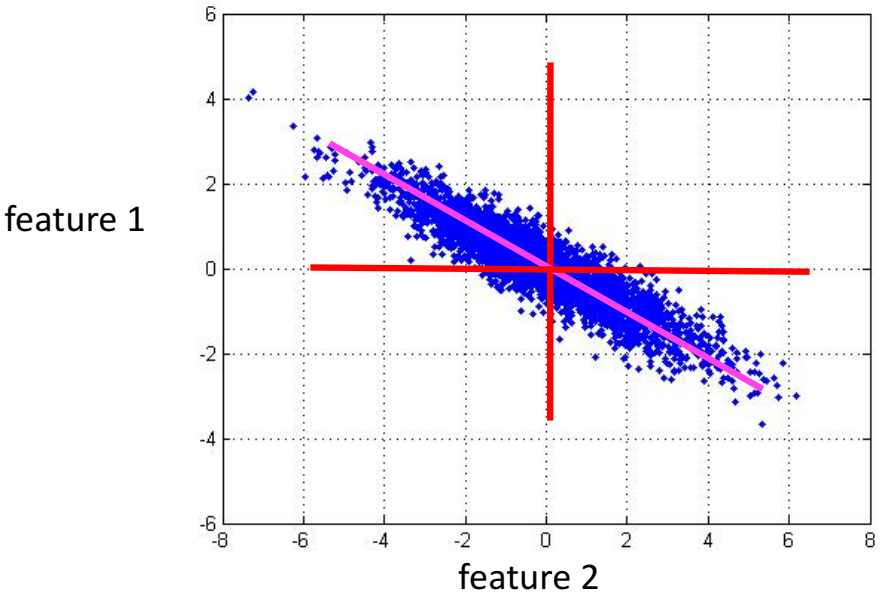
Dimensionality Reduction

Feature Selection

- New features have to be a subset of old features

Feature Transformation (e.g., PCA)

- New features may NOT be a subset of old features



Conclusion

Why feature selection?

Feature-selection methods

- Filter method
- Wrapper method
- Embedded method

Comparisons of the three methods

Assignment 9

Part 2: Feature Selection

- Task B. Filter-based Method
- Task C. Principal Component Analysis (PCA)

Deadline: 11:59pm, Mar 26th

http://tiny.cc/cmpt733-a9