# TIM 245 Lecture 6 (4/19/17)

## Agenda

1) Review Homework 1

2) Feature Selection using Information Gain

3) Wrapper based methods and general comments on feature selection

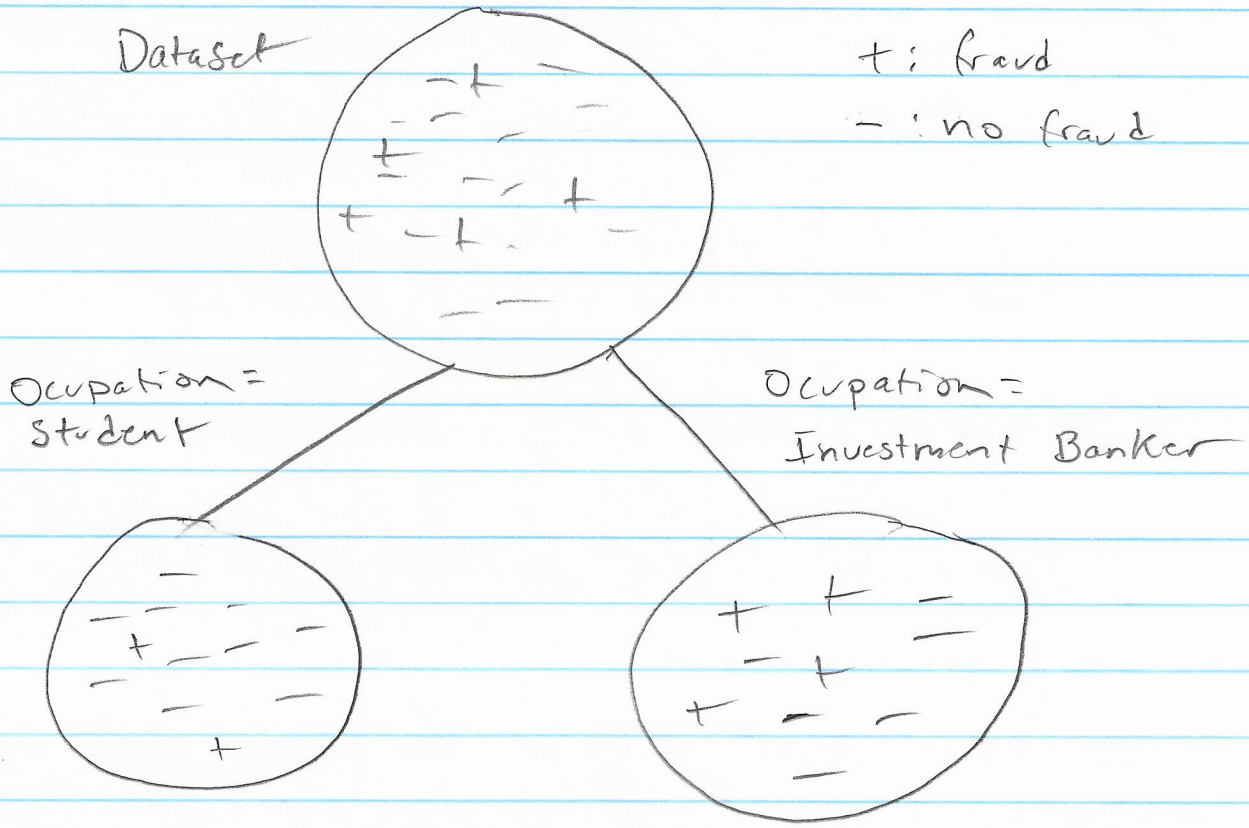
4) Dimensionality Reduction: PCA

5) Project Phase II and roadmap for the course

## ② Feature Selection using Information Gain

Each attribute gives us information on the target y.

Example : Fraud Detection




How can we select the attributes that provide the most information on y?

Let:

$c_i^Y \triangleq$ ith ($i = 1, 2, \ldots, r$) of the target $y$

$c_K^{A_j} \triangleq$ Kth ($K = 1, 2, \ldots s$) of $A_j$

$p(c_i^Y) \triangleq$ probability of $c_i^Y$ estimated as $|c_i^Y| / |D|$

Process for computing information gain of attribute $A_j$:

1) Compute the entropy of the complete dataset D

$$\text{Entropy}(D) = -\sum_{i=1}^{r} p(c_i^Y) \log_2 (p(c_i^Y))$$

2) Partition D into S subsets where partition $D_K$ contains only instances where $A_j = c_K^{A_j}$

$$\text{Entropy}_{A_j}(D) = \sum_{K=1}^{S} \frac{|D_K|}{|D|} \times \text{Entropy}(D_K)$$

3) Compute the information gain

$$\text{Gain}(A_j) = \text{Entropy}(D) - \text{Entropy}_{A_j}(D)$$

Information gain can be biased towards attributes with a large number of values.

Gain ratio is a normalized version of information gain.

$$SplitInfo_{A_j}(D) = -\sum_{K=1}^{S} \frac{|D_K|}{|D|} \times \log_2\left(\frac{|D_K|}{|D|}\right)$$

$$GainRatio(A_j) = \frac{Gain(A_j)}{SplitInfo_{A_j}(D)}$$

## ③ Wrapper Based Methods

Given a particular learning algorithm, e.g. kNN, we can search for the subset of attributes that provide the best performance

This is called a wrapper based method.

Search can be forwards or backwards through the attributes:

$$\{\emptyset\} \rightarrow \{A_1\} \rightarrow \{A_1, A_2\}$$

$$\{A_1, A_2, A_3\} \rightarrow \{A_1, A_2\}$$

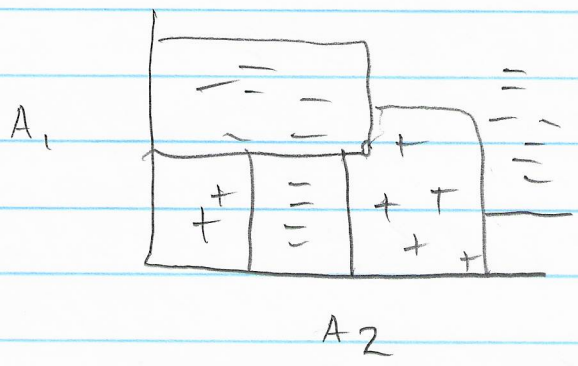
Solved using hill climb or greedy algorithm

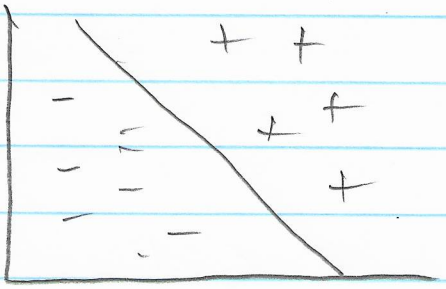
Result is the best set of attributes for the specific algorithm.

## General Comments on Feature Selection

1) Start with fast filter based methods to come up with a preliminary set of attributes

Correlation and information gain find different kinds of relationships



Better suited to information gain

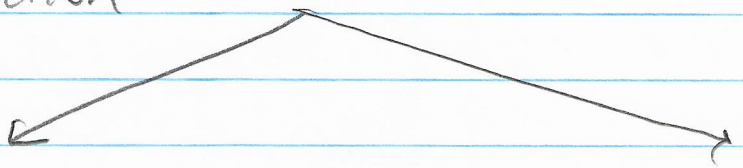
Better suited to correlation

2) Experiment with a variety of different learning algorithms

3) Apply wrapper based feature selection using the best learning algorithm

## ④ Dimensionality Reduction

Two general approaches to dimensionality reduction



**Linear**

Principle Components Analysis (PCA)

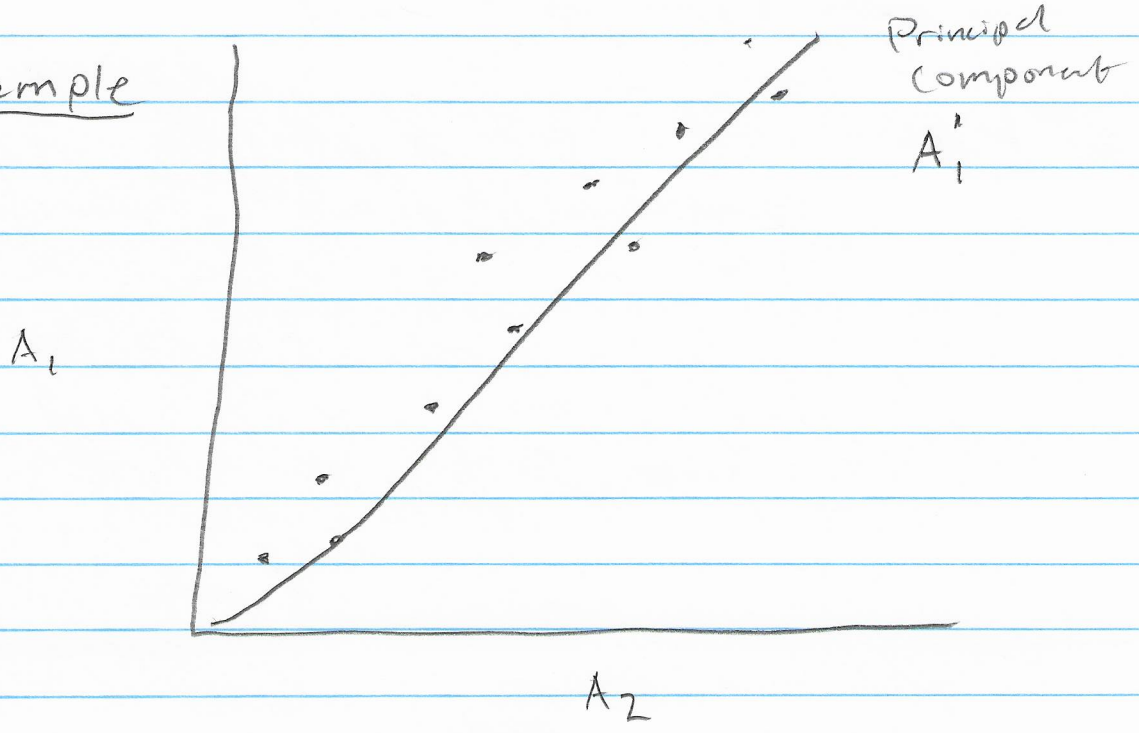
**Non-linear**

Neural Networks

### Principle Components Analysis

Find the internal "axes" of the data set. Each "axes" becomes a new high-level feature

**Example**



Assumptions:

1) Relationship between the variables is linear

2) Mean and covariance is important

3) Large variances have important dynamics