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

Dataset

+ + +
- - -
+ + -

Occupation = Student

Occupation = Investment Banker

+ + +
- - -
+ + +

How can we select the attributes that provide the most information on y?
Let:
\[ C_i^y \triangleq \text{ith (} i = 1, 2, \ldots, r \text{) of the target } y \]
\[ A_j^y \triangleq \text{kth (} k = 1, 2, \ldots, s \text{) of } A_j \]
\[ p(C_i^y) \triangleq \text{probability of } C_i^y \text{ estimated as } \frac{1}{n} \frac{1}{1 \text{ D}} \]

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 \( A_j = C_k^A \)

\[
\text{Entropy}_{A_j}(D) = \sum_{k=1}^{s} \frac{|D_k|}{|D|} \times \text{Entropy}(D_k)
\]

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.

\[
Split\text{–}Info_{A_j}(D) = -\sum_{k=1}^{s} \frac{|D_k|}{|D|} \times \log_2 \left( \frac{|D_k|}{|D|} \right)
\]

\[
Gain\text{–}Ratio(A_j) = \frac{Gain(A_j)}{Split\text{–}Info_{A_j}(D)}
\]
(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:

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

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

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.

A1

Better suited to information gain

A2

Better suited to correlation

2) Experiment with a variety of different learning algorithms

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

![Diagram showing two axes A1 and A2 with data points on a line]
Assumptions:

1) Relationship between the variables is linear.

2) Mean and covariance is important.

3) Large variances have important dynamics.