TIM 245 Lecture 5 (4/17/17)

Agenda

1) Comments on Project Phase I and Homework 1

2) Normalization (Complete Lecture 4)

3) Types of Data Reduction

4) Sampling (instance reduction)

5) Feature Engineering: feature selection (Attribute Reduction)

6) Work on the project (time permitting)

③ Types of Data Reduction

Why do we need to do data reduction?

Two cases

1) Processing the complete dataset is too expensive or time consuming (too many instances)

2) using the Complete dataset will result in a poor predictive model or patterns that are not useful (too many attributes)

Computational Complexity (case 1)

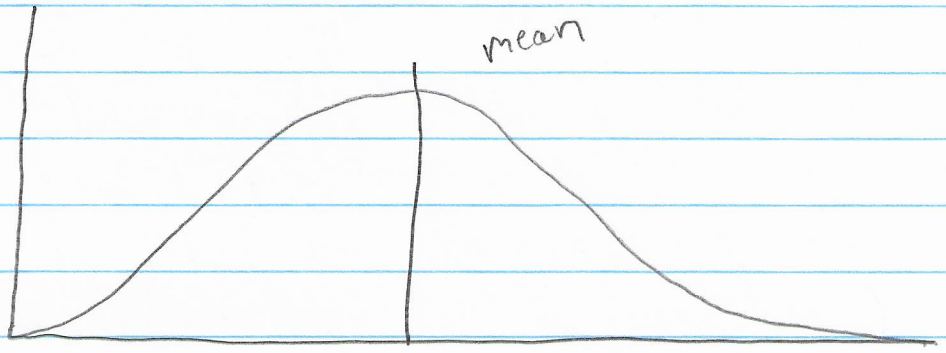
It can be useful to reduce the number of instances during model selection to enable faster iteration.

However, once the model selection process is finished, use either the entire dataset or use increasingly large samples until the performance stabilizes.

Curse of Dimensionality (case 2)

High dimensional data is unexpectedly sparse. All points become uniformally distant from eachother and the distinction between for and near becomes meaningless.

Example: Multivariale Normal



④ Sampling

The goal of Sampling is to get a representative subset of the instances $d_1, d_2, \ldots d_n$

Types of Sampling

1) Simple Random Sample (SRS): draw set of s instances from D ($s < n$), where the probability of drawing any instance is $\frac{1}{n}$. Can be done with or without replacement.

2) Stratified Sample: divide D into mutually disjoint parts or Stratas, and perform a SRS for each strata.

e.g. fraud = {yes, no}

perform SRS for each attribute value

(5) Feature Engineering

The goal of feature engineering is to get a compact representation of the attributes $A_1, A_2, \ldots A_m$

Three general approaches to feature engineering:

1) Manually remove and/or combine attributes based on domain knowledge
   e.g. id

2) Select a subset of the attributes that have the most predictive power
   (Feature Selection)

3) Transform the dataset into a lower dimensional space
   (Dimensionality Reduction)

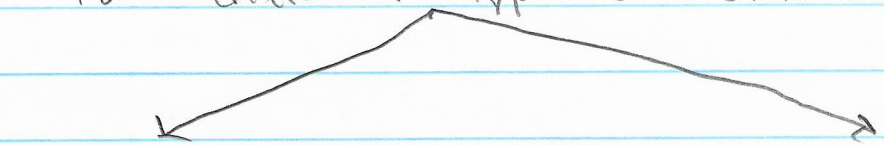
Recommended process: 1 → 2 → 3

Dimensionality reduction can be very powerful but turns the model into a "black box".

Two general approaches to feature selection:

1) Filter Methods: Score each attribute using a statistical measure that approximates the predictive power, and then select the highest scoring attributes

Two different types of statistical measures

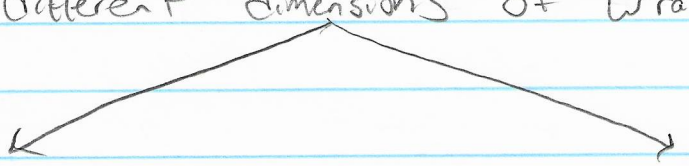


Correlation

Information Gain

2) Wrapper Methods: Evaluate the performance of models created using different subsets of attributes and then select the highest scoring subset

Two different dimensions of wrapper methods



Search method
- Best first
- Greedy Selection

Learning Algorithm
- SVM, DT, NN

Correlation Based Feature Selection

Numerical target $\rightarrow R^2$ correlation coefficient

Categorical target $\rightarrow \chi^2$ chi squared test

χ^2 statistic tests the hypothesis that attributes A_j and A_K are independent

Let:

$$A_j = \{ C_1^{A_j}, C_2^{A_j}, \ldots, C_r^{A_j} \}$$

$$A_K = \{ C_1^{A_K}, C_2^{A_K}, \ldots, C_s^{A_j} \}$$

$(C_i^{A_j}, C_L^{A_K}) \triangleq$ event attribute A_j takes value $C_i^{A_j}$ and A_K takes value $C_L^{A_K}$

o_{iL} = observed frequency of $(C_i^{A_j}, C_L^{A_K})$

e_{iL} = expected frequency of $(C_i^{A_j}, C_L^{A_K})$

$$= \frac{\text{count}(C_i^{A_j}) \times \text{count}(C_L^{A_K})}{n}$$

where n is the total number of instances.

Contingency Table: Location and Rainy Days

	San Jose	Santa Cruz	Total
Rain	108 (105.5)	103 (105.5)	211
No Rain	257 (259.5)	262 (259.5)	519
Total	365	365	730

O_{iL}

e_{iL}

$$\chi^2 = \sum_{i=1}^{r} \sum_{L=1}^{s} \frac{(O_{iL} - e_{iL})^2}{e_{iL}}$$

Test is based on a significance, 0.001, With $(r-1) \times (c-1)$ degrees of freedom.

Example:

$$\chi^2 = \frac{(108-105.5)^2}{105.5} + \ldots + \frac{(262-259.5)^2}{259.5}$$

$$= 0.1667$$

For $(2-1) \times (2-1) = 1$ degree of freedom, the χ^2 value to reject at 0.001 significance is 10.828. The values are independent.