Lecture 3 (4/10/17)

Agenda

1) General Comments on the Project

2) Formalization of the Data Mining tasks

3) Exploratory Data Analysis (EDA)

4) Outlier analysis

② Formalization of the Data Mining Tasks

Let:

$D \triangleq$ dataset of n instances and m attributes

$d_i \triangleq$ ith instance $(i=1,2,\dots n)$

$A_j \triangleq$ jth attribute $(j=1,2,\dots m)$

$X_{ij} \triangleq$ value of Attribute A_j for instance d_i

	A_1	A_2	A_3	...	A_m
d_1	X_{11}	X_{12}	X_{13}		X_{1m}
d_2					
$\vdots$					
d_n	X_{n1}	X_{n2}	X_{n3}		X_{nm}

Predictive Analysis (Supervised Learning)

Let attribute, A_j, be the target Y

$$d_i = (x_{i1}, x_{i2}, \dots x_{im}, y_i) \triangleq (X_i, y_i)$$

Learn function $\Phi : X \to Y$

classification : y is a nominal attribute

prediction : y is a numerical attribute

Descriptive Analysis (Unsupervised Learning)

Cluster Analysis

Organize the dataset D into groups such that all instances in a group are similar to eachother and dissimilar to instances in other groups based on some distance metric

Association Analysis

Let X and Y be sets of attribute values such that $X \cap Y = \emptyset$

Association rule $X \Rightarrow Y$ means that instances containing values from Y also contain values from X

③ Exploratory Data Analysis

The purpose of data cleaning is to remove noise and inconsistencies that do not reflect the real-world process that created the data.

In order to do this accurately, we first need to understand the underlying process that Created the data.

There are two ways to understand a dataset:

1) Domain Knowledge: fit the data to the process ("Inside-out")

2) Exploratory Data Analysis: fit a process to the data ("Outside-In")

EDA Objectives

1) What is a typical value?

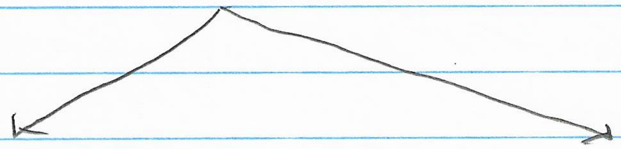
2) What is the uncertainity for a typical value?

3) What is a good distributional fit?

4) Does the attribute affect other attributes?

5) Does the data contain outliers?

Two general approaches to EDA



Visual Exploration (Visualization)

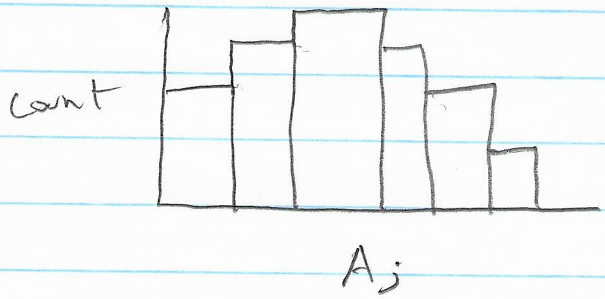
Quantitative Exploration (Descriptive Statistics)

Visual and quantitative exploration should be done in parallel

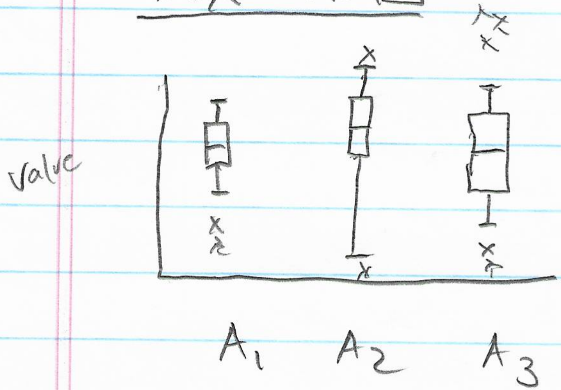
Visual Exploration

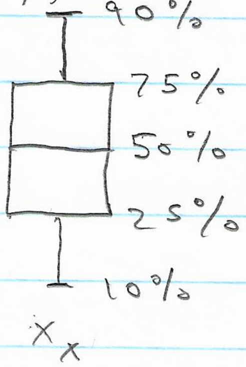
Always plot the data!

<u>Histogram</u> → understand typical values, uncertainity, and distribution

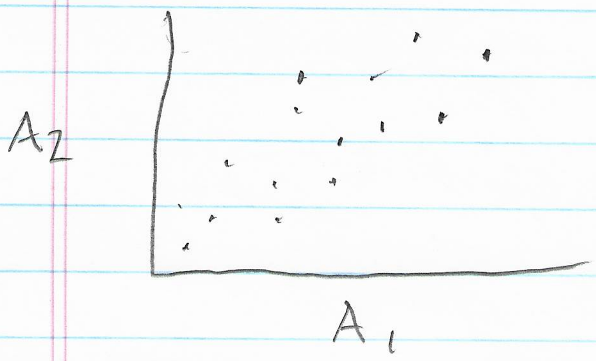


<u>Box Plots</u> → comparing typical values, uncertainity, outlier detection





<u>Scatter Plot</u> → understanding relationship between variables



Quantitative Exploration

Three general ways of describing how attribute, A_j, behaves:

1) Central Tendency: What is the typical value of A_j

2) Dipersion: What is the spread of A_j

3) Correlation: what is the relationship between A_j and another attribute A_K

Measures of Central Tendency

Mean (average) : $\frac{1}{n} \sum_{i=1}^{n} X_{ij}$

Median : middle valve of $X_{1j}, X_{2j}, \ldots X_{nj}$ (sorted)

Mode : most frequent valve of $X_{1j}, X_{2j} \ldots X_{nj}$

Mean uses all of the data but is sensitive to outliers

Median is robust to outliers but can be sensitive to small changes

Mode useful for nominal data but rarely used for numerical data

Measures of Dispersion

Variance $(\sigma_{A_j}^2) = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{A}_j)^2$

std dev $(\sigma_{A_j}) = \sqrt{\sigma_{A_j}^2}$

Median Absolute Deviation (MAD) $= \text{median} (|X_{1j} - \bar{A}_j|, |X_{2j} - \bar{A}_j|, \ldots |X_{nj} - \bar{A}_j|)$

Variance and std dev are sensitive to outliers

Median Absolute Deviation is useful for data sets with long-tails and/or outliers

Measures of Correlation

Pearson Correlation Coefficient (R^2)

$$\text{Covariance}(A_j\ A_K) = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{A}_j)(X_{iK} - \bar{A}_K)}{n-1}$$

$$\text{Correlation}(A_j, A_K) = \frac{\text{COV}(A_j\ A_K)}{\sigma_{A_j}\ \sigma_{A_K}}$$

Nominal Attributes → Chi Squared Test (x^2)

④ Outlier Analysis

Two general approaches to outlier analysis

Qualitative

Outliers are values that should not occur based on the underlying process (Domain Knowledge)

Quantitatively

outliers are values that are different from the rest of the data (EDA)

Two general approaches to quantitative outlier anaysis

Parametric

Data that is unlikely (low probability)

Non-Parametric

Data that is significantly far away from the central tendency

Process for Parametric Outlier Analysis

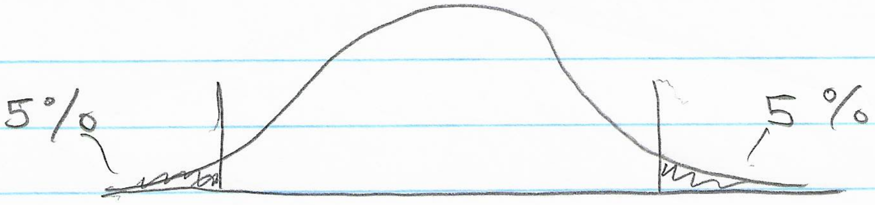
1) Fit the data to an appropriate theoretizal distribution

e.g. Normal, Binomial, Exponential

2) Select a probability cutoff for outliers

e.g. 10%

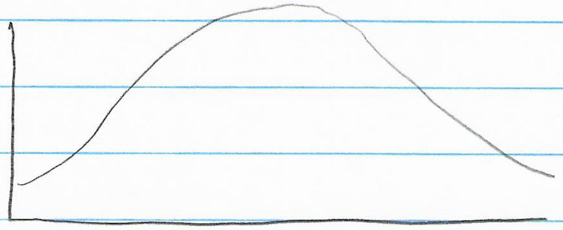
3) Determine the cutoff value based on the PDF/CDF



Process for Non-Parametric outlier Analysos

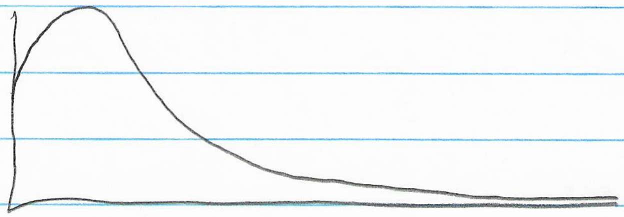
1) Determine the appropriate measure for central tendency and dispersion

Symmetric Histogram



→ mean
Std dev

Skewed / long-tail histogram



→ Median
MAD

2) Determine the appropriate threshold based on visual exploration and domain knowledge

Cutoff = central tendency ± K spread