TIM 245 Lecture 4 (4/12/17)

Agenda

1) Phase I Project Proposal and Homework 1

2) Review of Lecture 3

3) Motivating Example

4) Data Cleaning, Integration, and Transformation

③ Motivating Example

<u>Problem:</u> Predict if a tax return is fraudulent using historical data

SSN	Name	Age	Ocupation	Employer	Income	Fraud
111-11-1111	John Smith	22	Student	UCSC	30K	No
222-22-2222	Bob Smith	20	Student	UC Santa Cruz	200K	Yes
⋮						

Learn $\Phi : X \rightarrow Y$ Fraud = {yes, no}

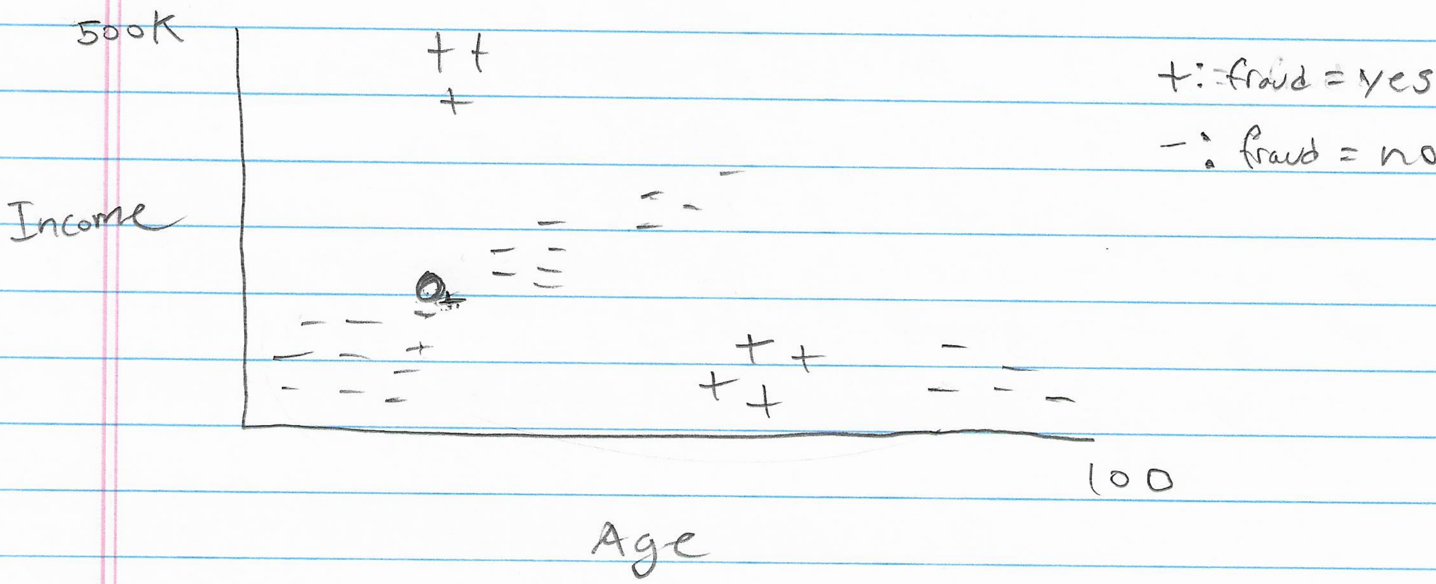
Age, Income, Employer ...

A Simple Classification Model

Use a majority vote of the closest instances in the historical dataset

This model is called K-Nearest-Neighbors. The value K controls the number of neighbors used in the vote.

Simplified KNN Model



Closeness is determined based on a distance function.

Example : Euclidean distance

$$dist(d_i, d_L) = \sqrt{\sum_{j=1}^{m} (X_{ij} - X_{Lj})^2}$$

Missing Values :

both missing → 1

one missing → Max($|X_{ij}|$, $|X_{Lj}|$)

Nominal Values :

equal → 0

not equal → 1

(4) Data Cleaning, Integration, Transformation

Data cleaning issues:

1) Missing values

2) Duplicates

3) Outliers

4) Noise / Inconsistencies

Missing Values

Example: 45K, 60K, NA, 35K

Approaches:

1) Ignore or remove

2) Fill in manually based on domain knowledge / EDA

   a) one instance rule
   b) rule for entire dataset

3) Fill in using central tendency of the attribute (mean, median, mode)

4) Predict based on other attributes e.g. regression

Duplicates

Example: John Smith, John Smith

Approaches:

1) use a unique identifier

2) Use a combination of attributes that approximate a unique identifier

Outliers

Example: 45K, 60K, 500K, 35K

Approaches:

1) Do nothing, leave in the data set

2) Remove instances from the data set

3) Replace attribute values (see missing value)

Noise / Inconsistencies

Noise = measurement error, random flucuations, variation that isn't relevant to the problem

e.g. 30,500, 31,250, 30,750

Inconsistency = human recording "error" in nominal values

e.g. UCSC, UC Santa Cruz, University of California Santa Cruz

Approaches to cleaning Noise:

1) Group the data into bins and perform local smoothing

   a) Smooth by bin mean
   b) Smooth by bin median
   c) Smooth by bin boundries

2) Predict based on other attributes, e.g. regression

Approaches to Cleaning Inconsistencies

1) Write rules based on domain Knowledge and EDA, i.e. find and replace

2) Used Named Entity Recognition (NER) to map to a canonical value

3) Cluster instances with similar values into groups and then write a rule for the entire group of instances

Edit distance is a useful similarity measure for clustering nominal string values

Edit distance $\triangleq$ minimum number of substitutions to transform X_{ij} into X_{kj}

Hammings distance : Substitutions
Levenshtein distance : deletions, insertions, and substitions

Example : ATT → AT&T = 1
UCSC → UC Santa Cruz = 9
ATT → UCSC = 4

Data Integration

Combine datasets from multiple sources into a single coherent data sets

Reasons for data integration

1) Bring in additional attributes to potentially improve model performance

2) Want to predict an attribute that is not currently in the data set

3) Find association rules for attributes not in the dataset

4) Want better cluster seperation

Approaches:

1) Join on a unique identifier, e.g. SSN

2) Use a combination of attributes that approximate a unique identifier and fuzzy matching (e.g. edit distance)

3) Use NER to create a unique identifier

Data Transformation

Changing attribute values to make them

1) More suitable for data mining algorithms

Balance the weight of numerical attributes in distance based models

Example: Age vs Income in Eucliden space

Methods: Min-Max normalization
Z score normalization

2) More suitable to human interpretation

Generalization of nominal attributes to higher level concepts

Example: UCSC, UCB, UCSD
↓
university

Methods: EDA, Domain knowledge, rules

Min-Max Normalization

Linear transformation that maps the values of a numerical attribute to the range [a, b]

Let x'_{ij} denote the normalization of x_{ij}, then

$$x'_{ij} = \frac{x_{ij} - \text{Min}(A_j)}{\text{Max}(A_j) - \text{Min}(A_j)} (b-a) + a$$

Example: Suppose that the minimum and maximum income are \$5,000 and \$500,000, respectively.

We want to map income to the range [0, 1]

$$x_{ij} = 45{,}000$$

$$x'_{ij} = \frac{45{,}000 - 5{,}000}{500{,}000 - 5{,}000} (1-0) + 0$$

$$= 0.08$$

Z Score Normalization

Values are normalized based on the mean and standard deviation of A_j

$$X'_{ij} = \frac{X_{ij} - \bar{A}_j}{\sigma_{A_j}}$$

where $\bar{A}_j$ and σ_{A_j} are the mean and std dev of A_j, respectively

Example: $\bar{A}_j = 52{,}000$

$\sigma_{A_j} = 5{,}000$

$X'_{ij} = -1.4$