**DATA PREPROCESSING**

**1. Handling Missing Data**

1.1 Delete Records Containing Missing Values

Pros: Easiest way

Cons:

* Not necessarily best approach
* Pattern of missing values may be systematic
* Deleting records creates biased subset
* Valuable information in other fields lost

1.2 Replace Missing Values with User-defined Constant (0 for numerical and "missing" for categorical)

Pros:

Cons:

1.3 Replace Missing Values with Mode or Mean

Pros:

Cons:

- Mean not always best choice for “typical” value

- Resulting confidence levels for statistical inference become overoptimistic

- Domain experts should be consulted regarding approach to replace missing values

1.4 Replace Missing Values with Random Values

Pros:

* Method superior compared to mean substitution

Cons:

* No guarantee resulting records make sense

**2. Identify Misclassification**

**2.1 Dealing with Outlier**

A histogram examines values of numeric fields

Two-dimensional scatter plots help determine outliers between variable pairs

**2.2 Data Transformation**

Variables with greater ranges tend to have larger influence on data model’s results. Therefore, numeric field values should be normalized.

* Min-max normalization:

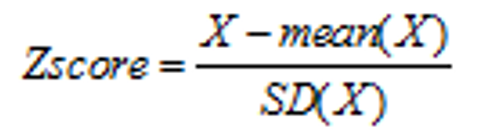
[0,1]

Text, letter

Description automatically generated with medium confidence

* Z-score standardization

Z-score means how many SDs below or above the mean



* Decimal scaling

[-1,1]

Chart

Description automatically generated

Measuring the skewness of the distribution informs us of symmetry



**Binning Numerical Variables**

Equal width (affected by outliers), Equal frequency, Clustering

**Remove Not Useful Variables**

Unary (include single value), with 90% missing, strongly correlated, duplicate records

ID field filtered out, not remove

**EXPLORATORY DATA ANALYSIS CODE**

<https://jupyter.utoronto.ca/user/ac08bcce-7102-46af-bb12-ed9c350736f7/notebooks/Exploratory%20Data%20Analysis.ipynb>

**K NEAREST NEIGHBOR**

Classification is a supervised method

**1. Distance Function**

Continuous data values should be normalized using Min-Max Normalization or Z-Score Standardization to avoid overwhelming



Categorical data



The use of different normalization techniques resulted in different nearest result.

Example: 50-year-old male: a 20-year-old male or a 50-year-old female

When calculating the distance between records containing both numeric and categorical attributes, the use of Min-Max Normalization is preferred

**2. Simple Unweighted Voting**

**3. Weighted voting:** closer neighbors have a larger voice in the classification decision than do more distant neighbors

Table

Description automatically generated with medium confidence

Chart

Description automatically generated

**4. Stretching the Axes**



Stretching the axes leads to improved accuracy by quantifying the relevance of each variable used in the classification decision

**5. Database Considerations**

A screenshot of a computer

Description automatically generated with medium confidence

Table

Description automatically generated

Table

Description automatically generated

Calculation



**6. Choosing K**

• Small k:

- Advantage:

- a) The computation power is relatively low with small k, which means the model is easy to run.

- Disadvantage:

- a) It is possible that the classification or estimation may be unduly affected by outliers or unusual observations (“noise”).

- b) The algorithm will simply return the target value of the nearest observation, which may lead the algorithm toward overfitting, tending to memorize the training dataset at the expense of generalizability.

• Large k:

- Advantage:

- a) Reduce the risk of overfitting issue.

- b) Larger values will tend to smooth out idiosyncratic or obscure data values in the training dataset.

- Disadvantage:

- a) If the k is too large, locally interesting values will be overlooked.

- b) The model becomes too generalized and fails to accurately predict the data points in both train and test sets, which may cause the underfitting issue.

**DECISION TREE**



Decision Tree is supervised classification method.

* CART
* C4.5

CART

1. Requirement of using:

1.Require preclassified target variables. A training data set must be supplied which provides the algorithm with the values of the target variable

2.This training data set should be rich and varied, providing the algorithm with a healthy cross section of the types of records for which classification may be needed in the future.

3. Only formulas and examples for a discrete target attributeare discussed here

2. Classification and Regression Trees (CART)

2.1 Splits at decision nodes are binary

2.2 Measurement:

* + Optimality measure maximizes *Φ*(*s|t*) over all possible splits, *s*, at node *t*



* + Gini index: Reaches its minimum (zero) when all cases in the node fall into a single target category

Diagram

Description automatically generated

Calculation of GINI INDEX



3. Classification error rate

分错组的概率

4. Pruning

Fully grown tree has lowest classification error rate but overfits the training set

C4.5

1. Intro

* C4.5 builds tree by recursively visiting decision nodes and choosing optimal splits, until no further splits possible
* Key Differences Between CART and C4.5
  + Unlike CART, C4.5 is not limited to binary splits and produces tree with variable shape
  + C4.5 produces branch for each categorical value. This may result in “bushiness”
  + C4.5 uses different algorithm to measure homogeneity occurring at leaf nodes
  + C4.5 uses information gainor entropy reduction to select optimal split at each decision node

2. Entropy

–log2(*p*)

Entropy of X :

Text, letter

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Calculation of ENTROPY



Graphical user interface, text, application, letter

Description automatically generated

**NEURAL NETWORK**

Supervised learning

1. Input and output encoding

* Output: between 0 and 1 using sigmoid function (combines nearly linear behavior, curvilinear behavior, and nearly constant behavior, depending on the value of the input)
* Input: between 0 and 1 with standardization
  + Continuous: min-max normalization: 
  + Catagorical:

Graphical user interface, text, application, email

Description automatically generated

2. Example of ANN

* ANN consists of a *layered, feedforward, completely connected* network of artificial neurons
* Every node in a given layer is connected to every node in the next layer with some weights (randomly assigned between 0 and 1)
* The number of layers and nodes may cause overfitting or underfitting issue.

Diagram

Description automatically generated

Calculate the output



3. Back propagation

通过调整weight减少SSE

3.1 Gradient Descent Method

- The slope and direction of movement has negative relations

- The large the slope, the large the movement

Δ*w*current = –η (∂SSE/∂*w*current). where η, 0 < η < 1, is the learning rate: a tuning parameter chosen to help us move the network weights toward a (global) minimum for SSE

A picture containing diagram

Description automatically generated

3.2 Back-Propagation Rules

Graphical user interface, text, application

Description automatically generated

Calculation



4. When to terminate?

Split data into training and CV. Keep training the on training data and apply the weight on CV data. Monitor the “current” weight and the “best” weight (so far has the lowest SSE) on CV data. When the “current” weights has significantly greater SSE than “best” weights, then terminate the algorithm

5. Learning rate

Learning rate should be large initially to quickly approach the neighbor of min SSE. Then reduce to avoid overshooting the min point

6. Momentum term: alpha represents inertial

Text

Description automatically generated with low confidence

Small alpha means small momentum, might undershot the min SSE, while large alpha may overshoot the min SSE

7. Sensitivity analysis

Graphical user interface, text, application, email

Description automatically generated

**TEXT MINING**

Text is unstructured data

1. Intro

* Text mining is classifying and extracting meaning from documents to interpret it like human language
* Classification (labeling) and Clustering
  + Spam filtering
  + Language identification
  + Genre classification

2. bag-of-word

* Grammar, syntax, punctuation, word order are ignored
* Effective when the goal is to decide which category or cluster a document falls in
* Requires lots of documents

3. spreadsheet model of text

Table

Description automatically generated

4. Tokenization

The process by which you identify delimiters and use them to separate terms is called *tokenization*

* Goal – reduction of text (also called vocabulary reduction) without losing meaning or predictive power
* Not case sensitive, stemming the words, frequency filter to eliminate terms that nearly or hardly appear
* Punctuation, stopwords, <= 2 character long words are dropped
* Normalization: email token, url token

5. TF-IDF

Graphical user interface, text, application

Description automatically generated

**CLUSTERING**

No target variable, not to classify, estimate or predict the target variable, but to segment entire dataset into homogeneous subgroups that max the similarity within groups.

Goal: *between-cluster* variation is large compared to the *within-cluster* variation

1. Cluster issues

1.1 Measuring distance between records

Euclidean, city-block, minkowski

Text, letter

Description automatically generated

1.2 Recode categorical variables



1.3 Normalize numeric variable

A picture containing text

Description automatically generated

2. Hierarchical Clustering Methods

* treelike cluster structure (*dendrogram*)
* *Agglomerative* methods: initialize each observation to be a tiny cluster of its own, and combine existing clusters to create tree
* *Divisive* methods: begin with all records in one big cluster and split off most dissimilar records into separate clusters

3. Determine the distance between clusters

* **Single linkage**: Based on the *nearest*-neighbor, or minimum distance between records. form long slender clusters

将两个组合数据点中距离最近的两个数据点间的距离作为这两个组合数据点的距离。这种方法容易受到极端值的影响。两个很相似的组合数据点可能由于其中的某个极端的数据点距离较近而组合在一起

* **Complete linkage**: Based on the *farthest*-neighbor, or maximum distance between records. form compact sphere-like clusters

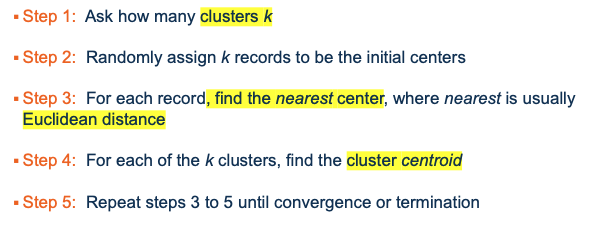
将两个组合数据点中距离最远的两个数据点间的距离作为这两个组合数据点的距离。问题：两个不相似的组合数据点可能由于其中的极端值距离较远而无法组合在一起

* **Average linkage**: Based on the average distance between all records in one cluster to all records in another cluster. Reduces dependence on extreme values

Calculation



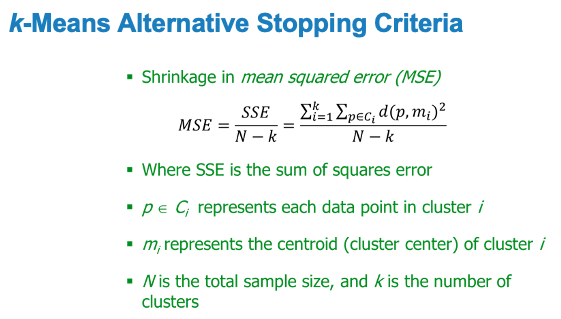
4. K-means cluster



Good clusters have large pseudo-F Statistic

Cluster i里面每个点到centroid的距离之和，再求几个组的和。

Mse越小越好，ssb越大越好，因为要组内差异小，组间差异大

Text, letter

Description automatically generated

K-mean example



5. Sum up

* K-mean not guarantee to find global max F
* May improve by changing the initial k
* May benefit from axis-stretching methods to quantify attribute relevance

6. KNN vs K-means

* *K*NN is a supervised learning algorithm mainly used for classification problems, whereas *K*-Means is an unsupervised learning algorithm.
* *K* in *K*-Means refers to the number of clusters, whereas *K* in *K*NN is the number of nearest neighbors (based on the chosen distance metric).

CLUSTER EVALUATION



**BIRCH CLUSTERING**

Graphical user interface, text

Description automatically generated

SEGMENTATION MODELS