STA 529 2.0 Data Mining

Dr Thiyanga S. Talagala

June 4, 2022

Lecture 1

What is Data Mining?

- Process of discovering **interesting patterns** of knowledge from huge amounts of data.
- **KDD: K**nowledge **D**iscovery from **D**ata/ **K**nowledge **D**iscovery from **D**ata Mining
- **Process**: Automatic or Semi-automatic
- **Interesting patterns:** Valid, Novel, Useful, Understandable

What do we mean by interesting patterns?

Example

- Retailers collect data about customer purchases at the checkout counters
- Customer purchasing patterns: Identify which items are frequently sold together?
- Products that are likely to be purchased together.

Why it is useful?

- Can make a purchase suggestion to their customers
- Gives an idea that how we can arrange items in a store to as a strategy for boosting sales.

Characteristics of Big Data: 5 V's of Big Data

1. Volume: size
2. Velocity: how quickly data is generated?
3. Variety: diversity
4. Veracity: quality of data
5. Value: how useful?

What motivates the development of data mining field?

- Scalability
- High dimensionality
- Heterogeneous and complex data
- Data ownership and distribution

Scalability: Example

Scalability: Example (cont.)

Scalability: Example (cont.)

Data Mining Tasks

1. **Predictive tasks:** Predict the value of a particular attribute based on the values of other attributes
2. **Descriptive tasks:** Find human-interpretable patterns that describe data

Data

Variables: Characteristic of an object

Features, Attributes, Dimension, Field

Object: Collection of attributes describe an object

Entity, Instance, Event case, Record, Observation

Question

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Spe
1	5.1	3.5	1.4	0.2	se
2	4.9	3.0	1.4	0.2	se
3	4.7	3.2	1.3	0.2	se
4	4.6	3.1	1.5	0.2	se
5	5.0	3.6	1.4	0.2	se

Data Quality

1. Range: How narrow or wide of the scope of these data?
2. Relevancy: Is the data relevant to the problem?
3. Recency: How recent the data is generated?
4. Robustness: Signal to noise ratio
5. Reliability: How accurate?

The Data Mining Process

The CRISP Data Mining Process

Applications

1. Web mining: recommendation systems
2. Screening images: Early warning of ecological disasters
3. Marketing and sales
4. Diagnosis
5. Load forecasting
6. Decision involving judgement

Many more...

Machine Learning Algorithms

1. Supervised learning algorithms
   Deals with labelled dataset
2. Unsupervised learning algorithm
   Deals with labelled dataset

Evaluating Predicting Performance

1. Training set
2. Validation set
3. Test set

Hyperparameters

- Parameter whose value is used to control the learning process
- These values are set before training the model

Hyperparameters - Example

Decision trees levels

Supervised learning algorithms

Outcome could be

- Numeric
- Categorical
- Probability

Loss function

- Function that calculates loss for a single data point

$e_i = y - \hat{y}$

$e_i^2 = (y - \hat{y})^2$

Cost function

- Calculates loss for the entire data sets

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_i$$

Numeric outcome: Evaluations

Prediction accuracy measures (cost functions)

Mean Error

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_i$$

- Error can be both negative and positive. So they can cancel each other during the summation.

Mean Absolute Error (L1 loss)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

Mean Squared Error (L2 loss)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Mean Percentage Error

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i}$$

Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{e_i}{y_i}|$$

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

Visualizaion of error distribution

Graphical representations reveal more than metrics alone.

Accuracy Measures on Training Set vs Test Set

Accuracy measure on training set: Tells about the model fit

Accuracy measure on test set: Model ability to predict new data

Example

Evaluate Classifier Against Benchmark

Naive approach: approach relies soley on Y

Outcome: Numeric

Naive Benchmark: Average ($\bar{Y}$)

A good prediction model should outperform the benchmark criterion in terms of predictive accuracy.

Outcome: Categorical

Can you give an example for a Naive rule?

Accuracy evaluation: Categorical

Confusion matrix/ Classification matrix

		Actual	
		Yes	No
Predicted	Yes	a	c
	No	b	d

$$\text{error} = \frac{c+b}{n}$$

$$\text{accuracy} = \frac{a+d}{n}$$

Performance in Case of Unequal Importance of Classes

Suppose the most important class is "Yes"

$$\text{sensitivity} = \frac{a}{a+b}$$

$$\text{specificity} = \frac{d}{c+d}$$

$$\text{False Discovery Rate} = \frac{b}{a+b}$$

$$\text{False Omission Rate} = \frac{c}{c+d}$$

Your turn

What is ROC curve?

Accuracy measures for class imbalance datasets?