Chapter 2 Business Problems and DS Solutions

10/25/19 12:31 AM

Transform business problems into data mining tasks

Classification problems (supervised)
- Classify individuals into several exclusive classes
- Very close to scoring models, which estimate the probability

Regression problems (supervised)
- Estimate a value (continuous)

Similarly matching problem (both)
- Identify similar individuals, most common for making product recommendations (based on similarity of individuals to recommend product purchased by another person)

Clustering (unsupervised)
- Group individuals based on similarity but not for any purpose
- Useful in preliminary domain exploration, to observe whether there are natural segments or groups formed, in order to find potential data mining problems

Co-occurrence grouping (unsupervised)
- Identify entities that appear together (usually in transaction), e.g. ground meet and hot sauce
- Can also be used in recommendation system

Profiling (unsupervised)
- Characterize behavior of certain individuals or groups. Behavior can be more than one description
- Often used in anomaly detection such as fraud detection (based on consumption behavior, detect fraud activity)

Link prediction (both)
- Decide whether a link between data should exist or not. E.g. have lots of common friends, suggest they should be friends as well
- Can also estimate the stength of a link

Data reduction (both)
- Reduce size of data, to have smaller set but contain most important information from original data set. Usually will involve trade-off

Causal modelling (supervised)
- Determine the causal relationship, commonly by randomly control

Supervised v.s. Unsupervised Methods

Supervised learning: requires a specific target to be provided
- Target must exist in data

Unsupervised learning: grouping but no guarantee to be meaningful for specific purpose

General data mining process and considerations

Will be discussed in details in later chapters