Data Analytics

Bryan Hoang

2021-11-14

Table Of Contents

1 Lecture 6

2021-10-20

1.1 Introduction

- Using basketball dataset (win loss & margins) found on OnQ under week 6.
- What's the point of these datasets?
  - The use of a model made form this data depends from case to case. There's no single rule for the usefulness of a model.
  - Hindsight?

1.2 Design = Good Choices

Figure 1.1: Design diagram

- Generic workflow: file read → binning & normalization & missing value inputation → model → scorer
  - Diagram (Based James?)
  - Random forests and Bagging/boosting diverge from the more typical workflows.
- When would one use binning?

  - Mapping numerical values to categorical values which better reflect meaning (e.g., numeric ages to categorical age ranges)
- When would one use normalization?
  - Depends on what will happen later in the workflow.
  - e.g., using neural networks, k-nearest neighbour
- Should try to design from ”Right to left” of workflow since we’re usually focused on the end result.
- Next stage:
  - Partitioning
  - cross-validation (e.g., x-partition, x-aggregator)
  - Bootstrap sample or Out Of Bag (OOB)
- Then we build the model
- Finally, score the results
- Clustering is thrown into all of that somehow
- Types of Models:
  - One R (Never use)
  - Bayes
    * Bad for redundant attributes
    * Good when there are many attributes that may be useful
  - k-NN
    * Bad for datasets with a large number of datapoints (big n), since scales based off of the size of the dataset
    * Great for datasets where the classes aren’t clumped together (geometrically). Hard to determine with testing it.
  - Decision trees (Never use)
  - SVMs
    * Bad for problems with a large number of classes
    * Pretty good go to option otherwise (always)
  - Rules (Never use)
  - Random forests
    * No immediate major downsides
    * Therefore, practically always a good option
  - Neural networks
    * Bad if training cost is an important factor (big n)

    * Great when there's a small amount of info in each attribute. e.g., images, audio

- From the left side, use knowledge of dataset to make decisions on the workflow
- From the right side, use knowledge of the problem to make decisions on the workflow

2 Lecture 7

2021-10-26

- Looking at basketballwinloss data
- RF out of bag sampling
  - Looking at prediction confidences can help determine why model might be misbehaving.
  - Attribute statistics help see what the RF is determining to be a significant attribute.
- Tree Ensemble learner is a generalization of RF (not the same thing, doesn't select attributes).
  - Options: Tree Options & Ensemble configuration.
  - Out of bag predictions: model count should avg about a third
- A method for determining the most important attributes involves using a math function & sorting the important (splits times candidates)
  - Iteratively remove unimportant attributes to try and increase prediction accuracy.

3 Lecture 8

2021-11-3

- Used Z-scoring method of Normalizer node to improve performance of Tree Ensemble Learner.
- Explaining why clustering is useful.
- Using k-Means clustering in KNIME, analyzing how wins/losses are clustered (ain't prefect).
- Certain parts of Scatter matrix help analyze effectiveness of clusters (include win/loss and cluster columns).
- Scatter plot with high jitter in appearance also helps

4 Lecture 9

2021-11-10

- Using provided Matlab script on data to make visual graphs
-