NO.: Date:

James 62m

TAE Wk 9 – Bagging & Random Forest

9.1 Bootstrapping

A resampling method used to quantify the uncertainty associated with a statistical learning method.
(Other resampling method: cross-validation)

Goal: establish empirical distribution functions for a wide spread range of scenarios.

- How
  - We obtain distinct datasets by repeatedly sampling observations from the original dataset with replacement
  - Each bootstrap dataset is created by sampling with replacement, it is the same size as our original dataset.

- Terminology
  - Z^{*1} : 1st bootstrap data set
  - $\hat{\alpha}^{*1}$: bootstrap estimate for α
  - B : No. of different bootstrap datasets
  - Standard error of bootstrap estimates, $SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*})^2}$

- Where it can be applied
  It can be used to tackle the bias-variance trade-off of some statistical learning methods (such as Decision tree)

8.2 Bagging (Bootstrap aggregation)

Intuition: Decision Tree suffer from high variance. (Say if we split the training data into 2 parts our random, & fit a decision tree to both halves, the results would be quite diff)

Method:

- Take repeated samples from the training dataset (Generate B diff bootstrapped training datasets)
- Train model $f^{*b}(x)$ on the b-th bootstrapped training dataset (repeat for all B datasets)
- We average all the predictions as follows:

$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x)$$

(take majority vote for classification tree)

Entire process is called Bagging

NO.: Date:

8.3 Out-Of-Bag (OOB) Error Estimation

We use OOB Error Estimation to test the error of the bagged model.

- To obtain a single prediction for the i^{th} observation, we average predicted responses (for regression) or, majority vote (for classification)

  ∴ this leads to a single OOB prediction for the i^{th} observation.

- An OOB prediction can be obtained in this way for each of the n observations, from which we compute the overall OOB MSE / classification error

- When B is sufficiently large, we have a decent alternative to cross-validation

8.4 Random Forest

Random forest builds from the idea of considering a random sample of m predictors from the full set of predictors P

→ this is to overcome the problem in DT bagging where trees tend to be correlated.

Idea: (total, subset)

- for each split, we consider only a subset of predictors (P c m)

↳ ∴ on average $(p-m)$ predictors are not considered, so other predictors will have a chance.
(This process decorrelates a tree)

Algorithm:

1. Generate B bootstrapped training datasets.
2. For each dataset, form a decision tree (At each split, use a subset m of p available pred
3. Average the predictions from the B trees
   (For classification, use majority voting)

value of m:

- Regression, $m = \frac{p}{3}$
- Classification, $m = \sqrt{p}$

Hyperparameters Tuning.

- NO. of trees, B
- NO. of predictors used at each split, m

NO.:

Jones 62

Date:

TAE WK 9 - Ethics of data analytics

Ethical issues in data science

① Bias, discrimination & exclusion
Algos & AI create exclusion towards individuals & groups of people.

② Algorithmic Profiling
The misuse of personality can affect the key collective principles like democratic & cultural pluralism & risk sharing in the realm of insurance

③ Protecting massive files while enhancing A.I.
Data protection laws are rooted in the belief that individuals' rights regarding their personal data must be protected.

④ Quality, quantity & relevance
The acceptance of the existence of potential bias in datasets curated to train algo is of paramount importance

How data visualisation is abused

1) Omitting the baseline.
(Not starting at the origin).

2) Manipulating the y-axis
(changing the scale of y-axis to exaggerate or conceal trends)

3) Cherry picking Data.

4) Using the wrong graph
(wrong type of visualisation).

NO.: Date:

5) Improper intervals or units

6) Omitting data

7) Extrapolation

8) Unnecessary complexity.