Special Topics in Decision Tree

CS7DS1

Bahman Honari

Overfitting – A Bugbear

What overfitting is.

- Overfitting occurs when a precise model is built more on the random noise presence in the training dataset, instead of the underlying relationship between the independent(s) and dependent variables.
- The problem is we do not really know what is noise and what is underlying relationship!!

Why overfitting occurs.

- The training set that is used to build the model is just a sample. The model, however, is expected to work for the entire dataset.

The Elephant in Dark, by Rumi

Some Hindus have an elephant to show, No one here has ever seen an elephant. They bring it at night to a dark room.

One by one, we go in the dark and come out, saying how we experience the animal.

One of us happens to touch the trunk, A water-pipe kind of creature.

Another, the ear. A very strong, always moving, back and forth, fan-animal.

Another, the leg. I find it still, like a column on a temple.

Another touches the curved back, A leathery throne.

Another the cleverest, feels the tusk. A rounded sword made of porcelain. He is proud of his description.

Each of us touches one place, and understands the whole that way.

The palm and the fingers feeling in the dark, are how the senses explore the reality of the elephant.

If each of us held a candle there, and if we went in together, we could see it.

Overfitting – Cont.

- A problem of avoiding overfitting might be to end up with a model of the opposite problem: underfitting!!
- Bias-Variance Trade off

- Don't forget that:

A simpler model developed using a larger dataset, would perform better that a complex model developed using a smaller dataset.

Bias-Variance Trade off

- An overfitted model would generate different responses for different datasets. (Variance)
  - High variance indicates that the algorithm produces a model that is very specific to the training dataset.
- An underfitted model would generate similar responses for difference datasets. (Bias)
  - High bias indicates that the model does not understand some important pattern in different datasets.

Overfitting Potential Solutions

- **Cross Validation**
- **Regularisation**
- **Ensemble Learning**

Cross Validation

- Cross validation is a technique for model selection.
- A proper model is the one that performs well on a dataset that was not seen before. In other words, if a model performs well on the training dataset, this is not a guarantee to be a good model.
- It would therefore be an appropriate approach to use multiple training datasets to end up to the model that works well with various datasets.

Cross Validation – Cont.

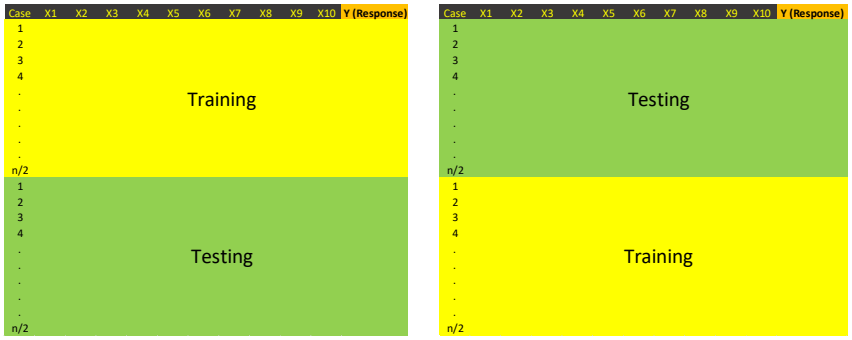
We therefore:

- Crate multiple training datasets, which all are a subset of the larger dataset.
- Keep some data aside to check the performance of the model.

These two steps together is what is called Cross Validation.

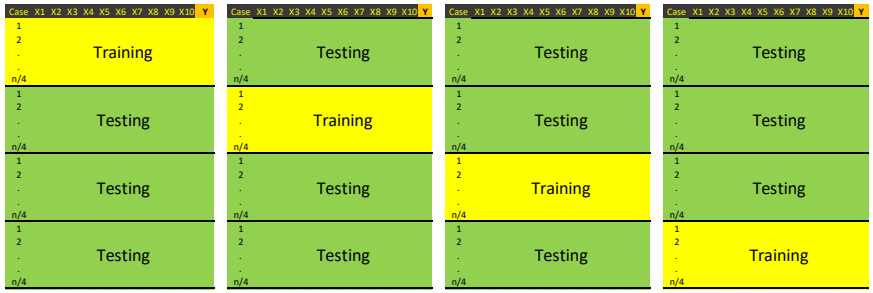
Cross Validation – Cont.

2-fold cross validation



Cross Validation – Cont.

4-fold cross validation



Cross Validation – Cont.

When to use Cross validation:

- To choose between algorithms.
- To choose the algorithm parameters.
- To choose the best set of IVs.

Cross Validation – Cont.

k-fold cross validation:

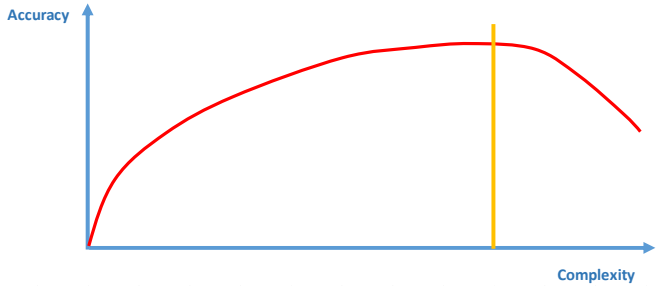
What is usual k?

- If you have a large dataset, then k=10 is usual.

- Leave 1 out, leave p out, Monte-Carlo are other variants of k-fold CV.

Regularisation

- What is Regularisation?
  - To penalise the models that are too complex (number of branches in DT or highest degree of a polynomial in regression analysis)



Regularisation – Cont.

- What is Regularisation?
  - To penalise the models that are too complex (number of branches in DT or highest degree of a polynomial in regression analysis)

$$E(New\ Model) = E(Model) + \alpha\, Reg(Complexity)$$

  - Adjusted R-square
  - Lasso regression

Ensemble Learning

What is Ensemble Learning?

Simply, it is using various models and combining their results.

Why does it work?

Because the overfitting component of several models cancel eachother when combined.

Ensemble Learning – Cont.

Ensemble learning could be:

- Using different techniques
- Using different training datasets
- Using different IVs
- Using different model parameters

Ensemble Learning – Cont.

The result would be:

- The most common outcome of different models (discrete outcome);
- The average of the outcomes of different models (continuous outcome).

We can also use a “Weighted” Mode or Mean of results.

Bagging, Boosting, Stacking

- Bagging (Bootstrap-Aggregating)
  - Bootstrap Sampling = Uniform Sampling with Replacement
- Boosting
  - Giving higher weight to misclassified cases.
  - Adaboost (Adaptive Boosting)
- Stacking
  - Machine-Learning approach to combine the outcome of the several models.

Random Forest

What Random Forest is.

- Simply Ensemble of Decision Trees.
  - The decision trees are developed based on both
    - Different Training sets (Bagging)
    - Different IVs (Random Subspace Method)
  - Allows us to understand which IVs are more important.