Decision Tree notes

- A simpler model built from a large dataset will perform better than a complex model built with a small dataset.

Overfitting

- Model is built to exactly match (perfectly predict) the dataset being used for training. Model reflects the noise in the dataset more than it does the relationship between independent and dependent variables.
- Would generate different responses for similar datasets.

Underfitting

- Model lacking sufficient training that it doesn't understand anything about the dataset and makes random, incoherent, predictions.
- Would generate different responses for similar datasets
- A model that fails to understand the underlying trends in the data

Overfitting solutions

- Cross Validation
- Regularisation
- Ensemble Learning

Cross validation

Splitting the data into **k** subsets.

- Model selection technique
- A good model is one which performs well on an unseen dataset i.e Something other than the training set

Workflow:

- Split the data into multiple subsets
- Keep some data aside to validate the performance of the model - Validation

Example(4-fold cross validation)

1. Split the data into 4 subsets
2. Take one of the subsets as the testing dataset. The other 3 subsets are the training datasets
3. Train your model and validate against the testing subset
4. repeat steps one to three using a different subset as the testing set each time

Regularisation

Penalises models that are too complex e.g high number of branches in decision tree or highest degree of a polynomial in regression analysis

E(New Model) = E(Model) + αReg(Complexity)

A form of regression that shrinks the coefficient estimates towards zero. For a refresher on regression just go down below

Some common regularization algorithms are *adjusted R-square* and *Lasso regression*

Adjusted R-square Since R-square is always positive and always increases as the number of terms in a model increases, it can be misleading as to the correctness of your model.

Adjusted R-square Increases only if the new term improves the model more than would be expected by chance.

Decreases when a predictor improves the model than less than would be expected by chance

Can be negative

Lasso Regression A shrinkage model that attempts to shrink the data points towards a central point, like the mean.

Meant to encourage simple, sparse models.

Ensemble Learning

Real simple, take various different models, train them on the data, and combine the results

Works because the overfitting component of various models cancel each other when combined

Various kinds of Ensemble learning could be

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

IV: fairly sure this is referring to *individual variables*. This is something like the α value used in regularisation to penalise overly complex models.

So what's the result of Ensemble learning?

- the most common outcome between all the different models (discrete outcome)

- The average of the outcomes of the different models (continuous outcome)

(brief) Regression refresher

Technique used to describe a relationship between a dependent variable and one or more independent variables

Before going into regression I'll first define the correlation coefficient r

Correlation Coefficient (r): ranging from -1 to 1 it measures the correlation of data from strong correlation (1, -1) to weak correlation (0)

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

where $x = x_i - \bar{x}$ and $y = y_i - \bar{y}$

Linear Regression: relation of an dependent variable to a single independent variable

$$\hat{y} = b_0 + b_1 x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable and $\hat{y}$ is the predicted value of the dependent variable

to find the the constant b_0 and the regression coefficient b_1 you use the following formulae

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},\ b_0 = \bar{y} - b_1 \bar{x}$$

I should mention that this is called the least squares method

- Minimizes the squared difference between values

I should also mention that your x and y values will be coming from your training data

There are many forms of regression but this should give an idea of how regression functions

Additional Notes

R-Squared

$R^2 = 1 - \frac{ExplainedVariation}{TotalVariation}$

A measure of how close the data is to the fitted regression line

Whereas correlation explains the strength of the relationship between an independent and dependent variable. R-Square explains to what extent variance in one variable explains variance in the second variable.

For example is the R^2 value of a model is 0.5 than approximately half of the observed variation can be explained by the model's inputs.

The higher the R-Square value the better the model fit (in general but this is not always the case)

Cannot assess whether the coefficient estimates and the predictions are biased –> Look at residual plots for this.

Residual plots

Can be used to access whether the observed error is consistent with the stochastic error.

stochastic is just a fancy way of saying random.

there are two parts to your model: deterministic and stochastic. Deterministic being the predictive part and the stochastic part being the rest. The radom error.

Looking at a residual plot the values should be symmetrically and randomly distributed with a mean of zero. There should be no predictive value in the plot. e.g I should not be able to say with confidence that x = 5 will produce a positive value.

If you get a non-random residual plot it is an indication that the deterministic part of your model is not capturing som explanatory information that is "leaking into the residuals" e.g a missing variable, higher-order term etc.