Machine Learning

Kiar Fatah

February 23, 2020

Contents

> If you can’t explain something in
> simple terms, you don’t understand it.
>
> *Richard Feynman*

1 Lecture 2

1.1 Decision Tree

A decision tree is a map of the possible outcomes of a series of related choices.

1.2 How to build a Tree

1. Choose the best question (according to the information gain), and split the input data into subsets.
2. Terminate: call branches with a unique class labels leaves (no need for further questions).
3. Grow: recursively extend other branches (with subsets bearing mixtures of labels).

1.3 How to ask the best questions?

The machine will automate the questions for the decision tree, however how does it obtain the best questions to ask? It does so by with the help of Gini impurity and information gain.

1.3.1 Gini impurity

Gini impurity is a measurement of uncertainty in a node. It is calculated by how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$1 - \sum_i p_i^2. \tag{1.1}$$

1.3.2 Entropy

Entropy is also a measurement of uncertainty in a node and can be applied instead of Gini impurity. It is calculated as

$$\sum_i -p_i log_2 p_i. \tag{1.2}$$

The information entropy is measured in bits, corresponding to the logarithmic base 2. Tossing a coin results in the entropy 1, hence it has 1 bit of information. For a real dice and a fake dice the value is 2.58 and 2.16 bits of information respectively.

1.3.3 Information gain

Information gain is how much the question reduces the uncertainty. The information is given by the first set of values for the internal node subtracted by the average value of the impurity of the split set.

1.3.4 Best question

The question with the highest value on the information gain will be the one to ask.

1.4 Overfitting

Overfitting is when the learned models are overly specialised for the training samples. This occurs when the data is too noisy, not representative and when the model is too complex. This can be tackled by choosing a simpler model.

1.4.1 Pruning

The idea of reduced error pruning is to consider each node in the tree as a candidate for removal. A node is removed if the resulting pruned tree performs at least as well as the original tree over a separate validation dataset. The pruning is done on validation set, until it is harmful for the model.

1.4.2 Validation set

Due to pruning there is a need for validation dataset that is not training nor test data set. Therefore if there is access to a rich data set, it is separated to three categories. In essence the validation set is used to estimate prediction error for model selection (i.e. to determine hyperparameters).

2 Lecture 3

2.1 Introduction

How does one determine the right model, f, from the data? The simple concept for classification is by calculating the misclassification rate in consideration to the data. Assume the following data is given $\mathcal{D} = (\vec{x_1}, y), ...$ is given. The misclassification rate is given by

$$err(f, \mathcal{D}) = \frac{1}{N} \sum Ind(f(\vec{x}) \neq y) \tag{2.1}$$

2.2 Curse of Dimensionality

1. Easy problems in low-dimensions are harder in high-dimensions, e.g training more complex models with limited sample data.
2. In high-dimensions everything is far from everything else, this is an issue in the nearest neighbour model, this causes issues with Nearest Neighbours.
3. Any method that attempts to produce locally varying functions in small isotropic neighbourhoods will run into problems in high dimensions.

2.3 The Bias-Variance Trade-off

Let us imagine we could repeat the modelling for many times – each time by gathering new set of training samples, D. The resulting models will have a range of predictions due to randomness in the underlying data set.

1. Variance: Refers to the amount by which $\hat{f}$, the prediction function, would change if we estimated it using a different training data set.
2. Bias: Refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

This can be showcased in figure 1 In essence the relationship between complexity, bias and variance can be summarised as

1. Low model complexity implies high bias, low variance.
2. High model complexity implies low bias, high variance.

Note that bias and variance are equally important as we are always dealing with a single realisation of the data set.

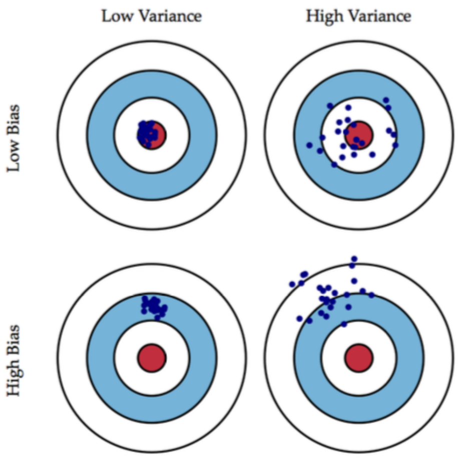

Figure 1: Variance versus Bias

3 Lecture 4

3.1 Linear Regression, A Parametric Method

Simple linear regression assumes theres a linear relationship between the output and input. Therefore a linear equation is assumed, parametric method. However, the coefficients are unknown and thus needs to be approximated.

The measurement of error is mean square. It is a convex function. Therefore the coefficients that minimises the sum of the squared error is given by the vector that sets the gradient of the sum of squared error to zero.

$$E_{in}(w) = ||Xw - Y||^2 \implies \frac{\partial E_{in}}{\partial w} = 2X^T(Xw - Y) = 0 \implies w = (X^TX)^{-1}X^TY \tag{3.1}$$

whereas w is the coefficients, X the data, Y the label and E_{in} the sum of squared errors.

3.2 RANdom SAmpling Consensus

Random sampling consensus or RANSAC is applied when the data set S contains outliers

1. Randomly select a (minimum number of) sample of s data points from S and instantiate the model from this subset.
2. Determine the set of data points S_i which are within a distance threshold t of the model. The set S_i is the consensus set of samples and defines the inliers of S.

3. If the subset of S_i is greater than some threshold T, re-estimate the model using all the points in Si and terminate
4. If the size of S_i is less than T, select a new subset and repeat the above.
5. After N trials the largest consensus set S_i is selected, and the model is re-estimated using all the points in the subset S_i

In essence RANSAC

3.3 Disadvantages with RANSAC

3.4 Differences with Parametric and Non-parametric Methods

1. If the parametric form is close to the true form of f, the parametric approach will outperform the non-parametric
2. As a general rule, parametric methods will tend to outperform non-parametric when there is a small number of observations per predictor (i.e. in a high dimension).
3. Interpretability stand point: Linear regression preferred to KNN if the test MSEs are similar or slightly lower.

3.5 k-NN Regression, A Non-parametric

The k-NN regression method is closely related to the k-NN classifier. Given a value for K and a prediction point x_0, k-NN regression first identiffies the K training observations that are closest to x_0 represented by N_0. It then estimates $f(x_0)$ usin the average of all the training responses in N_0. In other words

$$f(\hat{x}_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i. \tag{3.2}$$

Note larger values of k provide a smoother and less variable fit (lower variance!)

3.6 Parametric or Non-parametric Methods?

1. If the parametric form is close to the true form of f, the parametric approach will outperform the non-parametric
2. As a general rule, parametric methods will tend to outperform non-parametric when there is a small number of observations per predictor (i.e. in a high dimension).

3. Interpretability stand point: Linear regression preferred to KNN if the test MSEs are similar or slightly lower.

3.7 Shrinkage Methods

A common way to fit models is by appling least square, however an alternative is it is possible to fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently that shrinks the coefficient estimates towards zero.

1. Among a large number of variables X in the model there are generally many that have little (or no) effect on Y
2. Leaving these variables in the model makes it harder to see the big picture, i.e. the effect of the "important variables"
3. Would be easier to interpret the model by removing unimportant variables (setting the coefficients to zero)

3.8 Ride Regression

Similar to least squares but minimizes different quantity

1. Shrinkage penalty
2. The parameter λ
3. Interpation of the graphs

3.9 The Lasso

Similar to ridge regression but with slightly different term