# Machine Learning

Kiar Fatah

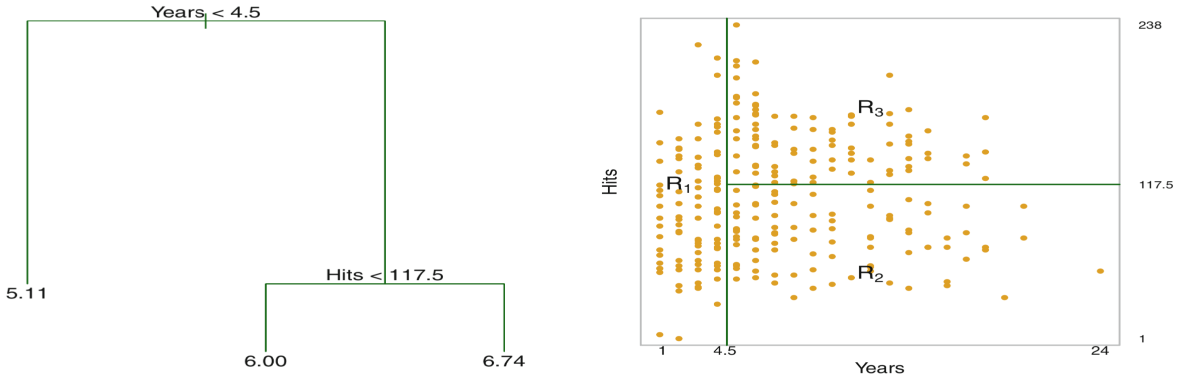
February 22, 2020

# Contents

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

# 1 Lecture 2

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



(a) A visual map of decision tree.

(b) The numerical effect of a decision tree.

Figure 1

## 1.1 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.

### 1.1.1 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.

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

### 1.1.2 Entropy

Entropy is also a measurement of uncertainty in a node. It is calculated as

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

### 1.1.3 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.

### 1.1.4 Best question

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

### 1.1.5 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.

### 1.1.6 Validation set

Due to pruning there is an 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.

# 2 Lecture 3

## 2.1 What is Statistical Learning

We assume that there is some relationship between the output $Y$ and the input $X = X_1, X_2, ...$, that can be written in the general form

$$Y = f(X) + \epsilon. \tag{2.1}$$

here f is some fixed but unknown function of the input and $\epsilon$ is a random error term, which is independent of the input and has a mean zero. Hence in essence, statistical learning refers to a set of approaches for estimating f.

## 2.2 Why Estimate f?

Two reasons to estimate the unknown function f

1. Prediction
2. Interference

### 2.2.1 Prediction

In common situations a set of inputs are available, however the output can not be easily obtained. In this setting, since the error term averages to zero, the output can be predicted as

$$\hat{Y} = \hat{f}(X). \tag{2.2}$$

Here $\hat{f}(X)$ represents our estimate for the function f, and $\hat{Y}$ the resulting prediction for the output. Therefore $\hat{f}(X)$ is treated as a black box, meaning that one is not concerned with the exact form of $\hat{f}$, provided it yields the accurate predictions for the output, Y.

In this case reducible and irreducible error are introduces. Reducible error means $\hat{f}$ is not a perfect estimate for the function f. This will introduce error, however this error is reducible and can be improved. Irreducible error cannot be improved and is introduced by $\epsilon$.

### 2.2.2 Inference

Often one is interested in knowing how the output is affected as the input variables change. The goal is to estimate f but not necessarily to predict Y. Hence $\hat{f}$ can not be treated as a black box. because it is needed to know the exact form. In this case the following questions needs answers

1. Which predictors are associated with the response?
2. What is the relationship between the response and each predictor?
3. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

## 2.3 How Do We Estimate f?

We wanna find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X,Y). Most statistical learning methods for this task can be characterized as either parametric or non-parametric.

### 2.3.1 Parametric Methods

In parametric methods there is a two-step model-based approach

1. Make an assumption about the functional form, or shapre, of f
2. After a model has been selected, there is a need for a procedure that uses the training data to fit or train the model.

In other words it reduces the problem of estimating f down to one of estimating a set of parameters. The potential disadvantage of a parametric approach is that the model that is chosen will often not match the true unknown form of f. This can be improved by choosing flexible models that can fit many different possible functional forms for f, however it requires a greater number of parameters. Greater number of parameters leads to more complex models that can result in overfitting the data, which means the approximation follows the errors, or noise, too closely.

### 2.3.2 Non-parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f . Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

## 2.4 Assessing Model Accuracy

In statistical learning there is not one model that dominates over all other models.

## 2.5 Measuring the Quality of Fit

The way to evaluate the performance of a statistical learning methods on a given data set, is through mean squared error (MSE) given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.3}$$

The MSE can be calculated for the training and the test set, however the interest lies in the accuracy of the model on unseen data, test set. Therefore the method that gives the lowest test MSE is the one to choose.

Note that there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data. This is due to the model is trying to hard to find patterns in the data but picks up patterns that are caused by chance rather than by true properties.

## 2.6 Curse of Dimensionality

1. Easy problems in low-dimensions are harder in high-dimensions
2. In high-dimensions everything is far from everything else, this is an issue in the nearest neighbour model
3. Any method that attempts to produce locally varying functions in small isotropic neighbourhoods will run into problems in high dimensions.

## 2.7 The Bias-Variance Trade Off

Let us imagine we could repeat the modeling 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 $\hat{f}$ 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 2 In essence the relationship betewen complexity, bias and variance can be summarized as

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

A low model complexity can be a decision tree with a low level depth and a high model complexity can be a decision tree with a high level depth.

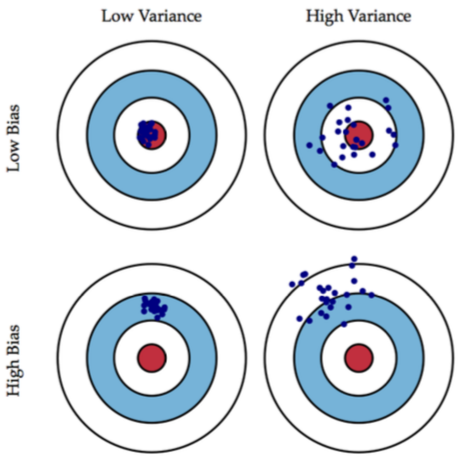


Figure 2: Variance versus Bias

### 2.8 The Validation Set Approach

The resulting validation set error rate provides an estimate of the test error rate.

## 3 Lecture 4

### 3.1 Linear Regression, A Parametric Method

Simple linear regression assumes theres a linear relationship between the output and input. Hence in the as mention earlier in the section parametric methods, a linear equation is guessed. However, in practice the coefficients are unknown and therefore needs to be approximated. The most common approach is done by least squared method.

### 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 $S_i$

## 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

$$f(\hat{x_0}) = \frac{1}{K} \sum_{x_i \in N_0} y_i. \tag{3.1}$$

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 $\lambda$
3. Interpation of the graphs

### 3.9 The Lasso

Similar to ridge regression but with slightly different term

## 4 Lecture 5

### 4.1 Probability Theory in Machine Learning

The advantages are

1. Interpretability: It can be more transparent and mathematically rigorous than other machine learning methods.
2. Transparency: Assumptions can be made more explicit than in other methods.

3. Efficiency: It can work in regimes that are data poor.
4. Flexibility: Easy to merge different parts of a complex system and to update current knowledge with new observations.
5. Encompassing: Aspects of learning and inference can be cast under the same theory.

The disadvantages are

1. Complicated: Often hard to derive closed solutions. Need to resort to computation and heuristic approximations.
2. Scalability: Not computationally scalable to large dataset.

## 4.2 Axiomatic definition of probabilities

Given an sample space $\omega$ of all possible outcomes, and an event E, or set of outcomes from $\omega$ then

1. $P(E) \geq 0$ for all $E \subseteq \omega$.
2. If $E = \omega$ then the probability of $\omega$ is one. In other words the probability that one event will occur is equal to one.
3. if the events are a ocuntable sequence of pairwise disjoint events then the probability of the union of each event is equal to the probability of the summation of every event.

## 4.3 Random (Stochastic) Variables

A random variale is neither random nor a variable, it is a function that does exactly what we need, $X : \omega \to \mathcal{R}$. The probability distribution function, so called pdf, of a random variable maps the range to positive numbers, $Pr(x) : X \to \mathbb{R}_{>0}$.

1. The probability of a random variables describes how the probability density is distributed over the range of X,

$$P[0 \leq X \leq 1] = \int_{x=0}^{1} Pr(x)dx. \tag{4.1}$$

2. X is distrubted Pr(x) is written ore compactly as

$$X\tilde{P}r(x). \tag{4.2}$$

## 4.4 Types of Random Variables

1. Discrete random variable: Countable set
2. Continuous random variable: Uncountable set

## 4.5 Joint Probabilities

Consider two random variables X and Y. Observe multiple paired instances, some paried outcomes will occur frequently. This information is encoded in the joint probability density function $Pr(x, y)$.

## 4.6 Marginalization

The probability distribution function, PDF, of any single variable can be recovered from a joint distribution by summin for the discrete case and integrating for the continuous case.

$$Pr(x) = \sum_y Pr(x, y), (Discrete) \tag{4.3}$$

## 4.7 Conditional Probabilities

The conditional probability of X given that Y takes a value y gives

$$P(A|B) = \frac{P(A, B)}{P(B)}. \tag{4.4}$$

If it turns out the events are independent then

$$P(A|B) = \frac{P(A)P(B)}{P(B)}. \tag{4.5}$$

Bayes's theorem gives us

$$P(A|B) = P(A|B)P(B) = P(B|A)P(A) \implies P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.6}$$

## 4.8 Common Distributions

1. Bernoulli: Binary variables
2. Categorical: Discrete variables
3. Gaussian: Univariate normal distrubtion

## 4.9 Central Limit Theorem

The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.1

## 4.10 Expectation

The expected value can be calculated according

$$\mathbb{E}[X] = \mu_X = \int xPr(x)dx \tag{4.7}$$

it is equal to the center of gravity of a distribution.

The variance is

$$Var[X] = \sigma_X^2 = \mathbb{E}[X - \mathbb{E}[X])^2] \tag{4.8}$$

It is interpreted as the spread of a distribution.

Covariance is given by

$$\sigma_{X,Y} = \mathbb{E}[X - \mathbb{E}[X]](Y - \mathbb{E}[Y])] \tag{4.9}$$

Shows how two variables vary together.

## 4.11 General Machine Learning Problem

if Y is discrete then it is classification, if it is continuous it is regression.

1. Learning: We estimate Pr($\boldsymbol{x}$, y) from data.
2. Inference: We estimate Pr(y— $\boldsymbol{X} = \boldsymbol{x}$) from data.

## 4.12 Bayes's Rule

1. $Pr(\boldsymbol{x}|Y = y)$: Likelihood represents the probability density of observing data $\boldsymbol{x}$ given the hypothesis Y = y
2. $Pr(Y = y)$: Prior repreents the knowledge about Y before any observation.
3. $Pr(y|\boldsymbol{X} = \boldsymbol{x})$: Posterior represents the probabiltiy density of hypothesis y given observation $\boldsymbol{X} = \boldsymbol{x}$.
4. $Pr(\boldsymbol{X} = \boldsymbol{x})$: Evidence describes how well the model fits the evidence.