**Chapter 2: Overview of Statistical Learning**

2.1 Introduction to regression models

* The standard model for regression is , where is the ***reducible* error** (the error in prediction can be potentially reduced by choosing a different model) and is the ***irreducible error*** (it originates from e.g. measurement errors or inherent stochasticity and cannot be reduced by changing the model).
* A good will allow us to make predictions of the value of at new points of .
* The ideal function will give the expectation value of at :

This ideal function is called the ***regression function***. This ideal function is optimal is the optimal predictor of in the sense that it will minimize the ***sum of squared errors***.

* Typically, there will be very few, if any, data points that correspond to a certain value of (for example, there may be no values for the output data at exactly ). In that case the expectation value cannot be determined.
* In this case, is estimated based on the points in the ***neighbourhood*** of :

Where is some neighbourhood around . This is called ***Nearest Neighbour or Local averaging***.

2.2 Dimensionality and Structured Models

* Nearest neighbour averaging becomes more problematic as the number of dimensions increases: For a high number of dimensions, the width of the neighbourhood for nearest neighbour averaging needs to increase to capture the same number of data points.
* As a result, nearest neighbour averaging becomes less local for as the number of dimensions increases. This is called the ***Curse of Dimensionality***.
* The curse of dimensionality can be circumvented by using ***structural models*** that do not depend local properties and nearest neighbour averaging. An example of a structural model is a linear model for which each of the parameters can be found by fitting it to all datapoints. As a result, these linear models do not depend on any local properties or nearest neighbour averaging.
* Structural models that are less ***flexible*** (for example, linear vs. quadratic models) typically have lower ***interpretability***, although they may provide a better fit to the data.
* It is important to choose a model that is not too flexible to prevent ***overfitting*** of the data.

2.3 Model Selection and Bias-Variance Tradeoff

* To see how well a model performs, it can be fitted to a set of ***training data***:

However, this can be ***biased towards more overfit models***. Instead, the value should be evaluated on a set of ***test data***:

* The performance of models with different flexibilities can be assessed by plotting against model flexibility. The value of will typically have a minimum value for a certain flexibility of the model, while will keep decreasing as the model becomes more flexible.
* The choice for a certain flexibility of the model is subject to a ***Bias-Variance Tradeoff***: the variance in predictions of a model will increase as the model becomes more flexible (the model will become more sensitive to the peculiar characteristics of the training dataset). In return, the bias will decrease as the model becomes more flexible.

2.4 Classification

* In classification, the response variable is ***qualitative*** and the aim is to build a classifier that can assign a class label from the set of labels for future observations of .
* An ideal classifier is the ***Bayes Optimal Classifier*** that minimizes the probability of misclassification.
* If there are elements in the set of class labels then the ***Conditional Class Probabilities*** are:

The Bayes Optimal Classifier at is given by:

In other words, the Bayes Optimal Classifier assigns the observation to the ***most class with the highest probability***.

* In classification, nearest neighbour averaging can be used as in logistic regression (take the conditional probabilities in the neighbourhood of ). In that case, the curse of dimensionality still applies. However, the curse of dimensionality has less impact on than on .
* The performance of the classifier is typically measured using the ***Misclassification Rate***:

In words: ***the error is the average number of mistakes***. This error is the smallest for the Bayes Classifier.

* ***Support Vector Machines*** build structured models for **.**

**Chapter 3: Linear Regression**

3.1 Simple Linear Regression

* Linear regression is a simple approach to ***Supervised Learning*** (it is based on the assumption that the dependence of on is linear).
* Simplest case of linear regression is a model with a single predictor:

When the model coefficients have been estimated, the output can be predicted by:

* The best values for the parameters can be estimated by ***least squares***.
* The r***esidual*** (which is the discrepancy of the actual outcome with the predicted outcome) for the prediction of based on the value of is given by:

And the ***Residual Sum of Squares*** is:

* The unique line with values for and that minimizes the RSS is the ***least squares fit***.
* The accuracy of the estimates for and can be determined using the ***Standard Error***. The standard Error will ***increase when the variance () increases*** and the Standard Error will ***decrease when the spread of x-values around their mean increases***.
* The standard Errors can be used to define ***Confidence Intervals*** (For example, an interval that has a 95% probability to contain the true slope value).

3.2 Hypothesis testing and Confidence Intervals

* The standard error can also be used for performing ***Hypothesis Testing***. The most common Hypothesis test involves testing the Null Hypothesis:

***Null Hypothesis*** (): There is no relationship between and .

***Alternative Hypothesis*** (): There is some relationship between and .

* The Null Hypothesis is tested by computing a ***t-statistic:***

If the Null Hypothesis is true, this will ***have a t-distribution with degrees of freedom***.

* The ***p-value*** is the probability to get ***at least*** the obtained value for .
* ***Hypothesis Testing and using Confidence Tests are equivalent***: if one of the two shows that there is a relation between and then so will the other and vice versa. However, the Confidence Test also tells you how big the effect is.
* The accuracy of the model can be determined by computing the ***Residual Squared Error***:

And by computing the  ***value*** (the describes the fraction of the variance in the data that is explained by the model). , where is the correlation between and .

3.3 Multiple Linear Regression

* Multiple Linear Regression is regression with more than one predictor:
* How to interpret the different predictors now:

-***If the predictors are not correlated data***, each predictor can be interpreted separately.

-***Usually, the different predictors are correlated,***and this makes their interpretation much more complex. In addition, ***the variance on all coefficients can increase*** when the predictors are correlated. This is because the relation of a predictor to a certain variable is less strong in that case (the predictors of different variables can be swapped without significantly changing the result). In this case ***Claims od Causality*** should be avoided.

* The ***Least Squares Estimates*** can be determined in much the same way as for a simple linear model ().
* Instead of a line, the function will now be a ***Hyperplane***. Therefore, in this case the Least Squares Estimates minimize the distance between each data point and its closest point on this hyperplane.
* For multiple Linear Regression, the standard error, t-statistic and p-value will tell you whether one variable affects the outcome ***in the presence of the other variables***. Although a variable may not have a significant affect in the presence of other variables, it could have a significant effect on its own.
* ***The correlation between the variables*** will tell you something about whether the information on a certain variable significantly improves the prediction of the model given that you have information on the other variables. In other words, data on a variable may have become redundant in the presence of another variable when they are strongly correlated.

3.4 Some Important Questions

* Which of the predictors is useful for the prediction of the response? And how accurate is the prediction?
* To determine if any of the predictors is useful to predict the response, the ***-statistic*** is used:

Here, the  ***is the Total Sum of Squares***, which is the sum of the residuals when no model (the predictor of the data is the mean) is used. is the Residual Sum of Squares, which are the residuals that remain when using our predictive model. is the number of parameters in the model (in total parameters are fitted to the data, the accounts for the intercept) and is the sample size.

* In the case that there is no effect of the predictors (the null hypothesis), the ***-statistic will follow an -distribution with degrees of freedom ()*.**
* The values for the -distribution can be looked up in tables. ***A large -statistic indicates that there is a strong effect of the predictors on the response prediction.***
* To decide whether a variable is important for the prediction of a linear regression model, ***All subsets or best subsets regression*** can be used.
* In all subsets regression, the least squares fits for all possible subsets of variables are made and the best subset is chosen based on a balance between training error and model size.
* The problem with all subsets regression is that the number of models that need to be evaluated grows exponentially with the number of parameters in the model.
* Two commonly used approaches for all subsets regression:

***Forward Selection***

Starts with the ***null model*** (a model with only the intercept and no predictors). Then each of the variables is added one at a time to the null model and simple linear regressions are performed. ***The variable that results in the lowest is added to the null model***. This is repeated for two-variable models, three variable models etc., until some stopping rule is reached (such as ***all the remaining variables have a p-value above some threshold***).

***Backward Selection***

Start with a model that contains ***all the variables***. Then remove the variable that has the largest p-value (the one that is the least statistically significant) from the model. Repeat this process for a model with variables, variables etc., until some stopping rule is reached (such as ***all the remaining variables have a p-value below some threshold***).

* How to deal with ***Qualitative Variables*** instead of quantitative variables: incorporate a ***dummy variable*** in the model (there will always be one fewer dummy variable than the number of levels in the set of qualitative variables).

Such that the equation of the model becomes:

In this case ­will tell you the effect of belonging to class A versus the ***baseline***, which in this case corresponds to belonging to class B. The value of the intercept () represents the average value for for class B. The value of will represent the strength of belonging to class A over belonging to class B and the p-value of will tell you if there is any significant difference in the expected value of depending on whether something belongs to class A vs class B.

3.5 Extensions of the Linear Model

* Make models that include ***Interactions*** and ***Nonlinearity*** instead of relying on the assumption that the effect of the different parameters are additive.
* To account for interactions, you include ***Product terms*** between the different parameters:

Instead of an independent coefficient for the variable , the coefficient for is now modified to be ***dependent on the second variable*** . Now the value and p-value of the coefficient  ***gives you information about the strength and significance of the interaction***.

* Sometimes the interaction terms between variables are very significant (low p-value), but their main effects do not have strong significance (that is, there is a significant contribution of the interaction between and on the response variable , but the contribution of either or alone does not). In that case, it is custom to follow the ***Hierarchy Principle***:

*If we include an interaction in a model we should also include the main effects, even if the p-values associated with their coefficients are not significant.*

The reason that this is done is because it is usually difficult to interpret the model when you do not include the main effects. Specifically, ***the*** ***interaction terms also contain main effects if there are no main effect terms added to the model***.

* Models can be modified in the following way to include ***interactions between a*** ***quantitative and a qualitative variable*** in the following way:

In this case where there is ***no interaction*** between the classes and the variables, ***only the intercept*** will be different depending on whether the datapoint belongs to class A or class B.

In this case where ***there is an interaction*** between the classes and the variables, ***both the intercept and the slope*** can be different depending on whether the datapoint belongs to class A or class B.

* The same approach can be used to accommodate ***polynomials*** in the model:

In this case the model is still called a ***linear model*** (and so we still do linear regression), because the model is still linear in the coefficients.