Aarhus University

Decision Support Systems

PROJECT REPORT

Author:

Lasse Lildholdt
(201507170)
Stinus Skovgaard
(201507170)
Daniel Tøttrup
(201507170)
Johan Vasegaard
(201507170)
Frederik Madsen
(201507170)

Supervisor:

Christian Fischer Pedersen

16. marts 2020

Indhold

Figurer

Kapitel 1

INTRODUCTION

In this project report, the reader will be presented with problem solutions for the course Decision support systems. Throughout the solution the reader will achieve knowledge on several different subject within the main area. Each topic will be presented with the theory along with solutions for appropiate exercise to validate the presented theory.

The main topics which will be handled in this report will be as follows:

- Simple linear regression / Multiple linear regression
- Logisitic regression / Linear discriminant analysis
- Cross validation / Bootstrap
- Subset selection
- Shrinkage methods / DImension reduction methods

Kapitel 2

SIMPLE LINEAR REGRESSION / MULTIPLE LINEAR REGRESSION

The simple Linear Regression approach is a quick and simple method for fitting a line through a 2-dimensional dataset. It is assumed that there is a approximately linear relationship between the two dimensions. This can be written mathematically as:

$$Y \approx \beta_0 + \beta_1 * X \tag{2.1}$$

eq. (2.1) can also be seen as "Regressing Y onto X". As an example the dataset Advertising.csv contains sales og a certain product and advertisement money spent on certain media platforms. X represents TV advertising and Y represents sales. It is the possible to regress sales onto TV.

In order to do this, we need to calculate the constants β_0 and β_1 which represents the intercept and slope terms in the linear model.

Kapitel 3

LOGISITIC REGRESSION / LINEAR DISCRIMINANT ANALYSIS

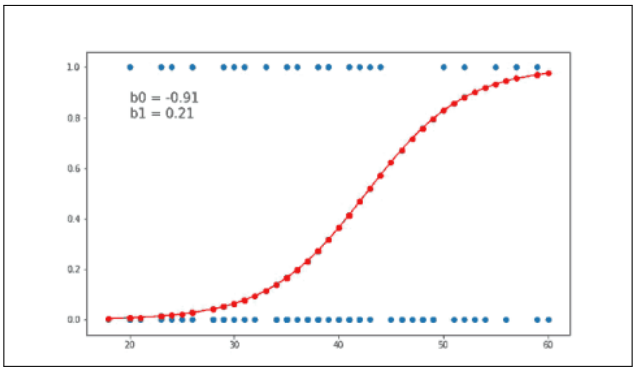
Kapitel 4

CROSS VALIDATION / BOOTSTRAP

When we are working with different machine learning models, we need two different data sets. The first set is often referred to as the training data set, and is used to create the model and define the curve seen in 4.1. The other dataset is referred to as the test set and is used to check if the model performs well. The data between the two sets can not be identical because we want to evaluate how the model performs on new data sets.

Because we often dont have well defined test sets available, we need some methods to gain test data. Here we have several diffenret options. Some methods will not provide new data, but instead use the training data to estimate the test error (the performance). Another approach, which is the one discussed here, is a method where you hold out some part of the data, in the training process and use the hold out data as a test data set.

When this is done with several different splits (different hold out sets) and we evaluate our model using the approach, we are achieving cross validation. The size of the hold our set can vary. If we choose the heaviest computational split, we only contains one sample in the hold out set. This is called leave-one-out-cross-validation LOOCV.



Figur 4.1: Logistic regression model illustration

When we have gained several different splits from using cross validation, we need to evaluate how our model performs in these different testing scenarios. We often use mean least square to evaluate our model. The formula for calculating this for our model with cross validation can be seen in 4.1.

$$CV_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k \tag{4.1}$$

Kapitel 5

SUBSET SELECTION

Kapitel 6

SHRINKAGE METHODS / DIMENSION REDUCTION METHODS