DAT405 Introduction to Data Science and AI

2019-2010, Reading Period 1

Assignment 2: Regression and classification (and ethics)

There will be an overall grade for this assignment. Everyone should attempt questions 1-3 Those aiming for a higher grade should also attempt question 4.

1. A company is considering using a system that will allow management to track staff behaviour in real-time, gather data on when and who they email, who accesses and edits files, and who meets whom when. The HR (human resources, or personnel) department at the company is in favour of introducing this new system and believes it will benefit the staff at the company. The company's management is also in favour. Discuss whether introducing this system raises potential ethical issues. (See slides from lecture 2 for advice on how to approach this.)

2. The following web page lists the selling prices of villas in Landvetter that were sold in the past 6 months. Find a linear regression model that relates the living area to the selling price. (You may transcribe the values from the web page into your program or into a data file by hand, or you can write a program to do this, but don't spend too much time doing this because "web scraping" is not the main objective of this assignment!)
   https://www.hemnet.se/salda/bostader?location_ids%5B%5D=940808&item_types%5B%5D=villa&sold_age=6m

   a. What are the values of the slope and intercept of the regression line?
   b. Use this model to predict the selling prices of houses which have living area 100 m^2, 150 m^2 and 200 m^2.
   c. Draw a residual plot.
   d. Discuss the results, and how the model could be improved.

3. Use a confusion matrix and 5-fold cross-validation to evaluate the use logistic regression to classify the iris data set.

4. Consider the classification models for the iris data set that are generated by k-nearest neighbours (with some different values for k, and with uniform and distance-based weights), by logistic regression and by support vector machines. Calculate confusion matrices for these models and discuss the performance of the various models.

Submitting work

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment.

If you upload a zip file, please also upload any PDF files separately (so that they can be viewed more conveniently in Canvas).

Deadline: Monday 16 September 2019 at 12:00 (noon).