UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

FYS-2021 - Exercise Set 4

Multivariate Methods
Department of Physics and Technology
Faculty of Science and Technology

From book (fourth edition in parenthesis)

5.7 (5.10)

Parametric methods

Problem 1

The data file `global-temperatures.csv` contains the average global annual temperatures spanning from the year 1880 to 2017.

(1a) Use the file from canvas and perform linear regression, with temperature as a function of years, and plot the results.

(1b) Briefly explain what the R^2 value tells you, and calculate the R^2 value for this model.

(1c) Assuming the regression model is on the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, then what is the interpretation of the estimator $\hat{\beta}_1$?

(1d) Plot the residuals and comment on the result with regards to the assumptions made in the regression model.

Multivariate methods

Problem 2

In this problem, we will use multiple linear regression to predict the fuel consumption of cars (measured in miles per gallon). The available predictors are: (i) cylinders, (ii) displacement, (iii) horsepower, (iv) weight, (v) acceleration and (vi) model year. The dataset is named `auto-mpg.csv`, and can be found in Canvas.

(2a) Implement your own function for estimating the linear-regression parameters, and use this to fit a regression model to the fuel-consumption data.

(2b) Explain why the magnitudes of the estimated coefficients are so different. What could be done to prevent this from happening?

(2c) Compute R^2 and use this to evaluate the quality of your model.

(2d) Experiment with removing different predictors from the model. Do you think all predictors are equally necessary? Explain.

Problem 3

In this problem you will use the Bayes' classifier described in Chapter 5.7 to create a system for detecting whether an SMS is spam or not, based on its contents. The Bag Of Words (BOW) representation[1] of 5574 text messages is provided in `sms-spam-bow.csv`. In this file, each row represents a single text message. The first element is the label (0 = not spam, 1 = spam), and the rest of the row is the BOW-representation of the message. The raw data can be found in `sms-spam.txt`. This corpus has been collected from free or free for research sources at the Web. More info can be found at http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

(3a) Implement your own version of the classifier described in Chapter 5.7. You can assume that the features are binary (Bernoulli).

(3b) Split the dataset into training- and test-sets (80 % training and 20 % test). Use the training data to estimate the parameters of the classifier, and use the test set to evaluate the performance. Report the confusion matrix and accuracy.

[1] You can read more about the bag of words representation in Chapter 5.7

Problem 4

Assume that the random vector $\boldsymbol{X} = [X_1, \ldots, X_d]^{\mathrm{T}}$ has distribution

$$p(\boldsymbol{x}) = \begin{cases} C(\boldsymbol{\lambda}) e^{-\boldsymbol{\lambda}^T \boldsymbol{x}}, & \min\{x_1, \ldots, x_d\} > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $C(\boldsymbol{\lambda})$ is a normalizing constant, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_d]^{\mathrm{T}}$ is a parameter vector with strictly positive elements.

(4a) Show that, if the distribution integrates to one, we have

$$C(\boldsymbol{\lambda}) = \prod_{j=1}^{d} \lambda_j$$

(4b) Argue that the X_i's are statistically independent.

(4c) Compute the Maximum Likelihood estimator for $\boldsymbol{\lambda}$. Compare your result to the Maximum Likelihood estimator for $\lambda = \frac{1}{\beta}$ in the univariate exponential distribution.

Problem 5

Suppose you have a classifier that classifies to the positive class for

$$\hat{\mathrm{P}}(C_1|x) > \theta, \quad 0 < \theta < 1,$$

and that the number of false positives, f_p, classified can be modelled as

$$f_p = N \cos\left(\frac{\pi\theta}{2}\right)$$

where N is the total of negative training samples. The true positives can be modelled as

$$t_p = T \cos\left(\frac{\pi\theta}{2}\right)\left(2 - \cos\left(\frac{\pi\theta}{2}\right)\right)$$

where T is the number of positive training samples.

(5a) Find the true positive and false positive rates.

(5b) Plot the receiver operating characteristics (ROC) curve.

(5c) Calculate the AUC score. Hint: it might be useful to obtain an expression for the true positive rate as a function of the false positive rate.

Problem 6

Different classifier algorithms has probability outputs as given in `soft-classifications-#.csv` (the `#` is a number).

All these files are structured identically: The first column is the true label of the classified data point and the second column is the classified probability of that data point belonging to the positive class. In the file header you can find more info of number of classes and datapoints.

(6a) Implement a script that takes a dataset and makes two plots: The first plot should contain the ROC curve and the second plot should visualize the accuracy, tp-rate and fp-rate as a function of threshold. The script should also calculate the AUC-score and estimate the optimal ROC cut-off point.

(6b) What can you say about the classifiers from the plots? Which ones are good and which ones are bad, and why?