The Elements of Statistical Learning, WS 2018/19

Jilles Vreeken and Tobias Marschall
Michael Scherer, Fawaz Dabbaghie, Aryan Kamal
Center for Bioinformatics & Max Planck Institute for Informatics
CISPA Helmholtz Center i.G. & Cluster of Excellence MMCI

2018-11-15

Problem Set 3

Deadline: Thursday, November 29. 2018, 10:00 a.m.

Please read and follow the following requirements to generate a valid submission.
This problem set is worth 50 points. You may submit your solutions in groups of two students. The solutions to the theoretical problems should be submitted either digitally (in .pdf format) to `mscherer@mpi-inf.mpg.de` or as a hard copy before the lecture. **Label your hard copy submissions with your name(s).**
Solutions to programming problems and resulting plots need to be submitted in digital format (.pdf). For the programming problems you have to submit an executable version of your code (R script).

For digital submissions the subject line of your email should have the following format:

`[SL][problem set 3] lastname1,firstname1;lastname2,firstname2`

Please include the numbers of the problems you submitted solutions to (both digitally and analogously) in the email's body. **Please make sure that all the files are attached to the email.** The attached files should only include an executable version of your code as .R file and **one** .pdf file with all the other solutions.

Problem 1 (T, 10 Points)

Derive the variance formula:

$$\mathrm{Var}(\frac{1}{k}\sum_{i=1}^{k} X_i) \; = \; \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

where X_i , $i = 1, \ldots, k$, are identically distributed random variables with positive pairwise correlation ρ and $\mathrm{Var}(X_i) = \sigma^2$ for $i = 1, \ldots, k$. This formula will play a central role in Chapter 8 of ISLR (Tree-Based Methods).

Problem 2 (T, 5 Points)

Logistic Regression
The book introduces the conditional probabilities and the log-odds for 2-way logistic regression. Extend this model to logistic regression for k response classes.

Problem 3 (T, 5 Points)

(a) (1P) What is specificity? Write down the formula and explain in your own words.

(b) (1P) What is sensitivity? Write down the formula and explain in your own words.

(c) (1P) Why is it useful to look at specificity and sensitivity instead of just considering the misclassification rate?

(d) (1P) What does the ROC curve display?

(e) (1P) What would the ROC curve of a perfect model look like? Why? Draw the plot. Don't forget to label the axes.

Problem 4 (T, 10 Points)

In this exercise we will investigate the so-called **curse of dimensionality**. (Exercise 2.3 in ESL)
Consider N data points uniformly distributed in a p-dimensional unit ball centered at the origin. Suppose we

consider a nearest-neighbor estimate at the origin. Show that the median distance from the origin to the closest data point is given by the expression:

$$d\left(p,N\right)=\left(1-\frac{1}{2}^{1/N}\right)^{1/p}$$

What does this mean for the k-nearest neighbor algorithm?

Hint: Consider that the volume of a p-dimensional sphere with radius r is given by $V(r,p) = G(p)r^p$, with $G(p)$ a dimension-dependent constant. The probability that a point falls into a sphere of radius r is proportional to the sphere's volume since the points are uniformly distributed.

Problem 5 (P, 20 Points)

Go through **4.6 Lab: Logistic Regression and LDA** (ISLR p. 154–167), doing this lab will make it easier to solve the following programming exercise. We will do classification using LDA and QDA on a speech recognition dataset. The dataset contains digitized pronunciation of five phonemes: `sh` as in "she", `dcl` as in "dark", `iy` as the vowel in "she", `aa` as the vowel in "dark", and `ao` as the first vowel in "water" which represent the responses/classes (column name `g`). The dataset contains 256 predictors (log-periodograms, which is a common method used in speech recognition to represent voice recordings).

(a) (2P) Download and load the phoneme data set (`phoneme.csv`) from the course website. Split the dataset into training and test set according to the `speaker` column. Be sure to exclude the row number, `speaker` and response columns from the features. *Useful functions:* `strsplit()`, `grepl()`

(b) (5P) Select the two phonemes `aa` and `ao`. Fit an logistic regression model on this data set and report train and test error. *Useful functions:* `glm()`

(c) (5P) Repeat step (b) using LDA and report your findings. Would you prefer logistic regression or LDA in this example? Why? *Useful functions:* `lda()` from the `MASS` package

(d) (3P) Generate confusion matrices for the logistic regression and the LDA model for `aa` and `ao`. Which differences can you observe between the models?

(e) (5P) Fit an LDA model using the full training data set (all phonemes). Report train and test error. Plot the projection of the training data onto the first two canonical coordinates of the LDA using the `plot()` function.