The Elements of Statistical Learning, WS 2017/18
Prof. Dr. Dr. Thomas Lengauer
Michael Scherer
D3: Computational Biology & Applied Algorithmics
Max Planck Institute for Informatics

2017-11-15

Problem Set 3

Deadline: Wednesday, November 29. 2017, 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, 8 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:

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

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 2 (T, 12 Points)

Logistic regression:

(a) (2P) In which setting is logistic regression applicable? Why is linear regression not applicable in such a setting?

(b) (2P) In general, what is the meaning of odds? Write down the formula and explain in your own words.

(c) (1P) Prove that Equation (4.2, ISLR) is equivalent to Equation (4.3, ISLR). Give all transformation steps!

(d) (1P) Show that a change in the predictor variable X_i from X_i to $X_i + \Delta$ leads to an odds ratio of $\exp(\beta_i\Delta)$, i.e.,

$$\frac{\mathrm{odd}(X_i + \Delta)}{\mathrm{odd}(X_i)} = \exp(\beta_i\Delta).$$

(e) (1P) Assume the number of features to be $p = 1$, i.e.,

$$Pr(Y = 1|X) = p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

How do we have to choose X if we want to have p(X)=0.5. What does this probability tell us?

(f) (5P) 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, 10 Points)

Linear Discriminant Analyses (LDA) and Quadratic Discriminant Analyses (QDA):

(a) (5P) (Exercise 4.7.2 in ISLR):
This problem relates to the LDA model. It was stated in the text that classifying an observation to the class for which (4.12, ISLR) is largest is equivalent to classifying an observation to the class for which Equation (4.13, ISLR) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k,\ \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

(b) (5P) (Exercise 4.7.3 in ISLR):
This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class-specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature. Suppose that we have k classes and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in Equation (4.11, ISLR). Show that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Problem 4 (P, 20 Points)

Go through **4.6 Lab: Logistic Regression, LDA, QDA, and KNN** (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) (3P) Fit an LDA model, compute and report train and test error. *Useful functions:* `lda()` from the `MASS` package

(c) (3P) Plot the projection of the training data onto the first two canonical coordinates of the LDA using the `plot()` function. Investigate the data projected on further dimensions using the `dimen` parameter.

(d) (3P) Select the two phonemes `aa` and `ao`. Fit an LDA model on this data set and repeat the steps done in (b).

(e) (6P) Repeat steps (b) and (d) using QDA and report your findings. Would you prefer LDA or QDA in this example? Why? *Useful functions:* `qda()` from the `MASS` package

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