Deadline: Thursday, November 28. 2019, 10:00 a.m.

Please read and adhere to 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 esl-ta@mpi-inf.mpg.de or as a hard copy before the lecture. **Label your hard copy submissions with your name(s), Matriculation number, and Exercise group.**

Solutions to programming problems, including resulting plots and answers to the questions, need to be submitted in digital format (PDF). For the programming problems you have to submit an R script (that means it needs to run with `Rscript your_file.R`). Write comments within the script to explain what your code is supposed to do.

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

[SL][problem set 3][*tutorial group*]_lastname1_firstname1_lastname2_firstname2

where *tutorial group* is either "Mo" or "We" depending on your assigned time slot. 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).   **Logistic regression**

1. (1.5P) In which setting is logistic regression applicable? Why is linear regression not applicable in such a setting?

2. (1P) What do we model with logistic regression? How are the independent variables and obtained probalities related?

3. (1P) In general, what is the meaning of odds? Write down the formula and explain in your own words.

4. (1.5P) Prove that Equation 4.2 (ISLR p. 132) is equivalent to Equation 4.4 (ISLR p. 132). What is the relationship between the logistic and the logit function?

5. (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.,
$$\frac{\text{odd}(X_i + \Delta)}{\text{odd}(X_i)} = \exp(\beta_i \Delta).$$

6. (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?

7. (3P) 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 2 (T, 10 Points).   **LDA and QDA**

1. (5P) **Linear discriminant analysis** (Exercise 4.7.2 in ISLR):
   It was mentioned that an observation to the class for which Equation 4.12 (ISLR p.139) is largest is
   equivalent to classifying an observation to the class for which Equation 4.13 (ISLR p.140) 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.

2. (5P) **Quadratic discriminant analysis** (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 p.139). Show that in this
   case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Problem 3 (T, 10 Points).   **Variance and Correlation**
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 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).

1. (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()`

2. (3P) Fit an LDA model, compute and report train and test error. *Useful functions:* `lda()` from the
   `MASS` package

3. (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.

4. (3P) Select the two phonemes `aa` and `ao`. Fit an LDA model on this data set and repeat the steps
   done in (b).

5. (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

6. (3P) Generate confusion matrices for the LDA and QDA model for `aa` and `ao`. Which differences
   can you observe between the models?