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

2018-10-31

Problem Set 2

Deadline: Thursday, November 15. 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 2] 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, 12 Points)

(Exercise 3.3 in ESL, cf. ESL, Section 3.3.2, p.51)
Consider all estimates $\tilde{\theta}$ of the linear combination of the parameters $\theta = a^T\beta$ that are unbiased, i.e. $\mathrm{E}\left(\tilde{\theta}\right) = \theta$.
Prove the **Gauss-Markov theorem**: The least squares estimate $\hat{\theta} = a^T\hat{\beta}$ has variance no bigger than that of any other linear unbiased estimate of θ that has the form $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$. Is it also the *best* (in terms of test error) linear unbiased estimate (argue with the bias-variance tradeoff)?
Hint: Consider $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ and the least squares estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Assume an arbitrary linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$ is unbiased for parameter $\theta = a^T\beta$ and calculate its variance: $Var(\tilde{\theta}) = Var(\hat{\theta} + (\tilde{\theta} - \hat{\theta}))$.

Problem 2 (T, 10 Points)

The R^2 statistic is a common measure of model fit corresponding to the fraction of variance in the data that is explained by the model. In general, R^2 is given by the formula

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

Show that for univariate regression, $R^2 = Cor(X, Y)^2$ holds.
Bonus: Show that in the univariate case, $R^2 = Cor(Y, \hat{Y})^2$ holds.

Problem 3 (T, 8 Points)

(Exercise 2.7 in ESL).
Suppose we have a sample of n pairs (x_i, y_i) drawn i.i.d. from the distribution characterized as follows:

$$\begin{aligned} x_i &\sim h(x), \quad \text{the underlying density} \\ y_i &= f(x_i) + \varepsilon_i, \quad \text{where } f \text{ is the regression function and } \mathrm{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2 \end{aligned}$$

We construct an estimator for f that is *linear* in the y_i:

$$\hat{f}(x_0) = \sum_{i=1}^{n} l_i(x_0; \mathcal{X}) y_i,$$

where the weights $l_i(x_0, \mathcal{X})$ do not depend on the y_i, but do depend on the entire training sequence of x_i, denoted here by $\mathcal{X}$. Show that linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $l_i(x_0; \mathcal{X})$ in each of these cases.

Problem 4 (P, 20 Points)

The book provides a practical guide for linear regression. Go through **3.6 Lab: Linear Regression** (ISLR p. 109–119), doing this lab will make it easier to solve the following programming exercise. This exercise uses the *Auto* data set which is contained in the R package *ISLR*. Install the R package *ISLR*.

(a) Create scatterplots between all the variables. Is the relationship between those variables linear? Describe the connection between the variables. (Exclude the *name* variable, which is qualitative.)

(b) Detect the variables in the scatterplots that appear to be most highly correlated and anti-correlated, respectively. Justify your choice using the *cor()* function.

(c) Perform simple linear regression with *mpg* as the response using the variables *cylinders, displacement, horsepower* and *year* as features. Which predictors appear to have a statistically significant relationship to the outcome and how good are the resulting models (measured using R^2)?

(d) Use the **lm()** function to perform a multiple linear regression with *mpg* as the response and all other variables except *name* as the predictors. Use the **summary()** function to print the results. Compare the full model to those generated in (c) in terms of their model fit. What can you observe in the different models concerning the significance of the relationship between response and individual predictors? What does the sign of the coefficient tell you about the relationship between the predictor and the response?

(e) Use the **plot()** function to produce diagnostic plots of the linear regression fit. Does the residual plot suggest any non-linearity in the data? Does the residual plot suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

(f) Generate three linear models that are based on all pairwise interaction terms (X_1X_2) for *cylinders*, *weight*, and *year* as well as on the non-linear transformations $\log(X)$, $\sqrt{X}$, X^2 for the *displacement* variable (one per linear model). Comment on your findings.