Exam questions:

Chapter 2:

- What is the difference between prediction and inference?
- Explain the reducible and irreducible error!
- In the context of this book: how would you explain the difference between parametric and non-parametric methods and how does it relate to model interpretability?
- Is it true that a non-parametric and flexible model is always resulting in a better prediction accuracy? Motivate your answer by providing an example.
- What are non-linear methods?
- Explain the difference between supervised and unsupervised learning.
- What is the difference between regression and classification? Is this distinction clear in the statistical methods or is there an overlap in the application with respect to qualitative and quantitative response variables?
- There exist many statistical methods and selecting the best approach for a given problem can be challenging? Which measure for assessing the model accuracy could you use? Is it important to compute this measure on previously unseen data? Why is that? How do we call these data sets and corresponding accuracy measures?
- Explain what is going on in the figure below. Explain the different lines. What is the meaning of the dashed horizontal line? Why is the grey line in the right panel always lower than the red line?

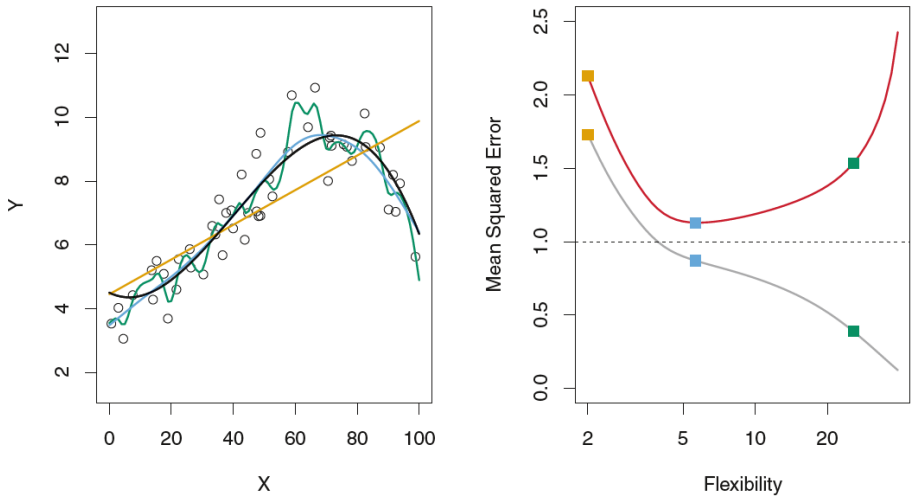

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

- Explain the Bias-Variance trade-off. Give the equation and use a figure to describe the concept. How does the bias-variance trade-off relate to the reducible and irreducible error? What is the ideal situation regarding the bias and the variance?
- Show that the relation in the equation below holds:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

- How do you compute the overall expected test MSE? Is it possible to disentangle the variance and bias component in a real world example.
- What is the relation between bias/variance and model flexibility? Explain by using a figure! Do models with low flexibility always lead to a high bias? What are the ideal quantities of bias and variance? Indicate this point on the figure.
- Give the definition and equation for the error rate.
- Explain the concept behind the Bayes classifier in the case of two classes. What is the Bayes decision boundary? How does the Bayes error rate relate to the irreducible error? Is this a practical classifier to use in a real world case.
- Explain the K-nearest neighbor classifier? How does it classify an unseen data point? Is there a model that is fitted in KNN? Is KNN based on underlying assumptions? If so, which assumptions? How are the KNN decision boundaries calculated? How do you alter the model flexibility?
- In the case of KNN, a low training MSE means also a low test MSE. Is this statement true? Explain. How does KNN relates to the Bayes classifier?
- Exercises 2.4 from 1 to 7.

Chapter 3:

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \quad (3.4)$$

- Proof by using equation 3.3 that 3.4 is true.
- Proof that equation 3.4 is equal to the Pearson correlation for standardized data.
- If you compare correlation with linear regression would you think of correlation as supervised or unsupervised method?
- What have a model trained on a training set and a sample mean in common?
- How should you interpret the 95% confidence interval of a parameter estimate?
- How do you compute the R^2 statistic? Should you prefer this one above the RSE? What is the difference?
- How would we test whether all parameters in a model are equal to zero? How would the test statistic change if we test for a subset of parameters that equal zero? How does it relate to the significance test of a single parameter?
- Can you make an argument, why we want use an omnibus test to test whether all parameters equal zero rather than to perform the tests individually for all the parameters?

- When fitting a linear model on a data set, one might expect that the RSS decreases when adding an additional predictor variable to the model? However, it might occur that the RSE will increase. How is this possible? In the context of machine learning, is it our optimal goal to achieve the lowest RSS possible?
- Given a multiple regression prediction model, what are the three sorts of uncertainty associated with the prediction?
- In a multiple regression model with significant interaction terms it is common practice to drop the main effects if the p-values associated with their coefficients are not significant. Doing so we increase the residual degrees of freedom (or decrease RSE) which is advantageous for our parameter estimates. True or false. Elaborate!
- What is the difference between an outlier and high leverage observation? What is a leverage statistic? How does it relate to the hat matrix → y_hat = H x y with $H = X(X^TX)^{-1}X^T$
    - more info can be found on https://stats.stackexchange.com/questions/208242/hat-matrix-and-leverages-in-classical-multiple-regression
- What is the meaning of these equations? How are they related?

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- What happens with the estimated standard errors when the error terms are correlated?
- What is collinearity? How does it affect the model fit? Explain by using a contour plot of RSS values two model coefficients.
- In KNN regression how would you decide on the optimal neighborhood. What does it mean if K is small/large?
- If you had to develop a small electronic device that needs to predict a certain value. Would you opt for a model-based regression approach of KNN regression? Motivate your choice! In which setting would a parametric approach outperform a non-parametric approach?
- In the context of machine learning, would you regard KNN regression as a lazy learner or an eager learner? Explain! Even if KNN would slightly outperform the linear regression, why would you opt for the linear regression?
- In KNN regression or classification, why do we express the number of nearest point as 1/k when plotting the test MSE?
- In general, it seems that the non-parametric methods, like KNN, seems to outperform parametric methods, like linear regression, especially when the true relation is non-linear. In practice it may happen that the linear regression still outperforms KNN even if the underlying relation is non-linear. When is this the case? It has to do with an unwanted n/p ratio! What is the term to describe this phenomena? Explain this in a graphical way!
- Exercises 3.7: 2