Exam questions:

Chapter 10:

- Describe in your own words the concept behind principal component analysis. What are score and loadings. Illustrate PCA by an example. What does it mean: “PCA provide low-dimensional surfaces closest to the observations”
- How many principal components can I compute for a dataset of *n* observations and *p* predictors? How can I decide on the optimal number of principal components to display my data?
- What is the effect of scaling the variable on the principal component analysis? When should scaling be applied and when should it be not recommended?
- What is a biplot? What is a scree plot?
- How would you use PCA in a regression setting? How to define the number of principal component to use? Where should we be cautious about when new test data comes at hand?
- Could PCA work on mixed data? Consider the predictor as binary, ordinal or categorical?
- What is K-means clustering. Explain graphically! What are the underlying assumptions and properties of the clustering? How would you decide on the optimal K? Is K-means invariant to scaling? Explain!
- What is the difference between clustering and PCA?
- What is the objective of clustering? What would you tend to cluster: observations or features?
- The K-means clustering algorithm aims to minimize the within-cluster variability over all K clusters is minimized. There is a trivial solution to this problem. Which one?

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

- Is K-means a deterministic algorithm? Does it find a global optimum?
- What is hierarchical clustering? How should we interpret the dendrogram? Indicate the leaves and branches and root of the tree. How would you cut the tree? What are the assumptions and properties of the algorithm?
- Why do we need the notion of linkage in hierarchical clustering? Which types of linkage mechanism do you know? Explain them graphically.
- What are the differences between correlation-based distances and say, Euclidean distances? Which one should be preferred?
- How could we gain confidence in our unsupervised results?
- Exercise 10.7.1, 10.7.4

Chapter 6:

- Which three main methodologies are available for feature selection!
- Is Lasso based on maximum likelihood estimation? How should we do inference? How do we obtain the lambda value?

- Explain BIC, Cp, AIC, adjusted R^2
- What is the deviance?
- Give the algorithm for best subset selection? What are the two main disadvantages when the number of predictors are large?
- How does forward selection work? How many models does it consider? What is the disadvantage?
- What happens in linear regression when p >>> n?
- Describe backward selection. What is an important requirement?
- Why is AIC proportional to Cp in case for least square models.
- In the algorithm for best, forward and backward subset selection one uses the RSS to select the most optimal model conditionally for *d* predictors. Why should one use RSS instead of Cp, AIC, BIC?
- Give the cost function of ridge regression. Why does it not apply a constrain on the intercept?
- Shrinkage methods are not scale equivariant. Explain!
- What is the main difference between ridge and lasso regression? What is the advantage of lasso?
- How does lasso regression relate to best subset selection? Explain by using the constraint lasso object function.
- How does ridge regression relate to PCR?
- Explain by using a graph why lasso regression sets coefficient estimates to zero?
- What happens with correlated predictor variables when using shrinkage methods?
- Can we do formal testing on the coefficient estimates when using shrinkage methods?
- What is the main assumption behind PCR? How is PLS different? Under which conditions will PCR operate optimally?
- Can PCR be used as a feature selection method?
- In unsupervised learning with principal component analysis does it matter if we focus on the rows or on the columns? What is the difference in interpretation if we focus on the rows/columns? Is this different for principle component regression?
- Should we include the response variable in the principal component regression?
- What can we say about the residuals of a least square regression when p > n?
- Can we use Cp, AIC, BIC or adjusted R^2 for model selection with high dimensional data?
- What is meant by the curse of dimensionality?
- Exercises 6.8.1, 6.8.2, 6.8.3, 6.8.4