# Question Booklet

- **The final exam contains 5 questions and is scheduled for 2.5 hours. At maximum you can earn 50 points.**
- **Please verify if this question booklet consists of 11 pages, and that all questions are readable, else contact the examiners immediately.**
- **This is an open-book exam. You are allowed to consult the books, slides, and lectures while writing it. You are not allowed to consult others. Plagiarism is not condoned.**
- **Answers without sufficient details are void: you get no points for "yes" or "no" answers, unless the question specifically asks this. Explain all answers in your own words. Clearly state any assumptions you make.**

**Problem 1** (On Errors and Models) **(10 points)**

(a) Consider Figure 1. Which of the following three options correctly describes what is happening in the figure? *You do not need to explain your answer.* (1 point)

i) Starting at *flexibility*= 1, with increasing flexibility, the increase in variance is smaller than the decrease in bias, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, the variance starts to increase more rapidly than the bias decreases, hence causing a upward trend.

ii) Starting at *flexibility*= 1, with increasing flexibility, the increase in bias is smaller than the decrease in variance, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, variance starts to increase more rapidly than the bias decreases, hence causing a upward trend.

iii) Starting at *flexibility*= 1, with increasing flexibility, the increase in bias is smaller than the decrease in variance, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, bias starts to increase more rapidly than the variance decreases, hence causing a upward trend.

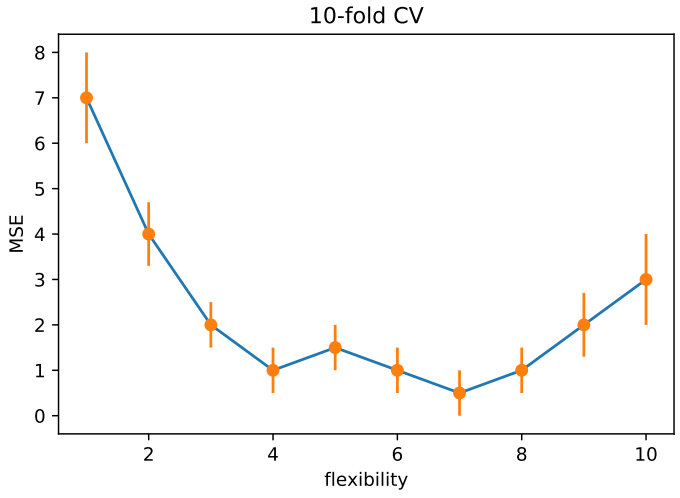


Figure 1: Test MSE for 10-fold Cross Validation (CV) for an unknown data set. Orange bars indicate the standard error.

(b) According to Figure 1, which model (flexibility), would you select? Explain why you chose this model. (2 points)

(c) State, for each of the three settings below, what will happen both in terms of bias and variance when we make the proposed change to the learning procedure. Indicate with a (+) if the given quantity increases, (–) if the given quantity decreases, and (=) if there is no change. Provide also an explanation, why the changes introduced in the specific model/method do (not) change the flexibility. (3 points)

| Action | Bias | Variance | Flexibility | Explanation |
|---|---|---|---|---|
| 1) Fitting data generated by $Y = \beta X + \mathcal{N}(0,1)$ instead of $Y = \beta X + \mathcal{N}(0,10)$. | | | | |
| 2) Changing from $K = 2$ to $K = 10$ in the K-nearest neighbor classifier. | | | | |
| 3) Setting budget $C$ in the Support Vector Classifier to a higher value. | | | | |

(d) In linear regression, why do we *in general* assume that the error (noise) term averages to zero? (1 point)

(e) Describe in your own words the difference between a parametric and a non-parametric method. (1 point)

(f) Would you then consider the following methods parametric or non-parametric? Explain why each of the models fall under the definition of parametric/non-parametric. (2 points)

1) Logistic Regression
2) Decision Trees
3) Linear Discriminant Analysis
4) $k$-Nearest Neighbors ($k$-NN)

**Problem 2** (Regression) **(15 points)**

You have been given data in Table 1 from a small experiment.

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| -2 | -3 | -1 |
| -2 | -1 | 1 |
| 1 | 2 | 2 |
| 3 | 2 | 3 |

Table 1: Observations for predictor variable $X_1$ and $X_2$ and target variable $Y$.

(a) You want to perform univariate linear regression of predictor $X_1$ on response $Y$. Recall that simple linear regression takes the form $Y = \beta_1 \mathbf{X_1} + \beta_0$, but that it is often convenient to formulate it as $\mathbf{X}\beta = \mathbf{Y}$ with $\beta = \begin{bmatrix}\beta_1 \\ \beta_0\end{bmatrix}$ and $\mathbf{X} = [\mathbf{X_1}; \mathbb{1}]$. Using the following conversion,

$$(\mathbf{X^T}\ \mathbf{X})^{-1} = \begin{bmatrix}0.06 & 0 \\ 0 & 0.25\end{bmatrix},$$

find the least square estimators $\hat{\beta}_1$ and $\hat{\beta}_0$. Explain the reasoning behind each step. (3 points)

Assume now that we collect a thousand data points to predict the response variable $Y$ using one million predictors.

(b) Is least squares linear regression appropriate in this setting? Explain your answer. (1 point)

(c) Assume we apply forward stepwise selection and the results indicate that 9 of these predictors lead to a good predictive model on a given training data set. Can we ensure that these 9 predictors are the optimal set, i.e., the set of 9 predictors with minimum MSE? Explain why (not)? (1 point)

We now want to perform Principal Component Analysis (PCA) for dimensionality reduction using the data in Table 1, i.e. we want to reduce the two dimensional data $X = [X_1, X_2]$ to one dimension.

(d) Compute the covariance matrix. Indicate all the steps that you follow to compute it. (The final result alone does not give any points.) *Hint:* The covariance matrix takes the form (2 points)

$$\begin{pmatrix}Cov(X_1, X_1) & Cov(X_2, X_1) \\ Cov(X_1, X_2) & Cov(X_2, X_2)\end{pmatrix}$$

(e) Now compute the first principal component using the covariance matrix computed before. Explain what you do in each step. *Hint:* (2 points)

$$det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - cb$$

*If you did not solve part (a), you may use this covariance matrix:*

$$\begin{pmatrix} 4.5 & 4 \\ 4 & 4.5 \end{pmatrix}$$

(f) Draw the first and second principal component in a plot. It is sufficient to indicate (i) the angle between the two principal components, (ii) the angle between each principal component and the x-axis and (iii) the module (length) of each principal component. (1 point)

(g) What important assumption does PCA make on the statistical relationship (i.e., dependence) between the predictor variables? (1 point)

(h) Do *in general* the LS estimator for the first predictor (e.g. $\hat{\beta}_1$ computed in a)) and the first principal component of PCA (e.g. computed in e)) point to the same direction? Explain why (not)? (1 point)

Suppose we have two data sets with the same $n = 500$ observations and the same $p = 20$ predictors, but with two different response variables $Y_1$ and $Y_2$. In data set 1 response variable $Y_1$ is a function of all the predictors, whereas in data set 2 the response variable $Y_2$ only depends on two of the predictors. We perform PCA for dimensionality reduction and extract the first five principle components for both the data sets.

(i) Will the principle components be the same for both data sets? Explain why (not)? (1 point)

(j) Next we perform Principal Component Regression (PCR) using the first five principle components that we extracted. For which data set would you expect to have a better accuracy? Explain why. (1 point)

(k) Assume you now also perform Partial Least Squares (PLS) regression with 5 principle components for both datasets. Which method, PCR or PLS do you expect to have the smaller training Residual Sum of Squares (RSS) for $Y_1$ and for $Y_2$? Explain why. (1 point)

**Problem 3** (Classification) **(10 points)**

Consider now a binary classification problem, i.e. the label can assume either $(Y = 1)$ or $(Y = 0)$.

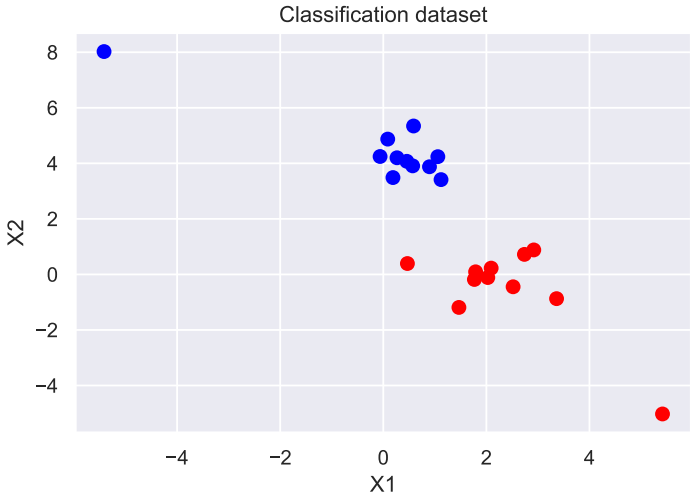


Figure 2: Classification data set with two classes.

Looking at Figure 2 you are wondering, if you should first pre-process the data to get rid of the outliers before training a classification model. To save time you first take a look at the different methods:

(a) Are the following methods sensitive to outliers, where outliers are seen as a set of predictors that are out of the predictor distribution (in Figure 2, the two observations in the top-left and bottom-right corners)? Explain why (not). (2 points)

1) Logistic Regression
2) Decision Trees
3) Support Vector Machines (SVM)
4) $k$-Nearest Neighbours ($k$-NN)

Let's take a closer look at Support Vector Machines (SVM). Recall that an SVM are defined as follow:

$$\begin{aligned}
&\underset{\beta_0,\ldots,\beta_p,\ \xi_1,\ldots,\xi_N}{\text{maximize}} && M && (3.1)\\
&\text{subject to} && \|\beta\| = 1 && (3.2)\\
& && \xi_i \geq 0 && (3.3)\\
& && y_i f(x_i) \geq M(1-\xi_i) \quad \text{for } i = 1,\ldots,N && (3.4)\\
& && \sum_{i=1}^{N} \xi_i \leq C && (3.5)
\end{aligned}$$

(b) Describe the constraints (3.3) and (3.5) in your own words. Which different values can $\xi_i$ take and what do the different values express? (2 points)

(c) Describe the purpose of constraints (3.2) and (3.4) in your own words. How are these two constraints related to each other? (3 points)

Now assume that there there would not be just two, but $K = 10$ classes. Let $f_k(x)$ denote the density function of $X$, i.e $\Pr(X = x \mid Y = k)$, for the observation that comes from the $k$th class. Recall, according to Bayes' Theorem: $\Pr(Y = k \mid X = x) \propto \pi_k \cdot f_k(x)$ . You suspect your data to follow an Poisson distribution with distinct $\lambda_k$ for each of the $k$ classes, where

$$f(x; \lambda_k) = \frac{(\lambda_k)^x e^{-\lambda_k}}{x!}$$

(d) Derive the discriminant function, in a similar way that we did for a Gaussian likelihood for Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). (2 points)

(e) Is the above derived discriminant function linear in terms of $x$? Explain why (not). (1 point)

**Problem 4** (Beyond Linear Regression) **(5 points)**

Now assume we are interested in how a predictor $X_1$ relates to the response variable $Y$. Prior analysis shows that this relationship is non-linear. One modeling option would be to use regression splines.

(a) How many degrees of freedom does a regression spline have if we use polynomials of degree $d = 3$, have $K = 10$ knots, and require the spline to be continuous at the knots up to the first derivative. Explain your answer. (1 point)

(b) Up to which derivative do we have to enforce continuity at the knots if we want 25 degrees of freedom, polynomials of degree $d = 4$, and have $K = 10$ knots? Explain your answer. (1 point)

Let's now consider Smoothing Splines.

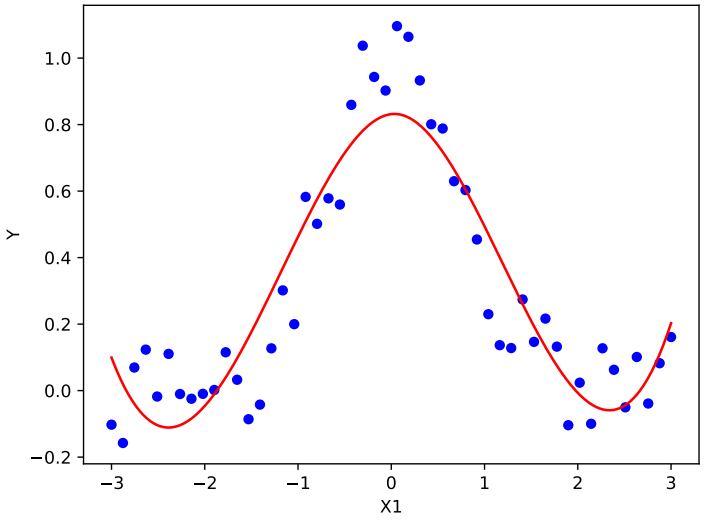


Figure 3: Smoothing spline with unknown parameter $\lambda$

(c) Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of $n$ points using the following formula:

$$\hat{g} = \arg\min_{g} \left( \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[ g'''(x) \right]^2 dx \right) \quad .$$

To which value do we have to set $\lambda$ to get a fit as shown in Figure 3 - i) $\lambda = 0$; ii) $\lambda = 1$; or, iii) $\lambda = \infty$? Explain your answer. (1 point)

Recall that a multiple linear regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Whereas a Generalized Additive Model (GAM) can be written as:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

(d) What are the functions $f_1 ... f_p$? Explain why they are introduced. (1 point)

(e) How does the logistic function for the Generalized Additive Model (GAM) formulation look like? Write it down. (1 point)

**Problem 5** (Unsupervised) **(10 points)**

Consider the following points that you wish to investigate for possible clusters:

| $i$ | $X_1$ | $X_2$ |
|---|---|---|
| 1 | 1 | 5 |
| 2 | 1 | 4 |
| 3 | 2 | 0 |
| 4 | -1 | 1 |
| 5 | 0 | 2 |

Table 2: Observations for predictor variable $X_1$ and $X_2$ for points $i = 1...5$.

(a) Can the following properties of the $k$-means clustering algorithm be seen as advantages or disadvantages? Explain why.
Properties: (1) algorithm convergence, (2) initialization procedure, (3) (in)sensitivity to outliers, (4) its (in)ability to cluster of new (unseen) samples. (2 points)

(b) Given the points in Table 2. Assume the $k$-means algorithm for $k = 2$ has not yet converged. At some step $i$ you observe the means of cluster 1: $\hat{\mu}_1^{(i)} = (0, 4)$ and cluster 2: $\hat{\mu}_2^{(i)} = (1, 1)$. Perform the next step (one step) of the $k$-means algorithm. Report (i) the cluster assignments (i.e. which points belong to each of the three clusters) and (ii) the coordinates of the new cluster means (i.e. $\hat{\mu}_1^{(i+1)}$ and $\hat{\mu}_2^{(i+1)}$). Make your calculations explicit. (2 points)

(c) Consider Figure 4 (next page), where final cluster assignments and means are indicated using $k$-means clustering. Give an (informal) description of how would you find the decision boundaries between the clusters. (1 point)

(d) What is the difference between $k$-means and $k$-medoids? Given the cluster assignments in Figure 4, what would be the medoids of the respective clusters using the Euclidean distance as dissimilarity metric? (2 points)

(e) To perform $k$-means clustering and $k$-medoids, do we need to know the coordinates of the elements to be clustered, or does it suffice to only know their mutual distances? Explain why. (2 points)

(f) If you are interested in clustering the observed data in terms of the sign of their feature values (not the actual values) - would you still use Euclidean distance as dissimilarity metric? Explain why (not)? If not, what dissimilarity metric would you use for clustering? (1 point)

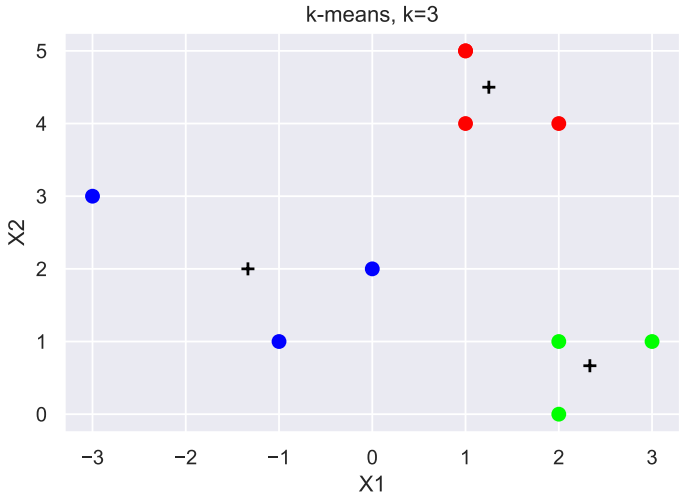


Figure 4: Converged $k$-means clustering with $k = 3$ using data from Table 2