Question Booklet

- **The final exam contains 5 questions and is scheduled for 3 hours. At maximum you can earn 50 points.**
- **Please verify if this question booklet consists of 9 pages, and that all questions are readable, else contact the examiners immediately.**
- **One A4-sized sheet of notes (handwritten on both sides of the sheet) is allowed. No other materials (other notes, books, course materials) or devices (calculator, notebook, tablet, cell phone) are allowed.**
- **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 (Introduction) **(10 points)**

1. Are the following statements correct or incorrect? Explain each answer. (7pts)
   (a) Among all classifiers presented in the lecture, Support Vector Machines always achieve the lowest test error, since it finds the maximum margin classifier. (1pt)
   (b) A feed-forward neural network without an activation function is equivalent to a linear model. (1pt)
   (c) The k-means algorithm is guaranteed to converge to the global optimum. (1pt)
   (d) The t-SNE embedding captures the directions of largest variance in the data. (1pt)
   (e) t-SNE is deterministic, i.e. for the same data it always gives the same result. (1pt)
   (f) k-fold cross-validation with $k < n - 1$ has higher bias than LOOCV. (1pt)
   (g) For data of a large enough sample size n, bootstrap sets that are sampled uniformly at random will be uncorrelated. (1pt)
2. Josephine is analyzing three datasets, but is not entirely happy with the results. Explain for each case how we can address her concerns. (3pts)
   (a) On Dataset 1 of $p = 50$ predictors for a real-valued outcome Y, Josephine applied polynomial regression ($d = 4$). She is concerned the model overfits and that some of the predictors are not relevant for predicting Y. (1pt)
   (b) On Dataset 2 of a single predictor X and real-valued outcome Y, Josephine fits a piecewise polynomial regression spline with $k = 5$ knots. When plotting the result, the fitted function looks discontinuous in the regions around the knots, making Josephine wonder how to make the model more smooth. (1pt)
   (c) On Dataset 3 of $n = 1000$ samples and $p = 100$ variables, Josephine applies k-means clustering to discover k clusters. How can she interpret the results, for example, to verify if the clusters make sense? (1pt)

Problem 2 (Linear Regression) **(10 points)**

1. Data scientists Ali Prediktørson and Omer Régrèssionaire investigate the effects of two predictors X_1, X_2 on a real-valued outcome Y. Using a linear regression model $\hat{f}$, they obtain the following coefficients and standard errors (intercept not shown),

	coefficient β	std. error $SE(\beta)$
X_1	3.31	0.04
X_2	-0.28	0.0012

   (a) What can you say about the relationship between X_1, X_2 and Y based on the coefficients β? (1 pt)

   (b) State the 95% confidence interval for the coefficient of X_1. (1 pt)

   (c) The dataset has $n = 103$ samples, a $TSS = 1500$, and the above model has an $RSS = 600$. Explain step by step how to design a hypothesis test to decide whether *at least one* of the variables X_1, X_2 is relevant for predicting Y. (1 pt)

2. Omer plots the residuals of their model for X_1 in Figure 1.

   (a) Ali argues that the points x_2 and x_3 have unusual values of X_1 compared to the population and suggests removing them from the dataset to fit a more reliable model. Do you agree? Why? (1 pt)

   (b) Ali suggests to do Leave-One-Out Cross-Validation (LOOCV) three times, each time leaving out one of the points x_1, x_2, and x_3. Based on the residual plot, which of these do you expect to result in the highest train, respectively test error? Why? (1 pt)

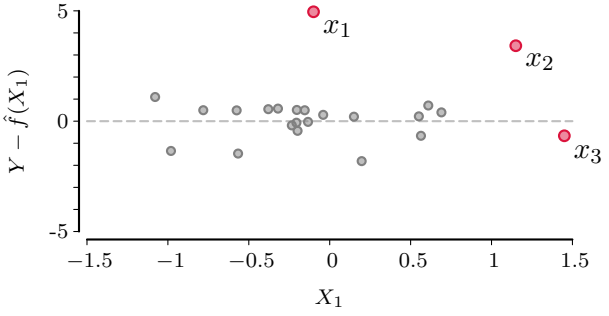
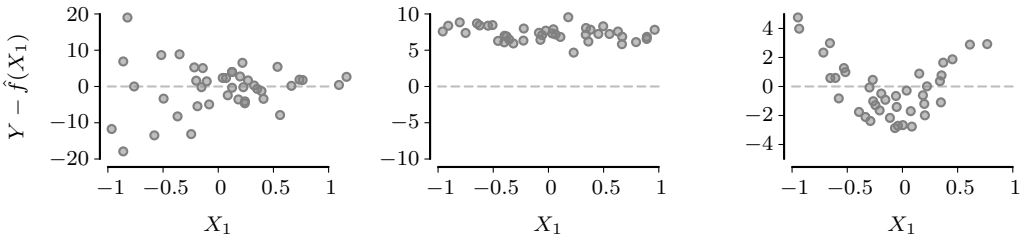


Figure 1: Residual Plot for Problem 2.2



(i) Residuals for D_1. (ii) Residuals for D_2. (iii) Residuals for D_3.

Figure 2: Residual Plots obtained by Filippo for Problem 2.4

3. Filippo Speçulatio is convinced that we should include another predictor X_3. He makes the following claims. For each, determine whether it is true or false, and give a brief explanation why. (2 pts)

   (i) The R^2 score of the model that includes X_3 in addition to X_1, X_2 will always be larger than that over the one with only X_1, X_2.

   (ii) As the predictor X_1 is significant in the model with only X_1, X_2, it will still be significant when we include X_3.

   (iii) We can check which predictors are useful by adding a regularization term.

   (iv) We can check whether the RSS values for any two predictors are correlated with each other to find out whether there is an interaction between two predictors. If so, we should add an interaction term to the model.

4. Filippo applies the linear model from Problem 2.1 to three different datasets, D_1, D_2, and D_3. He forgets to fit the intercept term. He plots the residual plots in Figure 2. (3 pts)

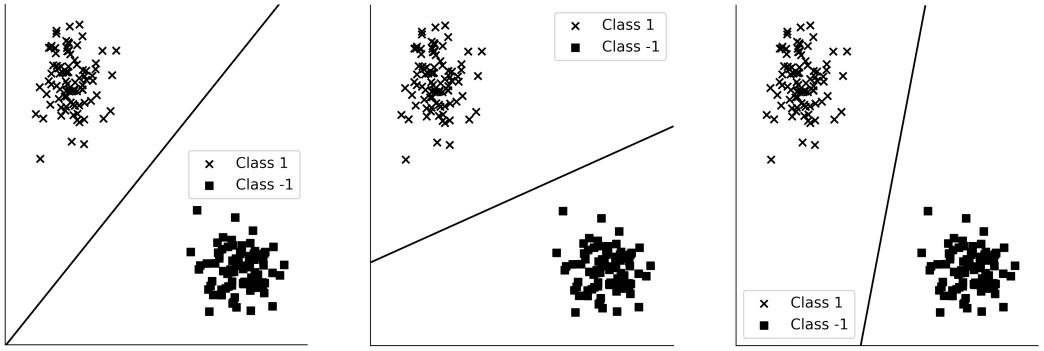
   (a) For each of the three datasets, explain why a linear model is suited or not. ($1\frac{1}{2}$ pts)

   (b) For each of the three datasets, propose an appropriate change to the model to improve the prediction accuracy. Explain your reasoning. ($1\frac{1}{2}$ pts)

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

1. (a) Give the decision boundary of a Maximal Margin Classifier as a function $g(x) = 0$, $x \in \mathbb{R}^p$. What is the geometric interpretation? (1 pt)

   (b) Consider Figure 3. Which of the three plots correspond to the decision boundary of a Soft-Margin Support Vector Classifier with $C = 0$? Explain your answer. (1 pt)



(i) (ii) (iii)

Figure 3: Three decision boundaries

   (c) What is the primary limitation of a hard-margin SVM and how does a soft-margin SVM resolve it? Sketch a dataset that showcases the problem. (2 pt)

2. (a) What problem does Christina face when she would apply Linear Discriminant Analysis (LDA) on the data depicted in Figure 4? Which classifier would you recommend her instead? Why? (1 pt)

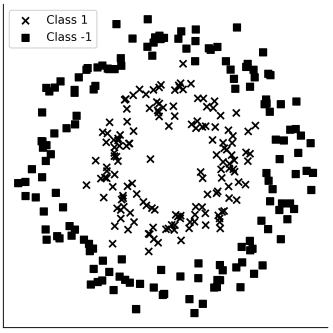


Figure 4: Plot belonging to Problem 3.2.

   (b) In the lecture, we derived LDA using Bayes' rule. Starting from (3 pt)

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)} \,,$$

   we assumed that $f_k(x)$ is a univariate Gaussian with the same variance across all classes.

Christina proposes that we should assume that each class follows a Rayleigh distribution. The density of the Rayleigh distribution is given by

$$f_k(x) = \begin{cases} \frac{x}{\sigma_k^2} e^{\frac{-x^2}{2\sigma_k^2}} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \text{, where } \sigma_k > 0.$$

Derive the discriminant **and** the decision boundary for $x \geq 0$. Is the discriminant linear in x? You can assume that the parameters are chosen such that we do not divide by 0.

3. Muhammed claims that Logistic Regression is a non-linear model and shows Figure 5 as evidence. Is he right? Explain your answer. (2 pt)

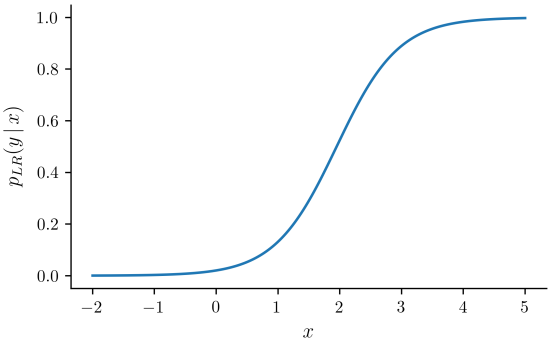


Figure 5: Plot for Problem 3.3

Problem 4 (Unsupervised) **(10 points)**

1. Jawad and Ahmed argue about what is the best laptop. Ahmed claims his is better because it has the new M4 processor, Jawad disagrees and says his is better because it has a touch display.

   As they are data scientists, they agree to settle the matter by examining a dataset $X \in \mathbb{R}^{n \times p}$ of $n = 10\,000$ laptops for which they collect $p = 1\,000$ different features.

   They first want to inspect the data visually. For this, they use PCA to reduce their dataset to $p = 2$ dimensions.

   (a) Briefly describe the general idea behind PCA. (1 pt)

   (b) We know that the principal components of X are the eigenvectors of the covariance matrix $X^T X$. Show that the unit eigenvector w_1 with the largest eigenvalue λ_1, where it holds that (3 pts)

   $$X^T X w_1 = \lambda_1 w_1 \, , \lambda_1 > \lambda_2 > \cdots > \lambda_p,$$

   is the principal component that maximizes the variance of the projected data.

   (c) Another view on PCA is that it tries to minimize the reconstruction error of the data. Explain how reconstruct the projected data back to the original space and what error the first principal component obtains in comparison to the other principal components. (1 pt)

2. After applying PCA, they use k-means to cluster the data into 5 clusters. (1 pt)

   Jawad runs K-means on his Microsoft Surface Pro and gets clustering with an in-cluster variation of 1000. Ahmed runs k-means on his M4 Macbook Pro and gets a clustering with an in-cluster variation of 500. Ahmed argues that his result is better because he is using Apple Silicon. Describe another reason why they could be getting different results that is not related to the processor.

3. Consider the following two initialization strategies for k-means:

   - Randomly pick K points from the dataset as the initial centroids.
   - Randomly assign each point to a cluster and compute the initial centroids as the averages of these clusters.

   (a) For each of the two strategies, where are the initial centroids expected to be on average, and how much variance would the centroids have between different initializations? (2 pts)

   (b) For each of the two strategies, describe an advantage over the other, when we employ them in standard k-means. (2 pts)

Problem 5 (Trees and Splines) **(10 points)**

1. You are given a dataset with points from three different classes and want to classify them based on a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols $[\times, \triangle, \circ]$) and the decision boundaries of a decision tree.

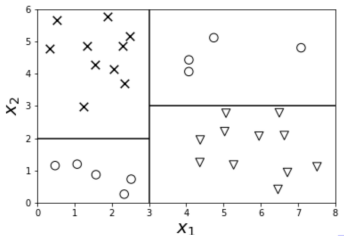


Figure 6: Decision tree Plot belonging to Problem 5.1.

   (a) Draw the corresponding decision tree. Make sure that you include the features (X_1 or X_2) and thresholds of the split. For each node in the tree, also give the number of training samples per class that arrive in that node. (1 pt)

   (b) Compute the Gini index for all internal nodes and all leaves in your decision tree. *Note: Your answer may contain improper fractions (e.g.* $\frac{117}{33}$*).* (2 pt)

2. (a) Assume we have a dataset of two predictive variables X_1 and X_2, with two different classes C_1 and C_2. The points from class C_1 are given by $A = \{(i, i^2) \mid i \in \{1 \dots 100\}\} \subseteq \mathbb{R}^2$, while the points from class C_2 are $B = \{(i, \frac{125}{i}) \mid i \in \{1 \dots 100\}\} \subseteq \mathbb{R}^2$. Construct a decision tree of minimal depth that assigns as many data points as possible to the correct class. Provide for each split the feature and corresponding thresholds. How many and which data points are misclassified? (2 pt)

   (b) Assume we have a dataset D of n samples (x_i, y_i) with $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits.
   Give an example of a small dataset D that contains at least one instance from each class, and for which the learned decision tree has no splits – the root node is a leaf. Justify your answer.
   Hint: you do not need more than a few instances. (1 pt)

3. Assume a linear spline of K knots, i.e. $f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \xi_k)_+$ where $(x_i - \xi_k)_+ = \max(x_i - \xi_k, 0)$ and b_k are the spline coefficients. We aim to minimize the sum of squared residuals

$$\text{minimize} \quad S = \sum_{i=1}^{N} ||y_i - f(x_i)||_2^2 \, .$$

   (a) What is the degree of freedom for this model? Explain your answer. (1 pt)

(b) To avoid overfitting, we usually introduce a penalty on the spline coefficients such as $\sum_{k=1}^{K} b_k^2$. Therefore we minimize a modified objective with regularization parameter λ, (3 pt)

$$\text{minimize} \quad S + \lambda \sum_{k=1}^{K} b_k^2 \, .$$

Derive the closed-form solution of the optimal parameters $\hat{\beta} = [\beta_0, \beta_1 \ldots, \beta_d, b_1 \ldots, b_k]^T$