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 10 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 your reasoning for each answer.

   (a) Using K-mean clustering for a given dataset, there is only one clustering that is the global optimum. (1 pt)

   (b) K-mediods always converges to a local optimum. (1 pt)

   (c) If the data is linearly separable, a hard margin classifier and support vector classifier find the same decision boundary. (1 pt)

   (d) Logistic regression minimizes the negative log-likelihood of the data. (1 pt)

   (e) Ordinary least squares always has a unique solution. (1 pt)

   (f) By Gauss Markov theorem, Ordinary Least Squares will always results in a smaller variance than biased Least Squares. (1 pt)

   (g) For large respectively small enough regularization parameter λ, the solutions of ridge regression and lasso regression will be the same. (1 pt)

   (h) Let $0 <= k < n$. The training error of a degree n polynomial is always strictly smaller than that of a degree k polynomial for the same dataset, when minimizing the Mean Squared Error. (1 pt)

   (i) k-NN is a parametric method because it takes k as parameter. (1 pt)

   (j) In a support vector machine (SVM) with a linear kernel, the decision boundary is always a hyperplane. (1 pt)

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

1. Ali and Omer want to refresh their linear regression skills. They consider a small dataset with a predictor X_1 and outcome Y,

X_1	Y
2	2
4	2
6	4
8	10

   (a) Derive the least squares solution $\widehat{\beta}$. You should use the following approximation (2 pts)

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{10}\begin{bmatrix} 3 & -1 \\ -1 & \frac{1}{10} \end{bmatrix}$$

   (b) Interpret the coefficients $\widehat{\beta}_0, \widehat{\beta}_1$ you obtained. (1 pt)

   (c) Which sample(s) of X_1 have the highest leverage? Does this necessarily suggest removing the sample before fitting a linear model is a good idea? Explain! (1 pt)

2. Filippo expresses doubts about whether the above model is suitable for the dataset.

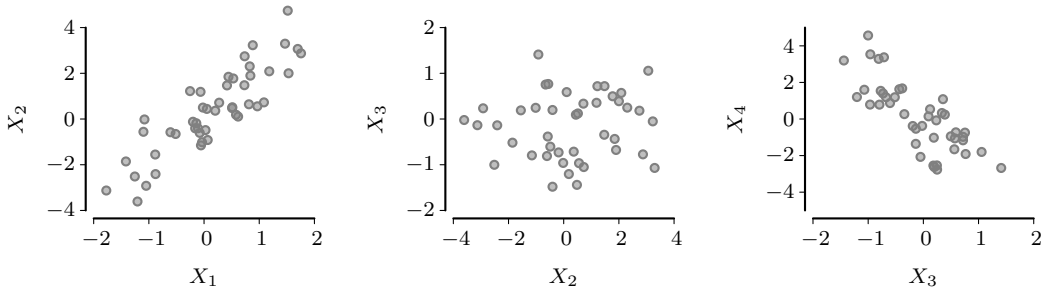
   (a) He computes the standard error of the estimated coefficients as $SE(\widehat{\beta}) = 2.627$. Sketch the distribution of the z-score. Using your drawing, explain how you can tell with 95% confidence that the trend given by $\widehat{\beta}$ holds. (2 pts)

   (b) Explain briefly how Filippo can (2 pts)
   - measure the goodness of fit of the current model,
   - quantify the uncertainty about the estimate $\widehat{\beta}$,
   - find out whether the underlying trend is instead nonlinear,
   - decide whether a given additional predictor X_2 is worth including.

3. Meanwhile, Omer found a larger dataset with multiple predictors X_1, X_2, X_3, X_4 for outcome Y. He plots some pairs of these predictors as shown in Figure 1.

   (a) From the plots, suggest one pair of predictors that would be suitable for linear regression with outcome Y. Explain your reasoning. (1 pt)

   (b) Outline a general approach for finding a subset of useful predictors for a given regression task. (1 pt)



(a) Predictors X_1 and X_2. (b) Predictors X_2 and X_3. (c) Predictors X_3 and X_4.

Figure 1: Pairwise plots of the predictors in Problem 2.3

Problem 3 (classification) **(10 points)**

1. What is the general idea of a Support Vector Classifier (SVC)? (1 pt)
2. Is it possible to apply an SVC (or any binary classifier) to a classification problem with more than two classes? Explain how you would do it or describe why it is not possible. (1 pt)
3. How does the objective of an SVC with a linear kernel relate to that of ridge regression? (1 pt)
4. Given a small dataset shown in Table 1, where X_1 and X_2 denote the coordinates of the data points and Y denotes the labels. Compute the parameters of the maximum margin classifier that perfectly classifies the dataset. (2 pts)
   Hint: Make a sketch and use basic linear algebra.

X_1	X_2	Y
2	3	1
6	-1	-1

Table 1: A simple dataset for Problem 3.2.

5. Figure 2 shows four different classification tasks with balanced classes. Assign each of these classification tasks to one of the following classifiers that is best-suited to solve the respective task: (2 pts)

   - Decision Tree
   - SVC with a RBF Kernel
   - SVC with a linear kernel
   - Linear Discriminant Analysis
   - Logistic Regression
   - Quadratic Discriminant Analysis

   Use each classifier only once. Briefly explain your reasoning.

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

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

   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 the so called YouCanDoIt distribution. The density of the YouCanDoIt distribution is given by

$$f_k(x) = \begin{cases} \frac{\mu_k}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\mu_k)^2}{2x}\right). & , x > 0 \\ 0 & , x \leq 0 \end{cases} \text{, where } \mu_k > 0.$$

   Derive the discriminant **and** the decision boundary for $x > 0$. Remove all terms that are independent of μ_k and π_k. Is the discriminant linear in x? You may assume that the parameters are chosen such that we do not divide by 0.

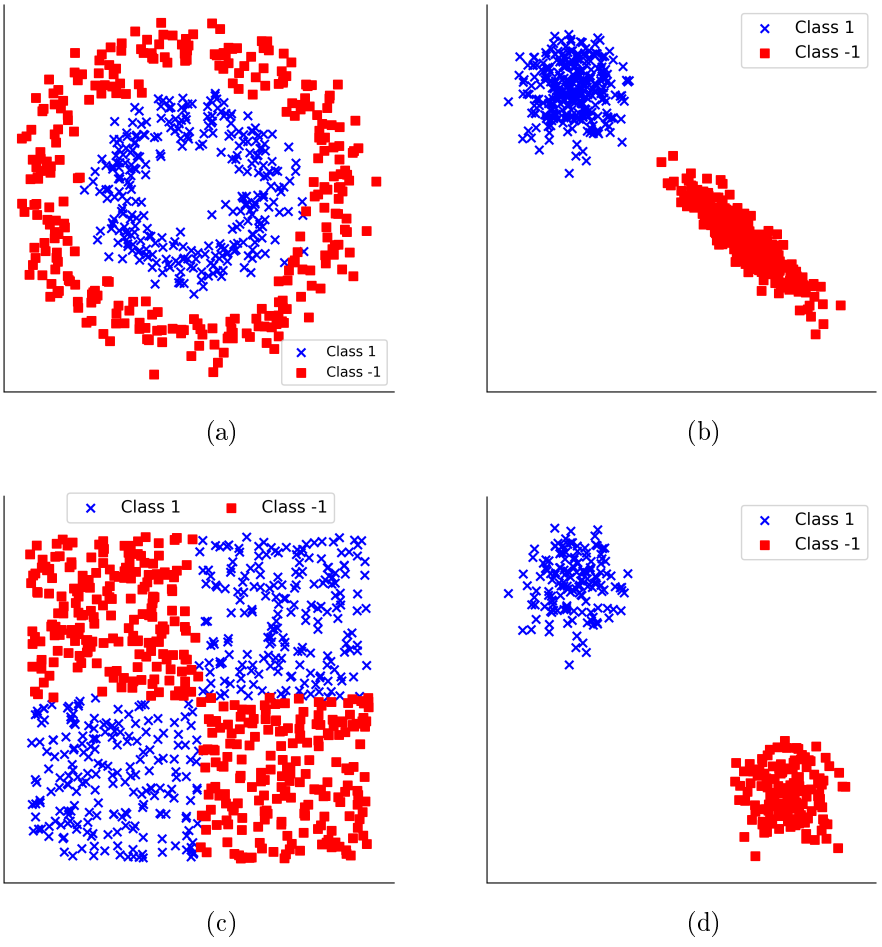


Figure 2: Four classification tasks for Problem 3.5.

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

1. It has been a month and Jawad and Ahmed are still debating about which laptop is best for their machine learning projects. Using the same dataset $X \in \mathbb{R}^{n \times p}$, with $n = 10{,}000$ datapoints and $p = 1000$ features, they want to inspect the data visually again. This time they decide to use t-SNE.

   (a) Describe one advantage of t-SNE over PCA. (1 pt)

   (b) The perplexity parameter in t-SNE controls the effective number of neighbors in the original space.

      i. Give the mathematical definition of perplexity in this context and explain how it is achieved in practice. (1 pt)

      ii. What impact does a low/high perplexity value have on the resulting visualization? (1 pt)

   (c) What is the crowding problem that stochastic neighbor embedding faces. How does the t-distributed variant, t-SNE, mitigate it? (2 pts)

2. With t-SNE, Jawad and Ahmed obtain the data visualization shown in Figure 3. They both agree that the data is best represented by two clusters. To confirm, they run hierarchical clustering, on the dimensionality reduced dataset, using single linkage and complete linkage. Both methods produce the same clustering for the top level (level with two clusters).

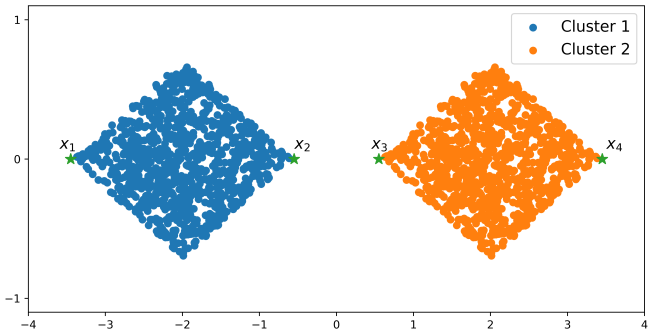


Figure 3: The dimensionality reduced dataset for Task 4.2a and 4.2b. Single linkage and complete linkage hierarchical clustering produce the same clustering.

   (a) Consider the top level of the single linkage and complete linkage hierarchical clustering. Which pairing of the points x_1, x_2, x_3, x_4 marked in Fig. 3 is used to determine the distance between the two clusters for each linkage method? (1 pt)

   (b) In the dataset of Figure 3, single linkage and complete linkage hierarchical clustering are able to recover the true underlying clusters. Assume there are now 1000 evenly spaced points between x_2 and x_3, which hierarchical clustering method would you use and why? (1 pt)

(c) Another commonly used linkage function is the group average linkage, that is the average distance between all pairs of points in the two clusters.

i. Given a pairwise distance matrix D, where D_{ij} is the distance between points x_i and x_j, and two clusters $G = \{x_i\}$ and $H = \{x_j\}$, give the formula to compute the distance between the two clusters $d(G, H)$. (1 pt)

ii. After merging two clusters G and H, we need to compute the distances of the new cluster $G \cup H$ to all other clusters K. Instead of doing this from scratch, show how we can instead use the previous distances $d(G, K)$ and $d(H, K)$ to compute the new distance $d(G \cup H, K)$. (2 pts)

Problem 5 (Trees and Model Selection) **(10 points)**

1. Consider the regression tree shown in Fig. 4. Sketch what the partition space for this tree would look like. (1 pt)

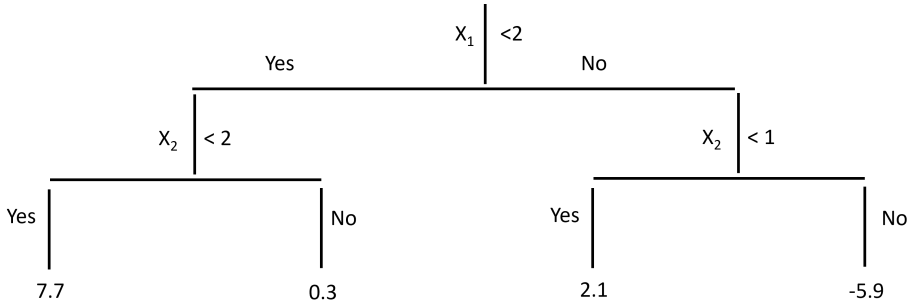


Figure 4: Regression tree for Question 5.1

2. Consider the four partition spaces for regression trees shown in Fig. 5. For each of the partition spaces, state why it is (not) possible to achieve this partition using the regression trees we have learned in this course. Give a short reason to support each of your answer. (2 pts)

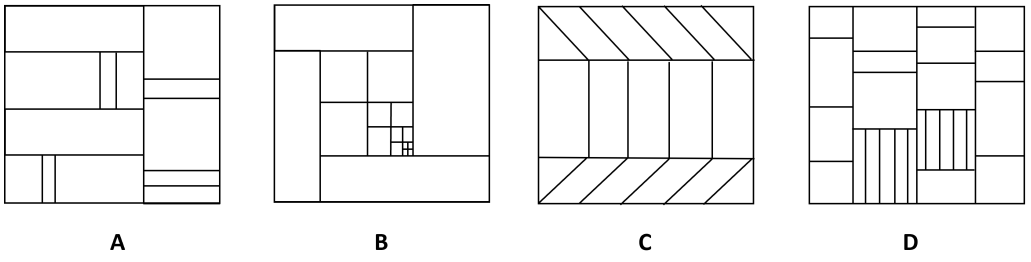


Figure 5: Predictor spaces for Question 5.2

3. Joscha and Osman want to learn tree-based classifier. Osman suggests they build a single decision tree as it has low bias. Joscha points out that a single decision tree will have high variance, therefore they should depth limit the tree to reduce variance. This makes Osman unhappy as depth-limited tree will have higher bias. Things are about to get heated before Nils intervenes and mediates a solution where they use bagging.

   (a) Explain how bagging works. (1 pt)

   (b) Explain one limitation of bagging and describe how you could overcome it. (1 pt)

4. Consider a dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, with $x^{(i)} \in \mathbb{R}^D$, $y^{(i)} \in \mathbb{R}$ and centered features so that $\sum_{i=1}^{N} x^{(i)} = 0$. There is one outlier in the dataset.

   (a) If we perform bootstrapping where each k subset of our dataset is of size N. What is the probability that at least one of them has an outlier? (1 pt)

   (b) Explain how this outlier affects scores of LOOCV. (1 pt)

   (c) Describe one alternate strategy that will not have the problem you described for LOOCV. (1 pt)

5. Assume that we are given a dataset, where each sample x_i and regression target y_i is generated according to the following process (2 pts)

$$x_i \sim \text{Uniform}(-10, 10)$$

$$y_i = ax_i^3 + bx_i^2 + cx_i + d + e_i, \quad \text{where} \quad e_i \sim \mathcal{N}(0, 1) \quad \text{and} \quad a, b, c, d \in \mathbb{R}.$$

   The regression algorithms below are applied to the given data. Describe what the bias and variance of these models are (low or high) with respect to the equation given above. Provide a 1-2 sentence explanation to each of your answers.

   (a) Linear Regression.

   (b) Polynomial regression with degree 3.

   (c) K-Nearest Neighbor Regression with $K = 1$.

   (d) K-Nearest Neighbor Regression with $K = n$ where n is the number of samples in dataset.