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)

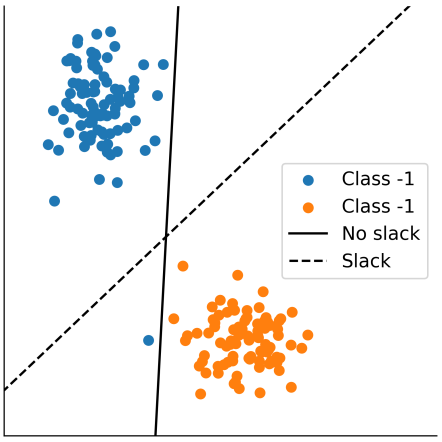


Figure 1: Solution to P1.c.

Solution.

1. (a) False. There can be multiple clustering that have the same globally optimal objective value. Consider a dataset with the four corner points of a square and $k = 2$. Then both solutions on opposing sides of the square are optimal for example.
   (b) True. The same convergence guarantee for K-means also hold for K-medians.
   (c) False. See Figure 1.
   (d) Yes, since minimizing the negative log-likelihood is equivalent to maximizing the likelihood.
   (e) False, OLS may have multiple solutions if the predictors are linearly dependent.
   (f) False, the Gauss Markov theorem addresses unbiased estimators and thus does not address a biased version such as regularized least squares.
   (g) True, for $\lambda = 0$, both are equivalent to OLS. For large λ the coefficients will be close to zero but not necessarily equal, but in the limit will approach zero.
   (h) False, the error is always smaller **or equal**, e.g. when MSE of $\hat{f}_k$is 0 then the MSE of $\hat{f}_n$ can not be smaller than 0.
   (i) False, k-NN is a non-parametric method. While it does take k as a parameter, it doesn't make assumptions about the underlying data distribution based on this parameter.
   (j) True, For a non linear decision boundary a non linear kernel has to be used.

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.

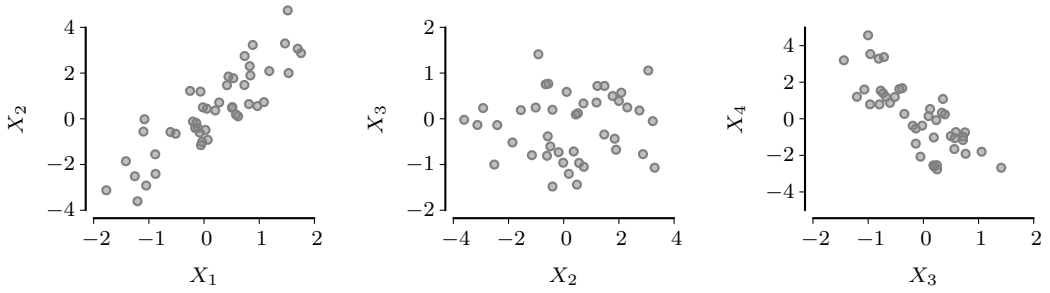
   (a) He computes the standard error of the estimated coefficients as $SE(\widehat{\beta}) = 2.627$. Sketch the distribution of the z-score. Using your drawing, explain how you can tell with 95% confidence that the trend given by $\widehat{\beta}$ holds. (2 pts)

   (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 2.

   (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 2: Pairwise plots of the predictors in Problem 2.3

Solution.

1. (a) Using the given hat matrix $\mathbf{P}$ as well as the matrix containing columns for X_1 and the unit vector,

$$\mathbf{P} = \frac{1}{10}\begin{bmatrix} 3 & -1 \\ -1 & \frac{1}{10} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}$$

we have

$$\mathbf{P}\mathbf{X}^T = -\frac{1}{10}\begin{bmatrix} -1 & 1 & 3 & 5 \\ 0.8 & 0.6 & 0.4 & 0.2 \end{bmatrix}.$$

This results in the estimates

$$\mathbf{P}\mathbf{X}^T Y = -\frac{1}{10}\begin{bmatrix} 62 & \frac{32}{5} \end{bmatrix} = \begin{bmatrix} -6.2 & -0.64 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 & \widehat{\beta}_1 \end{bmatrix}.$$

Note: The given $\mathbf{P}$ did not match the data X by mistake. For the given data, we have

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{4}\begin{bmatrix} 6 & -1 \\ -1 & \frac{1}{5} \end{bmatrix}$$

resulting in

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{4}\begin{bmatrix} 4 & 2 & 0 & -2 \\ -0.6 & -0.2 & 0.2 & 0.6 \end{bmatrix}$$

and in the coefficients

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \begin{bmatrix} -2 & 1.3 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 & \widehat{\beta}_1 \end{bmatrix}.$$

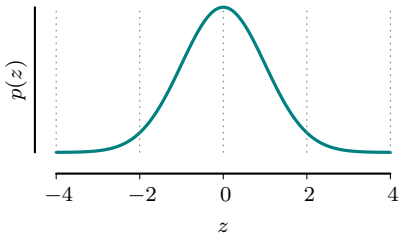
(b) We have that X_1 is negatively correlated with the outcome with intercept $\widehat{\beta}_0$.

(c) We obtain the leverage for a sample i using the row $X_1^{(i)}$ using

$$h_i = X_1^{(i)T} \mathbf{P} X_1^{(i)}$$

thus we get $h_1 = -0.06, h_2 = -0.34, h_3 = -0.54$ and $h_3 = -0.66$ *Note:* For the original data, we get $h_i = 0.3$ for $i = 2, 3$ and $h_i = 0.7$ for $i = 1, 4$.
A high leverage only means that the samples have high impact on the regression line but does not take into account the relationship of X_1 to the outcome Y, hence does not necessarily mean that we should remove them.

2. (a) The z-score $z = \frac{\widehat{\beta}}{SE(\widehat{\beta})}$ follows a normal distribution as sketched below,



(b)
- To judge the goodness of fit, direct measures such as (adjusted) R^2 are most commonly used. Model validation approaches such as k-fold or LOO-cross validation are also an option.
- To quantify the uncertainty of coefficients $\widehat{\beta}$ in linear regression, we can compute a confidence interval for $\widehat{\beta}$ from its standard error SE $(\widehat{\beta})$. Another approach is bootstrapping, where we resample repeatedly from the training set, fit a linear regression for each subsample, and obtain a histogram over the resulting values $\widehat{\beta}$. The the resulting empirical distribution over $\widehat{\beta}$ can also give insight into the uncertainty over $\widehat{\beta}$.
- Inspecting the residual plot can show nonlinearity as the errors will take on a nonlinear, U-shaped curve in that case. Note that directly comparing the RSS of a linear and nonlinear model may be misleading due to overfitting.
- We can use hypothesis testing to test whether the given additional predictor X_2 is useful or uninformative. To this end, the null hypothesis is $H_0 : \beta_2 = 0$, where β_2 is the OLS coefficient for X_2. The corresponding $F-$statistic is

$$\frac{\text{RSS}_1 - \text{RSS}_{12}}{\text{RSS}_{12}/n - 2},$$

using the residual squared errors of the model including X_2, RSS_{12}, compared to the one excluding it, RSS_1.

3.
- The predictors X_2 and X_3 appear to be uncorrelated, whereas X_1, X_2 are positively, X_1, X_2 negatively correlated. To meet the independence assumption in OLS, we should use predictors X_2 and X_3 (or X_2 and X_4 which are not shown).
- There are different approaches to doing so, such as best subset selection. Other options include shrinkage, that is, using Lasso regularization to encourage sparse models, or resampling or subsampling strategies.

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 3 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.

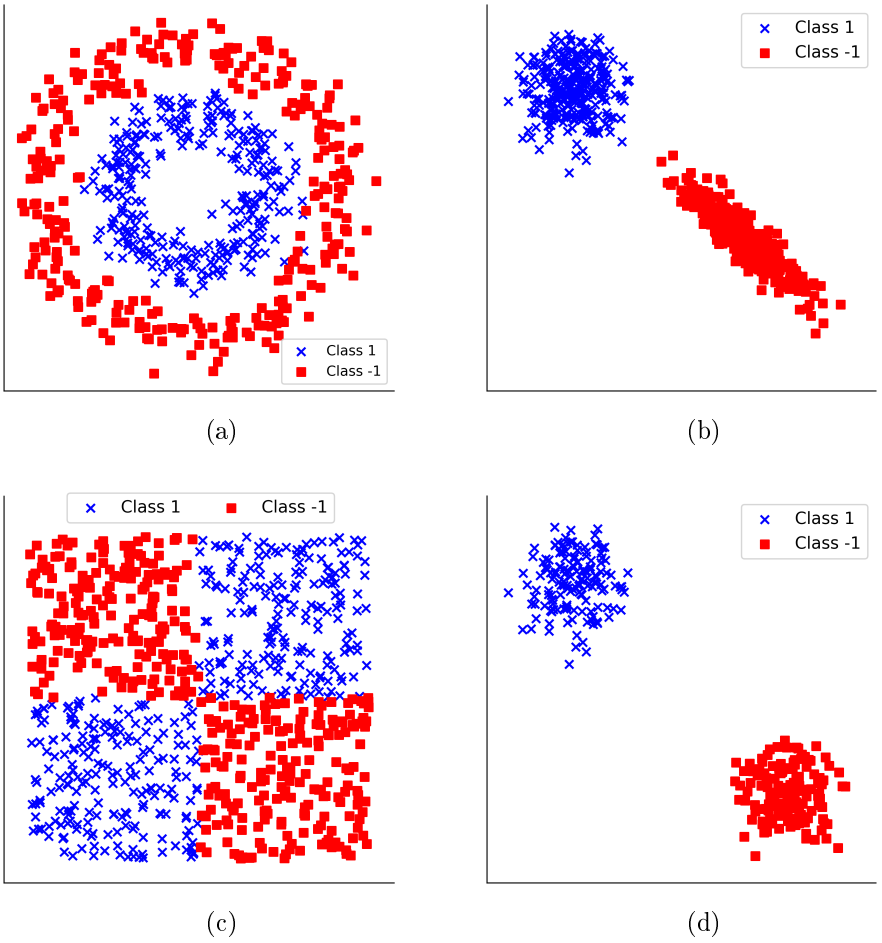


Figure 3: Four classification tasks for Problem 3.5.

Solution.

1. Support Vector Classifiers (SVCs) are a type of supervised machine learning algorithm used for classification. The main idea behind SVM is to find a hyperplane that best separates different classes in the feature space. In the context of binary classification, this hyperplane aims to maximize the margin between the two classes.

2. Yes, it is possible to apply Support Vector Classification (SVC) to a classification problem with more than two classes. The common approaches are one-vs-one (OvO) and one-vs-all (OvA).

   - **One-vs-One (OvO)**:
     For N classes, this strategy trains N * (N - 1) / 2 binary classifiers, each one distinguishing between two classes. During prediction, each classifier makes a prediction, and the class that wins the most pairwise competitions is chosen as the final predicted class.

- **One-vs-All (OvA or One-vs-Rest)**:
  For N classes, this strategy trains N binary classifiers, each one distinguishing between one class and the rest. During prediction, the classifier that assigns the highest confidence or score is selected as the predicted class.

3. The objective of SVM with a linear kernel involves minimizing the hinge loss function, which measures the classification error, along with a regularization term. Ridge Regression aims to minimize the residual sum of squares (RSS) between the predicted and actual values of the target variable. When a linear kernel is used in SVC, the optimization problem shares similarities with Ridge Regression, as both involve the minimization of a cost function that includes a term penalizing the squared magnitude of the coefficients.

4. Let $A = \begin{pmatrix}6 & -1\end{pmatrix}^T$ and $B = \begin{pmatrix}2 & 3\end{pmatrix}^T$. Then $\hat{w} = \overrightarrow{AB} = B - A = \begin{pmatrix}-4 & 4\end{pmatrix}^T$. Then $m_{AB} = A + 0.5\,\hat{w} = \begin{pmatrix}4 & 1\end{pmatrix}^T$. The hyperplane that perfectly classifies the data is given by:

$$\begin{pmatrix}-4\\4\end{pmatrix}\left[x - \begin{pmatrix}4\\1\end{pmatrix}\right] = 0$$

   The maximum margin classifier requires that $||\hat{w}|| = 1$. So, normalizing $\hat{w}$ yields $w = \begin{pmatrix}-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}^T$. Then the maximum margin classifier is given by,

$$f(x) = \text{sign}\left(\begin{pmatrix}-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{pmatrix}x + \frac{3}{\sqrt{2}}\right)$$

   To verify that the classifer indeed correctly classifies the data, we compute $f(A) = \text{sign}(-\frac{7}{\sqrt{2}} + \frac{3}{\sqrt{2}}) = -1$ and $f(B) = \text{sign}(-\frac{1}{\sqrt{2}} + \frac{3}{\sqrt{2}}) = 1$.

5. - SVM with a RBF kernel: Smooth non-linear decision boundary
   - SVC with a linear kernel: Can be well-seperated by a hyperplane, however class -1 is ansitropic, thus SVC with a linear kernel is best suited due to maximum margin decision boundary.
   - Decision Tree: Can be easily classified with axis-aligned cuts.
   - LDA, LR: Can be clearly classified by a hyperplane.

6. - Discriminant:

$$\begin{aligned}
&\ln\left(\pi_k \frac{\mu_k}{\sqrt{2\pi x^3}} \exp\left(-\frac{(x-\mu_k)^2}{2x}\right)\right)\\
&= \ln(\pi_k) + \ln(\mu_k) - \ln(2\pi x^3)^{1/2} - \frac{(x-\mu_k)^2}{2x}\\
&= \ln(\pi_k) + \ln(\mu_k) - \frac{x^2 - 2\mu_k x + \mu_k^2}{2x} - \frac{1}{2}\ln(2\pi x^3)\\
&= \ln(\pi_k) + \ln(\mu_k) + \mu_k - \frac{\mu_k^2}{2x} - \frac{x}{2} - \frac{1}{2}\ln(2\pi x^3)
\end{aligned}$$

   $\delta_k(x) = \ln(\pi_k) + \ln(\mu_k) + \mu_k - \frac{\mu_k^2}{2x} \rightarrow$ non-linear due to the last term

- Decision boundary:

$$\begin{aligned}
\ln(\pi_k) + \ln(\mu_k) + \mu_k - \frac{\mu_k^2}{2x} &= \ln(\pi_l) + \ln(\mu_l) + \mu_l - \frac{\mu_l^2}{2x} \\
\frac{\mu_l^2}{2x} - \frac{\mu_k^2}{2x} &= \ln(\pi_l) - \ln(\pi_k) + \ln(\mu_l) - \ln(\mu_k) + \mu_l - \mu_k \\
x^{-1} &= (\ln(\pi_l) - \ln(\pi_k) + \ln(\mu_l) - \ln(\mu_k) + \mu_l - \mu_k)(\frac{\mu_l^2}{2} - \frac{\mu_k^2}{2})^{-1} \\
x &= (\ln(\pi_l) - \ln(\pi_k) + \ln(\mu_l) - \ln(\mu_k) + \mu_l - \mu_k)^{-1}(\frac{\mu_l^2}{2} - \frac{\mu_k^2}{2})
\end{aligned}$$

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 4. 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).

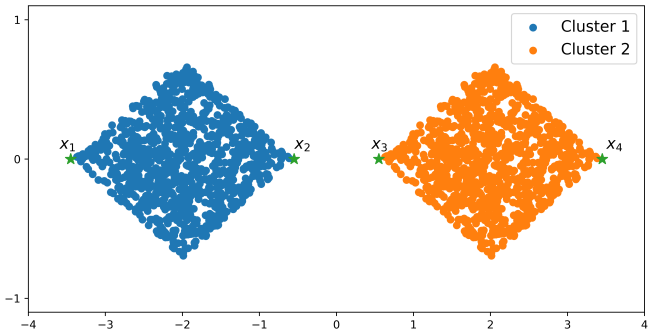


Figure 4: The dimensionality reduced dataset for Task 4.2a and 4.2b. Single linkage and complete linkage hierarchical clustering produce the same clustering.

   (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. 4 is used to determine the distance between the two clusters for each linkage method? (1 pt)

   (b) In the dataset of Figure 4, 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)

Solution.

1. (a) t-SNE tries to preserve the local structure of the data, and is also more resistant to noise due to the stochastic nature of the optimization. It is not restricted to linear transformations, and can capture non-linear structures in the data.

   (b) i. The perplexity is defined as the effective number of neighbors in the original space, and is given by

   $$\text{Perplexity}(p) = 2^{H(P_i)}$$

   where $H(P_i)$ is the Shannon entropy of the conditional distribution P_i. In practice, the perplexity is achieved by finding the value of the bandwidth of the gaussian kernel σ_i that results in the desired perplexity.

   ii. The higher the perplexity, the more non-local information is preserved, leading to a tendency to form larger clusters, and uncover global structure. A lower perplexity value focuses more on local structures being preserved, and tends to form smaller clusters.

   (c) The crowding problem is when the far distances between the embedded points are not preserved, and t-SNE tries to mitigate it by using a t-distributed variant that uses a heavy-tailed distribution to model the distances between the embedded points.

2. (a) For single linkage, the distance between x_2 and x_3 is the shortest distance between the two clusters. For complete linkage, the largest distance counts, hence it is the distance between x_1 and x_4.

   (b) A problem of single linkage is given by the so called chaining phenomenon. That is, as only the closest distance is relevant, any noise or outlier can connect two clusters. In this case, the 1000 evenly spaced points would connect the two clusters, and could hence lead us to obtain arbitrary clusters because now x_2 and x_3 are no longer the points which are merged last. Complete linkage avoids this problem, as the requirement of the largest distance between the clusters makes it prefer spherical clusters.

   (c) i. The average linkage is defined as

   $$d(G,H) = \frac{1}{|G||H|} \sum_{x_i \in G} \sum_{x_j \in H} D_{ij} \ .$$

   ii. The new distance $d(G \cup H, K)$ is defined as

   $$d(G \cup H, K) = \frac{1}{(|G| + |H|)|K|} \sum_{x_i \in G \cup H} \sum_{x_j \in K} D_{ij} \ .$$

   We can separate the sum into the sum over G and the sum over H, and use the previous distances to obtain

   $$d(G \cup H, K) = \frac{1}{(|G| + |H|)|K|} \sum_{x_i \in G} \sum_{x_j \in K} D_{ij} + \frac{1}{(|G| + |H|)|K|} \sum_{x_i \in H} \sum_{x_j \in K} D_{ij} \ .$$

Using the previous distances $d(G,K) = \frac{1}{|G||K|}\sum_{x_i \in G}\sum_{x_j \in K} D_{ij}$ and $d(H,K) = \frac{1}{|H||K|}\sum_{x_i \in H}\sum_{x_j \in K} D_{ij}$, we can write

$$d(G \cup H, K) = \frac{|G|}{|G|+|H|}d(G,K) + \frac{|H|}{|G|+|H|}d(H,K)\ .$$

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

1. Consider the regression tree shown in Fig. 5. Sketch what the partition space for this tree would look like. (1 pt)

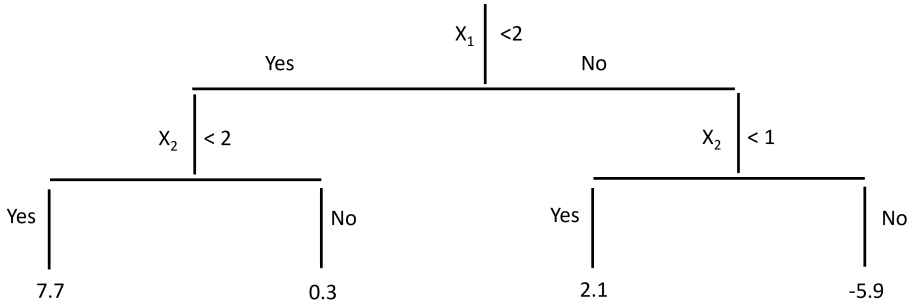


Figure 5: Regression tree for Question 5.1

2. Consider the four partition spaces for regression trees shown in Fig. 6. 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)

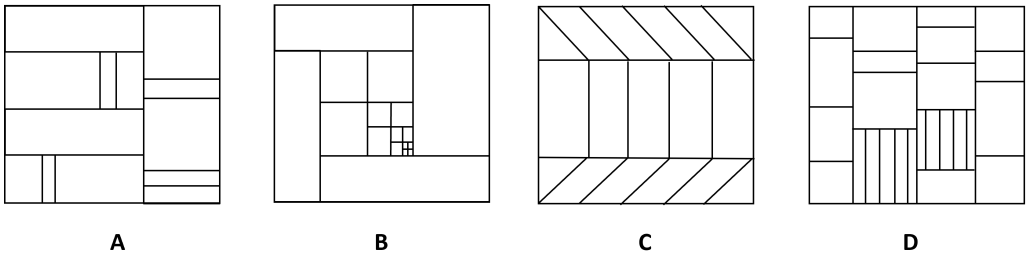


Figure 6: Predictor spaces for Question 5.2

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.

Solution.

1. The solution is the predictor space shown in Fig. 7. (1 pt)

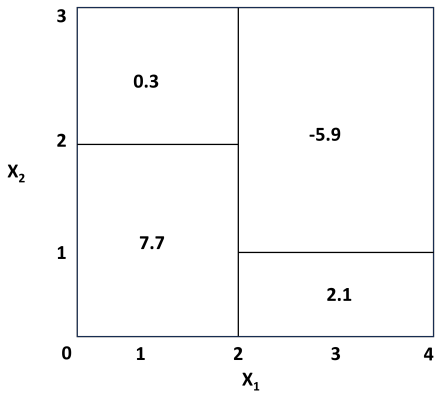


Figure 7: Predictor space for tree in Question 5.1

2. Partitions A and D can be modelled using the axis-aligned decision trees that we have learned in this course. Partition B would require modeling predictor spaces as rectangles. Partition C requires modeling splits as straight non-axis aligned lines due to the partition on its bottom right corner.

3. Bagging

   (a) Lecture 11, Slide 26

   (b) Bagged trees have a problem. Because bootstrap samples have a large overlap, bagged trees are highly correlated and the decrease of variance by averaging over them is thus not as large as desired. To over come this, instead of the full set of predictors, choose the best predictor out of a random sample of m predictors and train a tree using only those predictors. This is is the main idea behind the Random Forest predictor. (Lecture 11, Slide 27)

4. LOOCV

   (a) The probability that a subset does not have an outlier is $(1 - \frac{1}{N})^N$ then the probability that at least one of the k copies has an outlier is $1 - (1 - \frac{1}{N})^{Nk}$

   (b) LOOCV, when evaluated on outlier, results in a higher variance in its scores.

   (c) k-fold cross-validation ensures that the model is evaluated on other data points along with outlier, depending on the right choice of k. Thus k-fold cross-validation will have lower variance than LOOCV in this case. Alternate correct answers: train/test splits.

5. Bias-Variance

   (a) Bias: high. Variance: low.
   A straight line cannot capture a degree 3 polynomial (underfitting).

   (b) Bias: low. Variance: low.
   The model is same as the data-generating process. We can achieve a good fit.

   (c) Bias: low. Variance: high. The model will only the closest point to predict the value of a new point resulting in low bias, however adding new data points to the current dataset will result in high variance in prediction.

   (d) Bias: high. Variance: low. This model will underfit very strongly since the final predicted value will always be the average value of the target variable over dataset. This means that while the variance in prediction will be low, the bias towards the target mean would be very high.