Question Booklet

- **The final exam contains 5 questions and is scheduled for 2.5 hours. At maximum you can earn 50 points.**
- **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 (Linear regression and Regularization) **(10 points)**

1. Analyze, using linear regression, the data from the following table.

X_1	2	2.3	2.4	2.6	2.8	3
Y	14	14.6	14.8	15.2	15.6	16

Recall that linear regression takes the form $Y = \beta_0 + \beta_1 X_1$, but that it is often convenient to formulate it as $\mathbf{X}\beta = \mathbf{Y}$ with $\beta = [\beta_0\ \beta_1]^\top$ and $\mathbf{X} = [\mathbb{1}\ \mathbf{X_1}]$.

(a) Estimate the coefficients β_0 and β_1 using the following result.

$$(\mathbf{X^T}\ \mathbf{X})^{-1} = \begin{bmatrix} 9.93 & -3.88 \\ -3.88 & 1.54 \end{bmatrix} .$$

Explain and *justify* each step you take. (2 points)

(b) Based on the value of β_1: What can you say of the relationship between X_1 and Y? (1 point)

(c) If we know that the data was generated as $Y = \beta_1 X_1 + \beta_0 + \epsilon$, with $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = 0.5$, can you tell, 95%-confidently, that the trend given in (b) holds? Why? (2 points)

2. Assume that we have a one-dimensional dataset for which we perform ridge regression, where we fix $\beta_0 = 0$. That is, we solve the following optimization problem.

$$\min_{\beta_1} \sum_{n=1}^{N} (y_n - x_n\beta_1)^2 + \lambda\beta_1^2 .$$

(a) Derive, step by step, a closed-form solution for β_1. Make sure that the obtained expression is a minimizer of the optimization problem. (3 points)

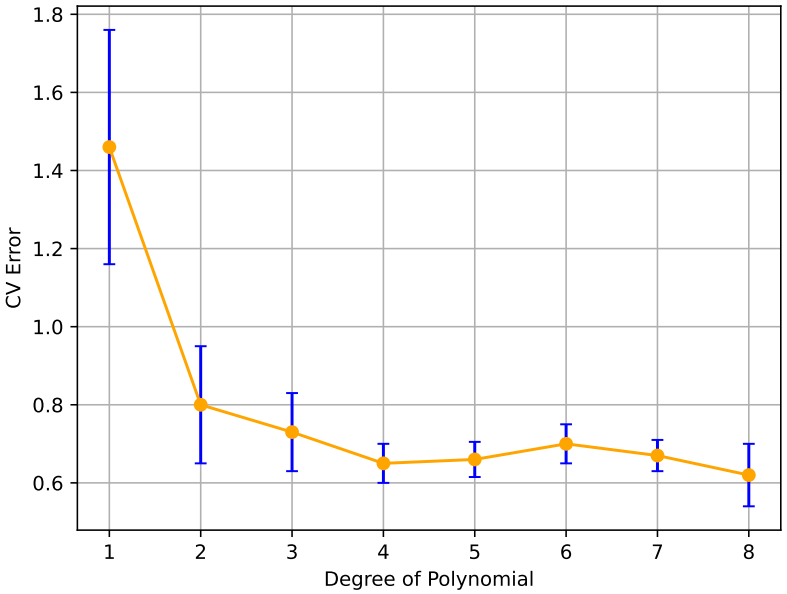
(b) Intuitively, what effect does λ have on the *bias* of the estimate of β_1? And on its *variance*? (2 points)

Problem 2 (Non-Linear regression and errors) **(10 points)**

1. We have been requested to build a model that can properly model some complex data. To this end, we have decided to use polynomial regression.

   (a) Describe in your own words the main idea behind polynomial regression. What is its main advantage over linear regression? How can we estimate its parameters? (2 points)

   (b) Given the cross-validation error as a function of the degree of the polynomial in the figure below (mean in yellow, standard deviation in blue): Which polynomial degree would you use? Justify your answer. (1 point)



2. We want to use a tree-based approach to learn data such as the one shown in the image below. For this task, we consider a simple regression tree, where the tree is created following the greedy *recursive binary splitting* algorithm seen in the lectures. Recall that, at each step, the algorithm chooses the predictor X_j and cut point s, creating two new regions $R_1(j,s) = \{X|X_j < s\}$ and $R_2(j,s) = \{X|X_j \geq s\}$ solving

$$\min_{j,s} \sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{2.1}$$

where $\hat{y}_{R_k}$ denotes the mean response for the training observations in $R_k(j,s)$.

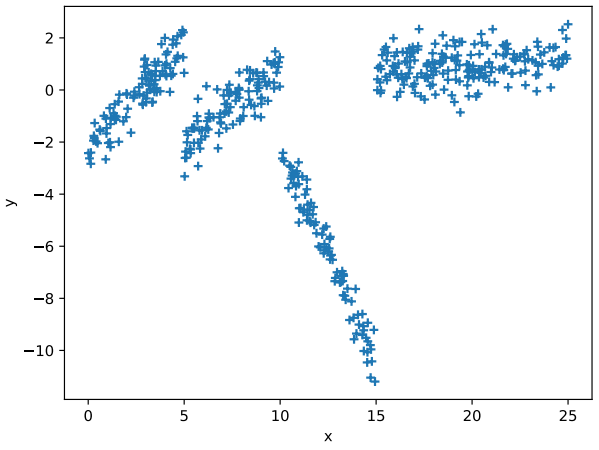
(a) Is such a tree a good model for the data shown below? Why (not)? (2 points)

(b) How would you reduce the training error to 0? And, in contrast, how would you avoid this from happening? Why would you be interested in avoiding zero training error? (2 points)

We decide to slightly change our approach and use, instead, a Linear Model Tree. This model differs from a standard regression tree in that $\hat{y}_{R_k}$ is replaced (in Eq. 2.1) by the prediction of a linear model fitted using the data from that region, that is, $\hat{y}_{R_k}$ is replaced by $\hat{y}_{i,R_k} = x_i a_{R_k} + b_{R_k}$.

(c) Is this a better model for our data? Why (not)? (1 point)

(d) Explain the steps to predict the response for a new data point. (2 points)



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

Our highly competent research team is dealing with a classification problem in which they want to predict the type of monkey from an NFT, out of K different monkey classes. However, the GPUs are broken, and their non-deep-learning skills are a bit rusty. As an external advisor, they demand your expertise in linear classifiers to make a decision on which model to use.

Ian Badfellow seeks for perfection, and claims that they should use the *ideal* classifier.

(a) Explain him in your own words what is the Bayes Classifier, in which sense it is ideal, and why we do not use it in practice so often. (2 points)

Jürgen Schüber, who complains that Bayes did not properly cite the work of Leibniz, claims that we need some assumptions if we want to solve the problem. To this end, he kindly reminds you that according to Bayes's theorem for each class k we have

$$\Pr(Y = k \mid X = x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l \cdot f_l(x)} ,$$

where $f_k(x)$ denotes the density function of X for the k-th class, $\Pr(X = x \mid Y = k)$.

(b) What is π_k in the above expression? How would you estimate π_k for a given dataset? (1 point)

(c) Assume that f_k is provided. How can you use Bayes' theorem to predict the class label of a new data point? (1 point)

Unfortunately, f_k is unknown in practice, and our experts really need your help to decide.

(d) Give the specific form of f_k, and list the extra assumptions made for: (2 points)

i) Multinomial Logistic Regression.
ii) Linear Discriminant Analysis.
iii) Quadratic Discriminant Analysis.

(e) Assume $K = 2$. What is the odds ratio? And the discriminant function? How do they relate? (2 points)

José Bengio is tired of talking and wants to step in. To keep funding coming, the team is interested in knowing whether the monkey is valuable ($Y = k$) or not ($Y \neq k$) (i.e., binary classification).

(f) In order to assess the best method, José wants to compute the ROC curve for all the previous methods. What are the axes of the plot, and how do you generate the curve (that is, which value would you change to generate the curve)? (2 points)

Problem 4 (Non-linear classification) **(10 points)**

1. The greedy algorithm seen in the course to build classification trees does not allow for partitions such as the one in Fig 2.

   (a) Draw the partition produced by the decision tree shown in Fig 1. Is the partition unique? Why? (2 points)

   (b) Why cannot we build a tree that produces the partition in Fig. 2? How would you change the model to allow such a partitioning? You do not have to explain how the changed model is fitted. (2 points)

   (c) Is the misclassification error a good loss function to generate a tree? Why? Justify your answer with an example. (2 points)

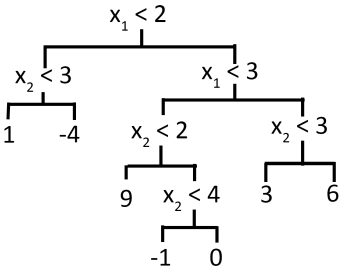


Fig 1. Decision tree for exercise (a).

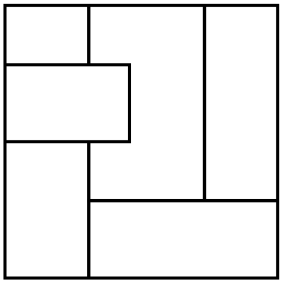


Fig 2. Partition for exercise (b).

2. Recall that the Support Vector Machine is defined as follows:

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

   (d) How does an SVM differ from a maximal margin classifier? (1 point)

   (e) Explain the purpose of the variable C in the optimization problem above. (1 point)

   (f) How does the kernel trick help an SVM classify non-linearly related data? What is the main advantage of this approach? (2 points)

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

1. Given the following set of points:

$$\mathbf{x}_1 = (7, \ 0); \quad \mathbf{x}_2 = (5, -3); \quad \mathbf{x}_3 = (1, \ 6);$$
$$\mathbf{x}_4 = (6, -1); \quad \mathbf{x}_5 = (5, \ 3); \quad \mathbf{x}_6 = (2, -3);$$

   Compute two full iterations of k-means clustering (Lloyd's algorithm) with initial clusters $\mu_1 = (-1, 2)$ and $\mu_2 = (3, 5)$. Use $d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$ as distance. Make sure to write down the necessary distances, explain the steps you follow, and to describe the resulting clusters (centroid and points) *at the end of both iterations*. (3 points)

2. Draw the dendrogram for the following dataset, using single linkage hierarchical clustering with the Manhattan distance, $d(\boldsymbol{a}, \boldsymbol{b}) = |a_1 - b_1| + |a_2 - b_2|$. Make sure to indicate the distances, the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram. (3 points)

Name	X_1	X_2
A	5	2
B	3.5	1
C	−3	2
D	2	4
E	7	−3
F	3	3.5

3. Suppose we have a dataset which has too many features, and thus we wish to perform dimensionality reduction on the dataset by applying PCA:

   (a) Does PCA perform feature selection? Why (not)? (2 points)

   (b) Say we use PCA to reduce the data dimensionality to a single feature. Give two different ways we can interpret the resulting feature. (2 points)