Question Booklet

- **This re-exam contains 5 problems 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 (Statistical Learning) **(10 points)**

1. We are asked to inspect the bias-variance trade-off for some unknown model based on the Figure 1 below.
   (a) Describe what happens when we slowly increase the flexibility from 1 to 7. (1 point)
   (b) Considering only natural numbers, which flexibility would you recommend? Why? What is the MSE of the model you chose? (2 points)

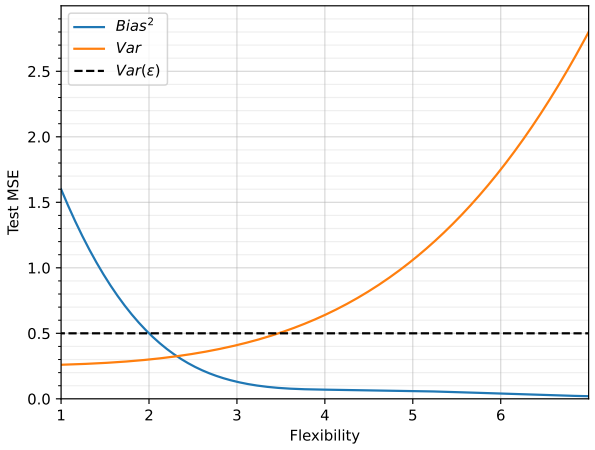

Figure 1: Test Mean Square Error (y-axis) for models of different flexibility (x-axis).

2. Explain for each of the three settings below what will happen, in terms of bias and variance, when we make the proposed change to the learning procedure. (3 Points)
   - We use local regression and change the fraction of training points from $s = 0.01$ to $s = 0.2$.
   - We replace the LDA classifier with the QDA classifier.
   - We use k-means clustering, and change the number of clusters k from 3 to 7.
3. In linear regression, we *generally* assume that the error (noise) term is zero on average. Why? (1 Point)
4. Are the following methods parametric or non-parametric? For each, explain why. (3 Points)
   - KNN (K-Nearest Neighbours).
   - SVM with a Radial Kernel.
   - A fully connected feed forward neural network of 3 layers of 25 nodes each, using standard sigmoidal activation functions.

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

We are asked to fit a quadratic polynomial (degree = 2) to the data from the following table. We are told that $\beta_0 = 0$ and hence do not have to estimate it.

X_1	-1	-0.7	-0.3	0	0.3	0.7	1
Y	4.09	1.33	0.99	-0.46	0.44	1.75	3.41

Recall that a quadratic polynomial (with fixed $\beta_0 = 0$) takes the form $Y = \beta_1 X_1 + \beta_2 X_1^2$, but that it is often convenient to formulate it as $\mathbf{X}\beta = \mathbf{Y}$ with $\beta = [\beta_1\ \beta_2]^\top$ and $\mathbf{X} = [\mathbf{X_1}\ \mathbf{X_1^2}]$. Note that the square in the last formula denotes the element-wise squaring of $\mathbf{X_1}$.

1. Estimate the coefficients β_1 and β_2 using the following result

$$(\mathbf{X^T}\ \mathbf{X})^{-1} = \begin{bmatrix} 0.32 & 0 \\ 0 & 0.4 \end{bmatrix}.$$

   Explain each step. (2 points)

2. Using an advanced fitting procedure, Prof. Vreeken obtained $\hat{\beta}_1 = -0.1$ and $\hat{\beta}_2 = 3.5$. Based on these estimates, what can you say of the relationship between X_1 and Y? Do you think a linear model (without the quadratic term) would give a better or worse fit, or is that impossible to say? Why? (2 point)

3. Sketch the residual plot for the given data, using the estimates provided in Problem 2.2. Interpret your plot. Does the plot support your conclusion from Problem 2.2? (2 points)

4. Compute the R^2 score, using again the parameters provided in Problem 2.2. What does the R^2 value, in general, tell you about the fit of your model? (2 points)

5. If we know that the data was generated as $Y = \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$, with $\mathbb{E}[\epsilon] = 0$ and $Var(\epsilon) = 0.15$, can you tell, 95%-confidently, that the trend given by β_2 in Problem 2.2 holds? Why? (2 points)

Problem 3 (Non-linear) **(10 points)**

1. Draw a decision tree that produces the partitioning shown in Figure 2. (2 points)

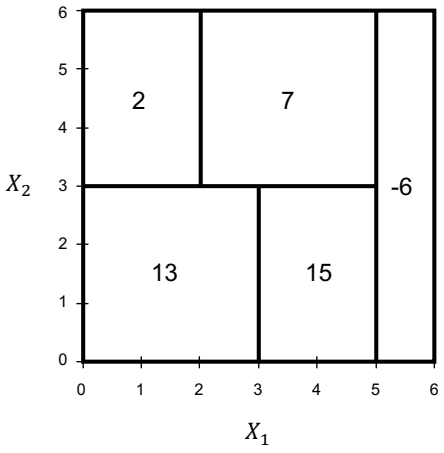

Figure 2: Partition produced by an unknown decision tree. The values within a region indicate the mean of Y within that region.

2. We consider regression splines with one predictor. (2 points)

   How many degrees of freedom has a regression spline model with two knots ($K = 2$) and linear functions ($d = 1$) as splines, when we as usual enforce continuity in derivatives at each knot up to degree $d - 1$. Explain the interpretation for each degree of freedom. **Do not just use the equation you might happen to know.**

   Example explanation of how each degree of freedom of a linear regression model can be interpreted: "The first degree of freedom specifies the intercept, the second specifies the slope of the linear function."

3. Suppose that we compute a curve $\hat{g}$ 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 flex his coding skills Prof. Vreeken implemented a fitting algorithm that allows values of λ of i) $\lambda = 0$, ii) $\lambda = 1$, and iii) $\lambda = \infty$. He applied the algorithm on some data and got the result shown below. Which of the three possible values of λ was used to obtain the fit shown in Figure 3? Explain your answer. (2 points)

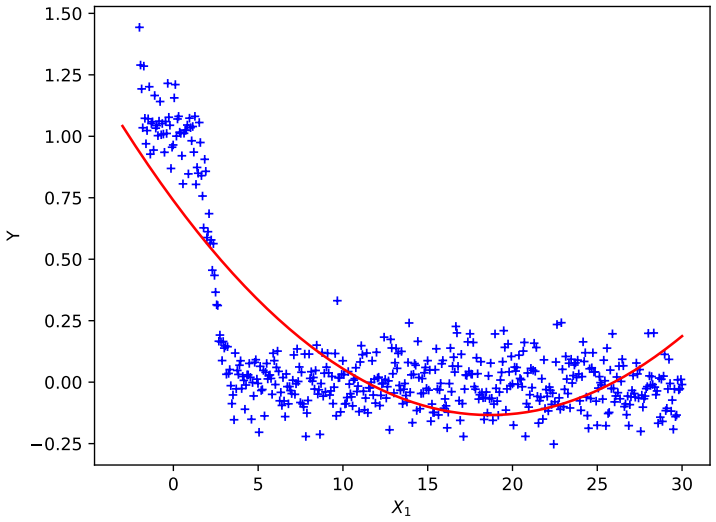

Figure 3: Smoothing spline with unknown parameter λ

4. To prepare for EML next year, Prof. Vreeken is studying an obscure regression method that is similar to Ridge Regression and LASSO, but has a different penalty term. It estimates $\hat{\beta}$ by minimizing the following term

$$RSS + \lambda \sum_{j=1}^{p} I(\beta_j \neq 0) \ ,$$

where $I(\beta_j \neq 0)$ is an indicator function that takes on a value of 1 if $\beta_j \neq 0$ and 0 otherwise. How does this method relate to best subset selection? (1 point)

5. Prof. Vreeken wants to fit a function to the data shown in Figure 4. He consider two models.

   As the first model (Model A) he considers a Generative Additive Model (GAM) with cubic polynomials (degree = 3) as basis functions

$$y_i = \beta_0 + \sum_{j=1}^{2} f_j(x_{ij}) \ ,$$

   where f_j is a degree 3 polynomial. As the second model (Model B), he considers a plain linear model over X_1, X_2, and the combination $X_1 * X_2$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

   Explain why one of the two models can fit the data well, while the other one cannot.

(2 points)

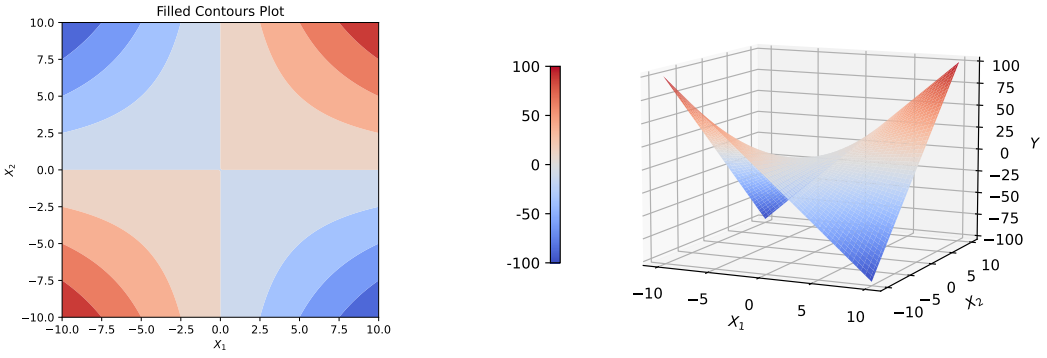

Figure 4: Same data visualized in two different ways. [Left] Contour plot where Y is indicated by the color, from deep blue representing Y values between -75 and -100 to dark red representing Y values from 75 to 100. [Right] 3D Plot of the same function. Y value shown on the vertical axis.

6. We want to learn a Linear Model Tree from the data shown in Figure 5. A Linear Model Tree is very similar to a standard regression tree. It is only different from a standard regression tree in that $\hat{y}_{R_k}$ now represents a linear model instead of the mean.

   At each step, the algorithm chooses that 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

   $$\min_{j,s} \sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \,, \tag{3.1}$$

   where $\hat{y}_{R_k}$ denotes the prediction of a linear model fitted using the data from that region. That is, $\hat{y}_{i,R_k} = x_i a_{R_k} + b_{R_k}$.

   Sketch[1] into Figure 5 how a Linear Model Tree would fit the shown data. Split the data into **at least** 3 and **at most** 5 regions. The resulting Linear Model Tree does not have to be balanced. Clearly indicate the boundaries of the region as well as $\hat{y}_{R_k}$ of each region. (1 point)

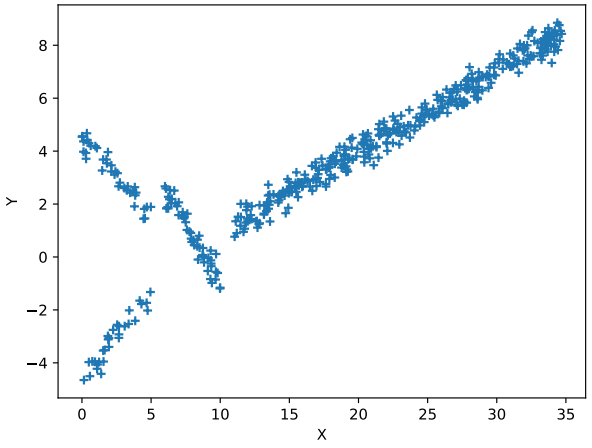

Figure 5: Data for the Linear Model Tree.

[1] You may also recreate the plot by hand, but do make sure you include all key details.

Problem 4 (Classification) **(10 points)**

1. In classification, values outside of the typical distribution of a predictor but on the "correct" side of the decision boundary are a special kind of outliers. Are the following methods sensitive to these special outliers? Explain why (not). (2 points)
   - Logistic regression.
   - Classification Decision Trees.
   - Support Vector Machines (SVM).
   - K-Nearest Neighbours (KNN).
2. You are given the following decision boundary of a Support Vector Classifier
   $$7 - 3x_1 - 2x_2 + 4x_3 + 7x_4 \ .$$
   Using this boundary, assign the following point to either the positive or negative class, $x = (-2, 3, 6, -4)$. (1 point)
3. Recall that according to the Bayes Theorem, $Pr(Y = k|X = x) \propto \pi_k * f_k(x)$, where π_k is the prior and f_k the likelihood function. A domain expert tells you that their data most certainly follows a Poisson distribution with distinct values λ_k for each of the K classes, where
   $$f(x, \lambda_k) = \frac{(\lambda_k)^x \, e^{-\lambda_k}}{x!} \ .$$
   Derive the discriminant function. Simplify the discriminant function as much as possible. (2 points)
4. We train a decision tree for a binary classification problem. We are given 20 data points out of which 16 belong to Class 1 and 4 to Class 2. Prof. Vreeken's implementation finds only one possible way to split the data: one region containing 10 points out of which 10 belong to Class 1, and a second region containing 10 points out of which 6 belong to Class 1. Does it make sense to split the data this way? Why (not)? How does this setting relate to Misclassification Error, Gini Index, and Cross Entropy? (2 points)
5. We train a decision tree for binary classification using the Gini Index as the quality measure. We are given a categorical predictor with q unordered values.
   (a) Show that there are $2^{q-1} - 1$ possible partitions into two groups, where each group has at least one element. (2 point)
   (b) We want to avoid testing all $2^{q-1}-1$ possible partitions. How can we nevertheless find the best partition? In other words, how can we find an order over the values that allows us to find the best partitioning. (1 point)

Problem 5 (Clustering and Dimensionality Reduction) **(10 points)**

Local French bakery "Les Deux Croissants" asked us to analyse their data.

1. We first consider the data given in Figure 6.

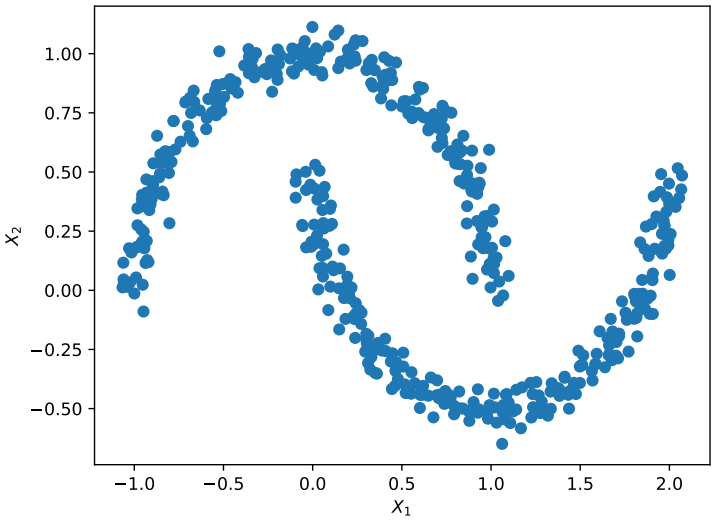

Figure 6: Data from bakery "Les Deux Croissants".

(a) Sketch into Figure 6 the clustering that k-means with $k = 2$ and using Euclidean distance is most likely to find. Explain why k-means is a good/bad choice for clustering this data. (2 points)

(b) Which linkage measure would you recommend for hierarchically clustering this data? Why? (1 point)

(c) Clustering high dimensional data is hard. It may help to first reduce the dimensionality, and then cluster. Consider the following two approaches for the data shown in Figure 6.

- Use PCA to reduce the dimensionality to one dimension, and then apply k-means with $k = 2$.
- Use t-SNE to reduce the dimensionality to one dimension, and then apply k-means with $k = 2$.

Describe the expected result for each. Which one would you choose and why? (2 points)

French bakers are very concerned with the butteriness of their croissants. They wish to predict this value, but although they are certain that only few predictors truly matter, they cannot agree which these possibly would be.

2. Suppose we are given $n = 1000$ datapoints and $p = 20$ predictors. Y is a linear function of 3 predictors. We consider using either PCR or PLS to reduce the dimensionality of this data to 5 dimensions. Describe a scenario in which PLS would work but PCR would fail to provide a meaningful result. (2 points)

At the last moment, the baker's assistant raises a serious concern about Problem 5.1a above. Does k-means even converge?

3. Explain why the k-means algorithm always converges. (2 points)
4. Is k-means, in general, sensitive to outliers in the data? Explain why (not). (1 point)