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 11 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.**
- **No cats were harmed in the preparation of this exam.**

Problem 1 (Linear Regression) **(10 points)**

1. Determine if the following statements are true or false. For every false statement, either correct it by replacing a *single* term (noun or adjective) *or* provide a counter example. (3pts)

   (i) Consider a linear model that consists of 3 predictors. We typically use the t-test to measure the collinearity of a single predictor with any of the rest.

   (ii) For a linear model that satisfies the least square assumptions, the R^2 statistic follows a t-distribution.

   (iii) Removing a high-leverage point always increases the accuracy of a linear model estimated using least squares.

   (iv) The least squares estimator can always be computed.

2. Not-yet-famous researcher Kanis Jalofolias is super interested in how the size (X_s) and weight (X_w) of a cat affects the loudness of its meow (Y). Each day he collects a dataset of all meows he encounters on random stray cats, normalises the predictors and creates a dataset. Since he knows that the relation must fulfil the OLS assumptions, he used this method on each dataset to fit a model (2pts)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_w X_w + \hat{\beta}_s X_s \,.$$

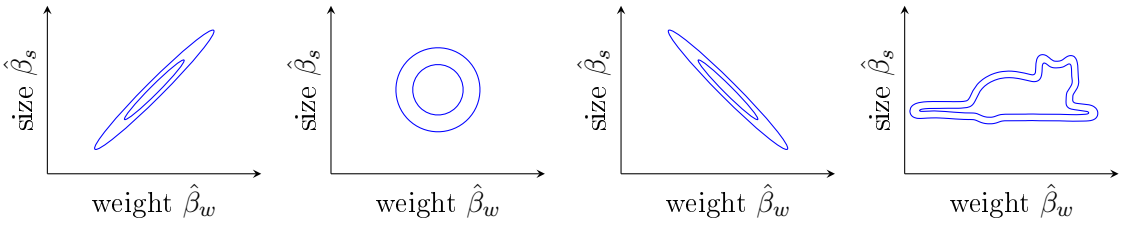
   After a few days he has a collection of models, and he creates a scatter plot where each point corresponds to the coefficients of each predictor for a given model. Which of the scatter plots below corresponds to the hypothetical case where:

   (i) The predictors *weight* and *size* are uncorrelated?

   (ii) The predictors *weight* and *size* are very positively correlated?

   (iii) The predictors *weight* and *size* are very negatively correlated?

   You are given 4 plots out of which you need to only use 3. Briefly explain all your choices and also why you left out the plot without a pair.



(a) Case A (b) Case B (c) Case C (d) Case D

Figure 1: Contour plots of the joint probability density of coefficient estimates $\mathbb{P}\left(\hat{\beta}_w, \hat{\beta}_s\right)$.

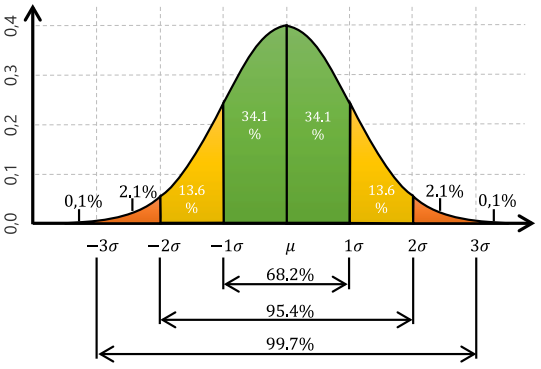
3. NopeAI recently received a large investment, and now wants to predict the added revenue based on two predictors X_{bf} and X_{tw}, corresponding to its advertising budget for two different media, Bacefook and Twutter.

   (a) How can we verify whether a linear model is the right choice? (1pt)

   (b) Give an example scenario where adding an interaction term between Bacefook and Twutter would lead to a better model. (1pt)

   (c) World-famous CEO of NopeAI, Melon Usk, says that the prediction accuracy will improve if we add predictor X_{bf}^{99} to the model. How can we determine if this improvement is significant? (1pt)

4. Not-yet-famous researcher Kavid Daltenpoth considers a linear regression task with only one predictor. Using OLS he obtains an estimated $\hat{\beta}$. The variance of the estimate is $\text{Var}(\hat{\beta}) = 0.49$. He looks at the following plot



and concludes:

> "*If X was uncorrelated with the outcome, with a probability* 2.2% *the value of $\hat{\beta}$ computed on a random dataset drawn from the true model would be greater than the one I just found.*"

   (a) Write down the equation that describes the fact Kavid stated. What is the common name for this probability? (1pt)

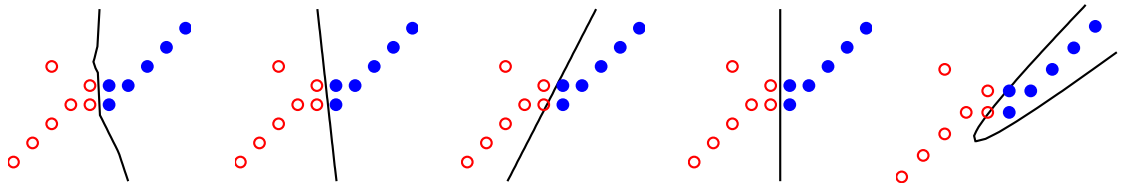
   (b) What is the value of $\hat{\beta}$? (1pt)

Problem 2 (Classification) **(10 points)**

1. A linear classifier that uses the predictors X_1, X_2, X_1X_2, X_1^2, and X_2^2 will have a decision boundary that falls into one out of 5 characteristic cases. Choose 4 out of the 5 cases, name them, and draw an example decision boundary for each. (1pt)

2. In Fig. 2 we show the decision boundaries for five different classification methods. Pair each of the boundaries to exactly one of the following classification methods and briefly explain each of your choices: (3pts)

   (i) LDA – Linear Discriminant Analysis
   (ii) QDA – Quadratic Discriminant Analysis
   (iii) LR – Logistic Regression
   (iv) 3-NN – k-Nearest neighbours with $k = 3$
   (v) SVC – Support Vector Classifier (hard margin, no kernel).



Boundary A Boundary B Boundary C Boundary D Boundary E

Figure 2: Different decision boundaries for the same dataset.

3. We consider the binary classification problem where based on a single predictor X we want to classify samples into one out of two classes '+' and '−'. We know that

   - $P(X|Y = \text{'+'}) = \mathcal{N}(0, 1)$ is a Gaussian with zero mean and unit variance,
   - $P(X|Y = \text{'−'})$ is uniform over the interval $[-\alpha, \alpha]$ with some parameter $\alpha > 0$,
   - and $P(Y = \text{'+'}) = P(Y = \text{'−'})$ the classes are equally likely.

   Without performing extensive computation, answer the following questions.

   (a) Draw the decision boundary of the Bayes optimal classifier for $\alpha = 2.5$ on the real line of the graded axes in your answer booklet (page 5). You may use rounding to 1 decimal point. (1pt)
   (b) Briefly explain what happens to the decision boundary if α increases slightly. (1pt)
   (c) Briefly explain what happens to the decision boundary if the prior $P(Y = \text{'+'})$ increases slightly. (1pt)

   You may find useful the Gaussian probability density function shown in Fig. 3.

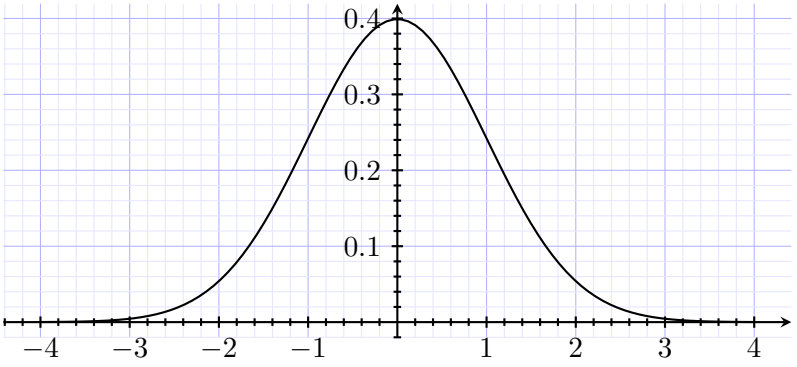


Figure 3: Probability density function of standard Gaussian $\mathcal{N}(0, 1)$.

4. You need to perform binary classification on the dataset shown in Fig. 4 that contains two predictors X_1 and X_2. You decide to use a support vector machine. For this, you consult for advice the not-yet-famous researcher Mara Saseche.

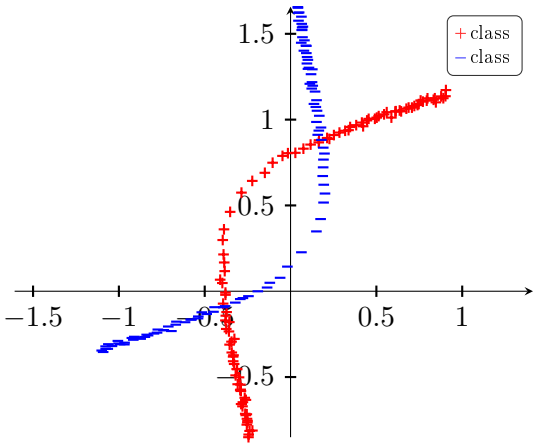
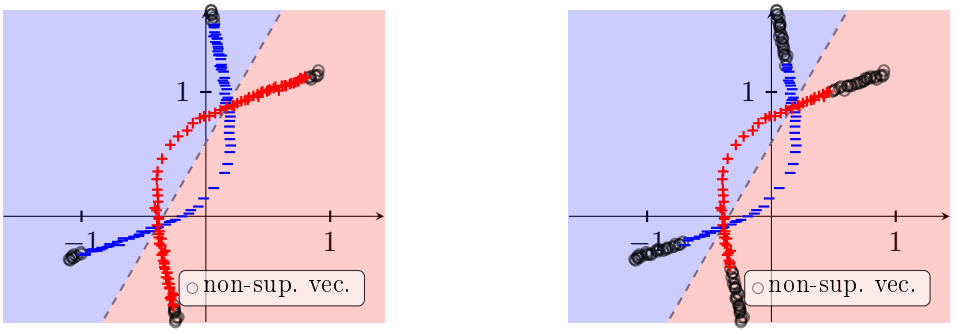


Figure 4: Dataset for support vector machine classification.

(a) You ask Mara to fit a hard margin classifier. Instead she fits the two soft margin classifiers shown in Fig. 5. Why couldn't Mara fulfill your request? How are the two models different? (1 pt)



(a) Model A (b) Model B

Figure 5: Mara's two models. The dashed line shows the decision boundary. The background color shows the predicted class. The '+' and '−' denote data points that are support vectors. The black circles show data points that are *not* support vectors.

(b) You asked Mara for a better classifier. She came up with a support vector classifier for which the decision boundary is shown in Fig. 6. Briefly explain how she achieved this decision boundary. (1 pt)

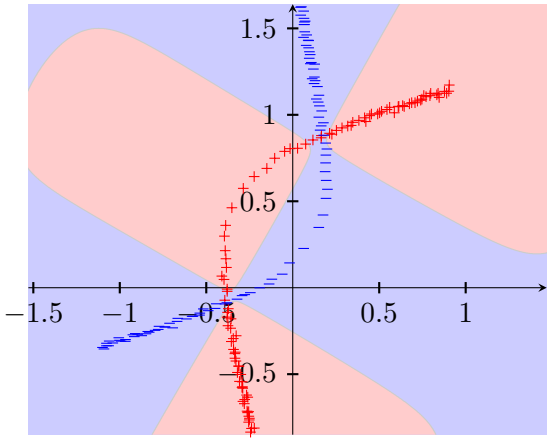


Figure 6: Support vector classifier with an improved decision boundary.

(c) As you discuss further, Mara informs you of the following correct statement. (1 pt)

> "*Using an SVM with a polynomial kernel of degree 2 is equivalent to using an SVM on a dataset that has been extended to additionally include predictors* X_1^2, X_1X_2, *and* X_2^2."

Given the above, why would anyone still want to use a polynomial kernel?

Problem 3 (Trees and Forests) **(10 points)**

1. Consider the following dataset of two predictors X_1, X_2 and one target Y. (1pt)

X_1	X_2	Y
1	1	1
1	2	1.5
2	1	11
2	2	10.5

   Construct a regression tree for the data such that each leaf contains precisely one data point. Use the approximation $(a - x)^2 \approx a^2 - 2ax$ for $x < 1$ when computing the MSE gains.

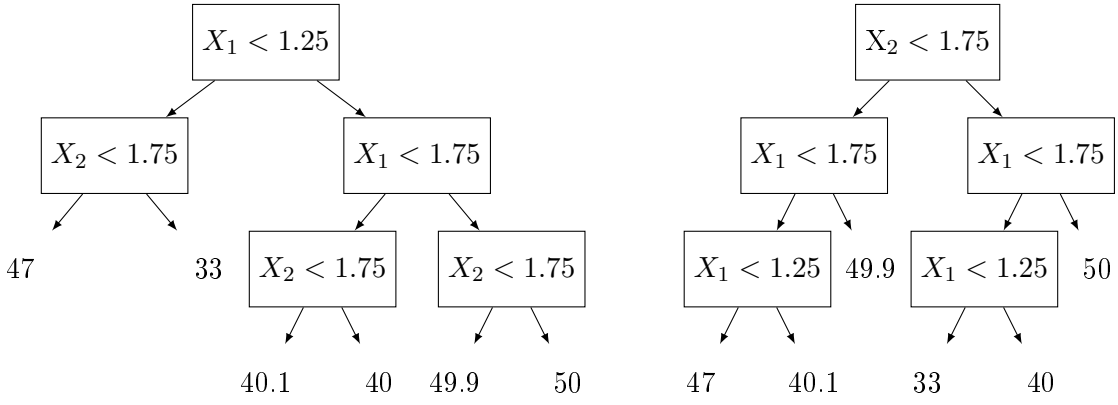
2. Consider the regression tree shown in Fig. 7a.

   (a) Draw the corresponding regions and indicate the value in each region. (1pt)

   (b) Consider the right-most split on $X_2 < 1.75$ in the tree. In terms of generalization, under which conditions does it makes sense to have this node in the tree? What would lead to having this node in the tree even when it does not make sense? (1pt)

   Now consider both regression trees shown in Fig. 7.

   (c) For both pruning and constraining tree depth, explain whether it would work better, worse, or equally well for the left tree vs. the right tree. Based on this, explain the shortcomings of both these approaches for constraining flexibility in regression trees. Explain your reasoning. (3pts)



(a) Regression tree 1 (b) Regression tree 2

Figure 7: Two equivalent regression trees.

3. World-famous ensemble learning researchers Roav Schreund and Yobert Fapire say that modeling data with only a single regression tree is a bad idea, and that we should instead use an ensemble of trees.

   (a) Explain how Bagging works, how Random Forests work, and how these two differ from each other. (1pt)

   (b) Explain how Boosting works, and how it differs from Bagging. (1pt)

   (c) Explain how variable importance is computed for a Random Forest. Can we use the same approach for Boosted Trees? Why (not)? (1pt)

4. Not-yet-famous researcher Kavid Daltenpoth makes the following statement: (1pt)

   "Bagging, Boosting, and Random Forests are all linear models".

   Give one argument or example in favor, and one example or argument against this statement.

Problem 4 (Model Selection) **(10 points)**

1. We are given two datasets, one of $n = 10$, and a second of $n = 1\,000$ samples, over ten predictors $X_1, \ldots, X_{10}$ and one continuous-valued target variable Y. We want to find out which are the relevant predictors for Y.

   (a) We fit a linear regression model using each possible subset of predictors. In Fig. 8, we show the BIC score of the regression model using the k best predictors. Which line corresponds to the small and which to the large dataset? Explain. (1pt)

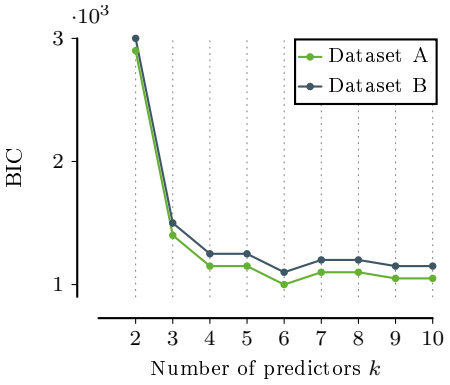


Figure 8: BIC scores of linear regression models that use the k best predictors.

   (b) Based on Fig. 8, how many predictors would you choose and why? (1pt)

   (c) Explain how BIC differs from AIC. Based on this, does the choice between BIC and AIC matter more for the small or for the large dataset? (1pt)

   (d) Briefly explain how we can use each of the following methods to select relevant predictors for Y, and give one advantage and one disadvantage: (3pts)
   - subset selection,
   - cross validation,
   - shrinkage.

2. Consider the following formulation to find the linear regression parameter $\hat{\beta} \in \mathbb{R}^d$,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \qquad \text{subject to } \|\beta\|_2^2 \leq K\,.$$

   (a) Explain the effect of the constraint in the above objective. (1pt)

   (b) How does the above formulation differ from OLS in terms of the number of parameters and in terms of the degrees of freedom? (1pt)

   (c) Consider varying the value of parameter K. What would be the effect on the bias and variance of the model? (1pt)

   (d) How could we modify the constraint to perform subset selection? (1pt)

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

1. Not-yet-famous researcher Kanis Jalofolias is still interested in cats. Looking at the plot in Fig. 9, he is starting to suspect that not all cats are created equal and wants to determine which cats are similar to each other.

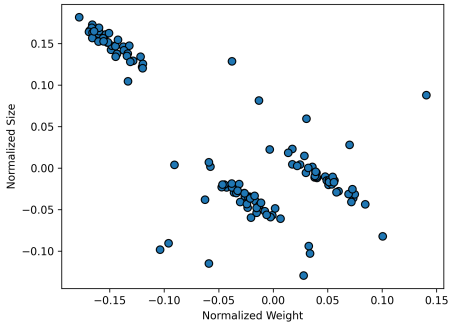


Figure 9: A scatter plot of two different predictors for different cats.

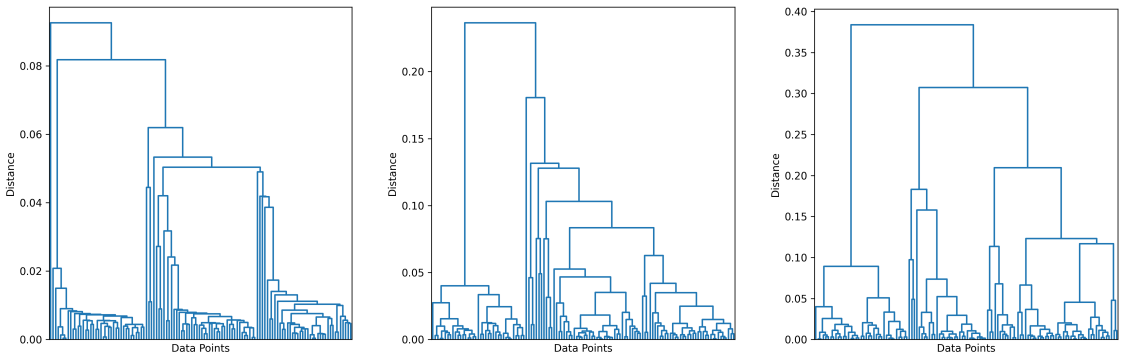
(a) Explain how the k-means clustering algorithm works. Would you expect it to work well on the data depicted in Fig. 9? Why (not)? (1pt)

(b) Kavid and Kanis both run k-means on the same data, with the same distance measure, and the same value for k. They get different results. What happened? (1pt)

2. Catvaid suggests that instead of using k-means, Kavid and Kanis should use hierarchical clustering.

(a) Explain how hierarchical clustering differs from k-means clustering. (1pt)

(b) For the three dendrograms in Fig. 10, explain which one corresponds to single, which one to average, and which one to complete linkage. (2pts)



(a) Dendrogram A (b) Dendrogram B (c) Dendrogram C

Figure 10: Dendrogram plots for different linkage methods.

3. Unhappy with the progress in his quest to understand cats, Kanis measures the genotype of every cat he comes across, resulting in a dataset of $d = 20\,000$ features (genes) and $n = 10\,000$ samples (cats).

   (a) Having gathered all this data, Kanis starts scratching his head, unsure how to deal with it. Explain the problem with the data set he has collected. (1pt)

   (b) In a discussion with Kavid, the two decide to use PCA to reduce the number of features. Kavid suggests they should use PCA *before* clustering the data, while Kanis is insistent on clustering the data first and then using PCA to visualize the resulting clusters. Give arguments both in favor and against either approach. (2pts)

   (c) Mara tells Kanis and Kavid that they are both wrong and should use t-SNE instead. Succinctly explain how t-SNE works and how it differs from PCA. (2pts)