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 8 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.**

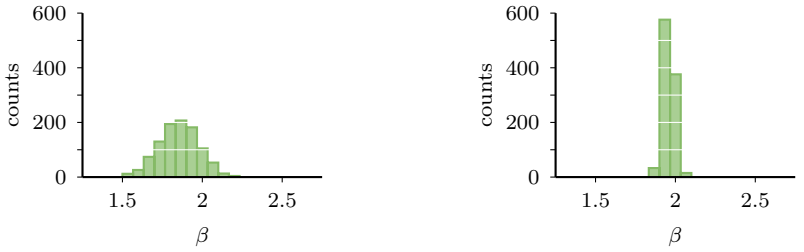
Problem 1 (Lots of Data) **(10 points)**

1. We want to predict target Y given a single predictor X. We collected two datasets from the same distribution, one of $n = 10$ samples, and a second of $n = 1000$ samples. (10 pts)

   We first fit a simple linear regression model.

   (a) What is a good approach to compare the least-squares estimate of β we get for the one dataset to the least-squares estimate of β we get for the other dataset? Explain in your own words. (1 pt)

   (b) We use bagging to obtain a thousand estimates of the parameter β for each dataset. We show the results in Fig. 1. Which of the two figures corresponds to the small dataset ($n = 10$) and which to the large ($n = 1000$) dataset? Explain your choice. (1 pt)



(a) Estimates for dataset A. (b) Estimates for dataset B.

Figure 1: Estimates of the linear coefficient β for two datasets.

   (c) We now consider polynomial regression with degree d. Compare the bias and variance of a model with degree $d = 3$ to a model with degree $d = 10$. (2 pts)

   (d) Explain how we can control the flexibility of

      i. a spline regression model, and (1 pt)

      ii. a regression tree. (1 pt)

   (e) Will fitting a more flexible model on the large dataset ($n = 1000$) *always* achieve a lower test error than fitting a less flexible model on the small dataset ($n = 10$)? If yes, explain your reasoning; if no, explain what modification we can make to the learning procedure to address this aspect. (2 pts)

   (f) To decide the flexibility (e.g., the degree d) of the above models, we consider using k-fold cross validation or leave-one-out cross validation (LOOCV).

      (a) Which do you recommend for the small, respectively the large dataset? Why? (1 pt)

      (b) Which do you recommend in general and why? (1 pt)

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

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

   (i) The Gauss-Markov theorem states there exists no estimator for the coefficients of the linear model that achieves lower variance than the least squares estimator.

   (ii) We should use the t-test to determine if a set of more than one predictor is significantly correlated with the outcome.

   (iii) Assume that we fit a linear model on a single predictor; if its coefficient is 0 then the outcome must be statistically independent with this predictor.

   (iv) Assume a dataset that comes from an underlying linear model $y = \beta x + \epsilon$, where ϵ is arbitrary noise. Then $t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$ follows a student t distribution, where $\hat{\beta}$ is the least squares estimate of β and $\text{SE}(\hat{\beta})$ its standard error.

2. One of our physicist friends is studying a phenomenon between a single predictor X and a target variable Y that can be described as a linear model satisfying the least squares assumptions. We have 50 datapoints shown on Fig 2.

   (a) Explain which out of x_a, x_b, x_c are *outliers*, which of these are *high leverage points*, and which are both? (1 pt)

   (b) Give a short explanation why our friend should be concerned about *outliers*, respectively about *high leverage points*? (1 pt)

   (c) Suppose we may remove *one* datapoint before we fit the least squares estimate. Which out of x_a, x_b or x_c would you remove? Why? (1 pt)

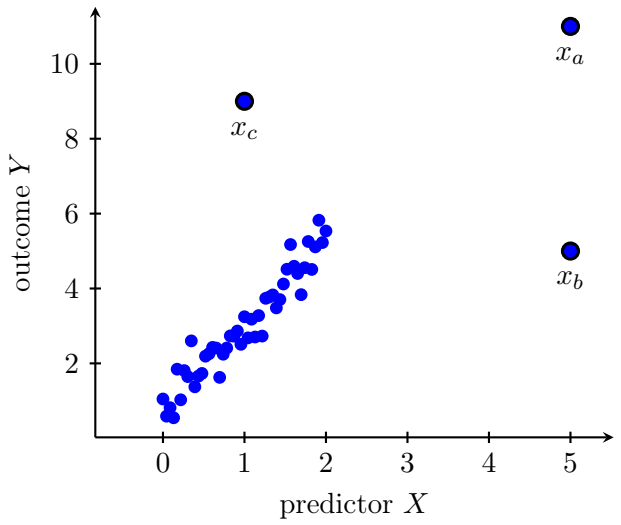


Figure 2: A dateset with 50 datapoints, where three points are annotated as x_a, x_b, x_c.

3. Consider linear models with a single predictor.

   (a) Give a counterexample to the following statement. (1 pt)

   Heteroskedastic noise leads to higher prediction error than homoskedastic noise.

   (b) When should we check for heteroskedasticity? How do we do so? (1 pt)

4. We are analyzing how sales are affected by advertising on three different media. By fitting a multiple linear regression model we get the following coefficients.

	intercept	YouToob	Bacefook	Twutter
coefficient	3.010	0.040	0.190	-0.010

   Based on these results, co-worker A suggests that the company should stop advertising on Twutter as it hurts sales. Co-worker B suggests the opposite.

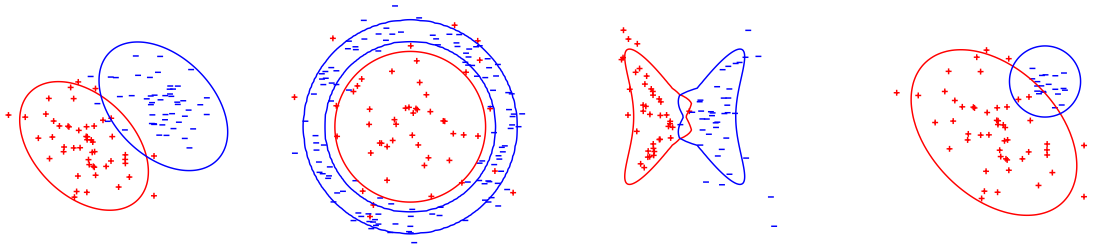
   (a) Give a brief explanation or a counter example why colleague A might be wrong. (1 pt)

   (b) Explain how we can test whether colleague B might be right. (1 pt)

Problem 3 (Classification) **(10 points)**

1. Assign each of the following classification methods to one dataset shown in Figure 3 such that they achieve the best possible accuracy. Briefly explain your choices. (3 pts)

   (a) logistic regression
   (b) linear discriminant analysis
   (c) quadratic discriminant analysis
   (d) nearest neighbours with $k = 1$



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

Figure 3: Likelihood contour plots for four binary classification datasets. Positive points in red (+), negative points in blue (−).

2. Consider the following statement. Is it correct? Why (not)? (2 pts)

   Logistic Regression is a linear model.

3. One of the most commonly used kernels in SVMs is the Gaussian RBF kernel $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$. Suppose we have three points z_1, z_2 and x. Assume $\sigma = 1$, z_1 is close to x and $||z_1 - x|| \ll \sigma$, and z_2 is far from x and $||z_2 - x|| \gg \sigma$. Here the symbols $\ll$ and $\gg$ mean "much smaller" and "much greater" respectively. (2 pts)

   What is the value of $k(z_1, x)$ and $k(z_2, x)$? Choose one of the following. Explain why.

   (i) $k(z_1, x)$ will be close to 1, and
   $k(z_2, x)$ will be close to 0.
   (ii) $k(z_1, x)$ will be close to 0, and
   $k(z_2, x)$ will be close to 1.
   (iii) $k(z_1, x)$ will be close to c_1 such that $c_1 \gg 1$, and
   $k(z_2, x)$ will be close to c_2 such that $c_2 \ll 0$ and $c_1, c_2 \in \mathbb{R}$.
   (iv) $k(z_1, x)$ will be close to c_1 such that $c_1 \ll 0$, and
   $k(z_2, x)$ will be close to c_2 such that $c_2 \gg 1$ and $c_1, c_2 \in \mathbb{R}$.

4. You are training a hard-margin SVM on the dataset shown in Fig. 4.

   (a) Find the optimal weight vector $\mathbf{w}$ and bias b. What is the equation corresponding to the decision boundary? (2 pts)

   (b) Circle the support vectors and draw the decision boundary. Note: Do *not* provide your answer here, but rather on Fig. 1 on page 10 of the *answer sheet*. (1 pt)

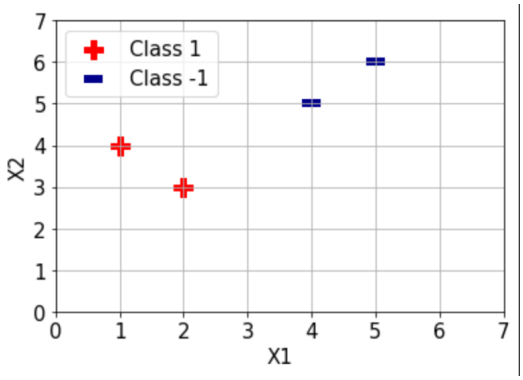
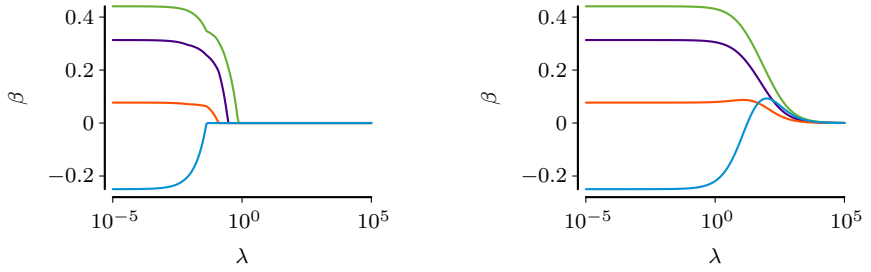


Figure 4: Dataset of two positive (+) and two negative (−) datapoints.

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

1. We have a dataset where we want to predict Y given four predictors, and consider a simple linear regression model with coefficients β.



(a) Regularization method A. (b) Regularization method B.

Figure 5: Linear coefficients β per predictor for varying regularization strength λ.

   (a) Which of the plots in Fig. 5a, 5b corresponds to Ridge and which to Lasso? Explain your reasoning. (2 pts)

   (b) Assume that we know that Y is influenced by only few predictors. Which of the two methods would you then prefer? Why? Explain how you would interpret the plots in Fig. 5 in this case. (2 pts)

   (c) Compare the behavior, in terms of bias and variance, of linear models with ridge regularization strengths $\lambda = 0.1, \lambda = 1$, and $\lambda = 10$. (1 pt)

   (d) Which approach can you use to select an appropriate tuning parameter λ? Explain this approach in your own words. (1 pt)

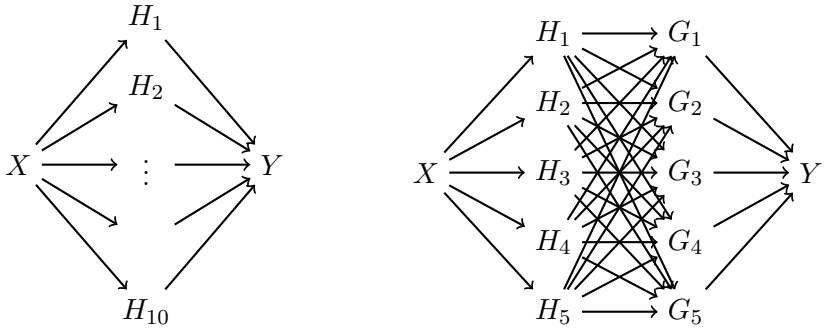
2. Consider the following regression objective for n datapoints and p predictors, (2 pts)

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 \right) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \,.$$

   Explain how this is different from the linear regression objectives you have encountered in the lecture, and what behavior you expect for this model.

3. Finally, we consider standard regression splines (i.e. not smoothing splines). Give two ways to constrain the flexibility of the model. Explain how this is similar or different to linear and polynomial regression. (2 pts)

Problem 5 (All Those Parameters) **(10 points)**



(a) Network with 1 hidden layer. (b) Network with 2 hidden layers.

Figure 6: Two neural networks with 10 hidden neurons (bias neurons not shown).

1. Consider the neural networks in Figure 6.
   (a) How many free parameters does the network in Fig. 6a have, if X and Y are univariate, and biases are *non-zero*. Explain your reasoning. (1 pts)
   (b) Explain which of the two networks is more expressive when using the sigmoid activation function $\sigma(t) = \frac{e^t}{1+e^t}$. (2 pts)
2. Yunn LeCann says that deeper networks are better, and proposes to use a neural network with 3 hidden layers and a total of 50 free parameters.
   (a) How many knots (K) would we have to pick for a cubic spline to have 50 free parameters? How many for a linear spline? Explain your reasoning. (2 pts)
   (b) The linear spline, the cubic spline, and the neural network that Yunn LeCann proposes all have 50 free parameters. Does this mean they will fit a given data set equally well? If so, explain why. If not, give a counter example on which one of them performs better than *at least* one of the other two models. (2 pts)
3. Let M_1 and M_2 be two model classes such that the ratio of number of free parameters $\frac{\text{FP}(M_2)}{\text{FP}(M_1)} > C$ is very large. Is it possible for M_1 to perform better in terms of generalization than M_2 even for arbitrarily large C? Why (not)? (1 pt)
4. We have two models that obtain exactly the same error on a held out test set. Give three reasons why one model may nevertheless be preferable to the other and explain your reasoning. (2 pts)