+1/1/60+

Final Exam

Introduction to Machine Learning for Bioengineers BIO-322

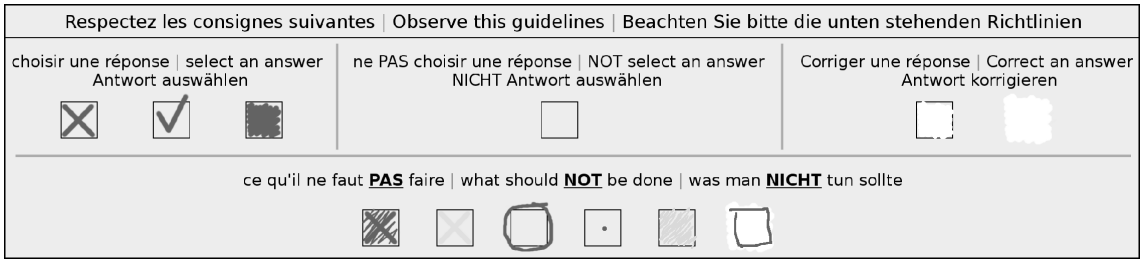
1

Firstname Lastname

123456

February 2, 2023

- The exam lasts 180 minutes.
- Write all your answers in English in a legible way on the exam (no extra sheets).
- Use a dark (e.g. black or blue) pen or pencil.
- Use the last 4 pages of this exam as scratch space.
- 1 page handwritten notes is allowed (A4 one side).
- The handwritten notes must be yours; copies from other students are not allowed.
- No calculator or other electronic device is allowed.
- Put your bag including computer and cell phone to the indicated places.
- Have your student card displayed before you on your desk.
- Do not write your name or sciper number on any page.
- Check that your exam has 14 pages.



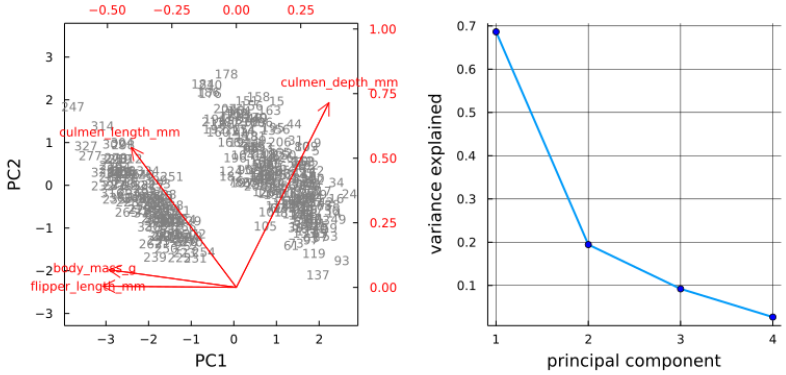
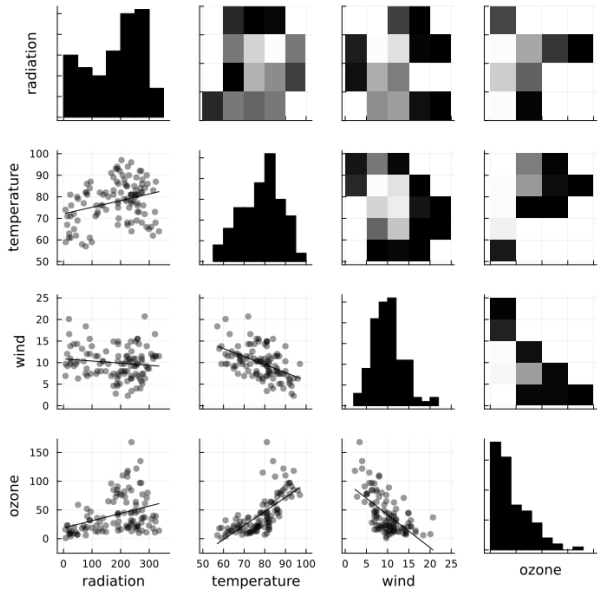


Figure 1: Principal Component Analysis of the Penguin Dataset.



The data contains measurements of solar radiation, temperature, wind speed, and cube root ozone concentration on 111 days at sites in the New York metropolitan region. Objective: predict ozone level.

- temperature in degree Fahrenheit.
- wind speed in miles per hour
- solar radiation in langleys
- cube root ozone level in $\text{ppb}^{\frac{1}{3}}$.

Figure 2: The ozone data set.

Choice Questions (36 Points)

For each of the following questions there is one correct answer and you have **two possibilities: tick one of the boxes or none**. Every correct answer gives 2 positive point, every wrong answer 1 negative point, and no answer no point.

features: 1 numerical, 1 categorical, 2 categorical

Question 1 We want to predict the height of 10 weeks old corn seedlings, based on the average temperature (in °C), the soil condition (dry or humid) and the fertilizer condition (none, biological, chemical). After one-hot coding relative to a standard level, the number of parameters of a linear regression (with intercept) is

+1

☐ 3 ☐ 4 ☒ 5 ☐ 6

Question 2 Suppose you did logistic regression and you get a test set $((x_1 = 1, y_1 = A), (x_2 = -3, y_2 = A), (x_3 = 4, y_3 = B), (x_4 = -1, y_4 = B))$. The predictions of your fitted machine on this test set are $P(y_1 = A|x_1) = 0.7, P(y_2 = A|x_2) = 0.7, P(y_3 = A|x_3) = 0.6, P(y_4 = A|x_4) = 0.6$. The AUC on this test set is

for thresholds between 0.6 and 0.7 the true positive rate is 1 and the false positive rate is 0.

☐ 0 ☐ 0.5 ☐ 0.6 ☒ 1

Question 3 Neural networks with strong L2 regularization tend to have a lower bias but a higher variance than neural networks without regularization.

reduces variance and increases bias.

☒ Wrong. ☐ Correct.

Question 4 Cross validation can be used to select the number of hidden layers in a neural network.

☐ Wrong. ☒ Correct.

Question 5 Consider the data generating process $y = \text{A}$ if $2x - 1 > 1$ and $y = \text{B}$ otherwise. The irreducible error of this generator is larger than 1.

This data generator is deterministic $\Rightarrow$ irreducible noise $= 0$.

☒ Wrong. ☐ Correct.

Question 6 Given data points $((x_1 = 0, y_1 = 0), (x_2 = 2, y_2 = 2), (x_3 = 4, y_3 = 0))$, the leave-one-out cross-validation estimate of the mean squared error for linear regression is

$\frac{1}{3}(4^2 + 2^2 + 4^2) = 12$

☐ 4 ☒ 12 ☐ another value

Question 7 With the Lasso with a large penalty λ, many parameters are exactly 0.

☐ Wrong. ☒ Correct.

Question 8 Assume the data set $((x_1 = 1, y_1 = 10), (x_2 = 3, y_2 = 0), (x_3 = 4, y_3 = 0), (x_4 = 8, y_4 = 0))$ should be fitted with a regression tree. In the first step of recursive binary splitting, the split is introduced at

☒ $x = 2$ ☐ $x = 3.5$ ☐ $x = 6$.

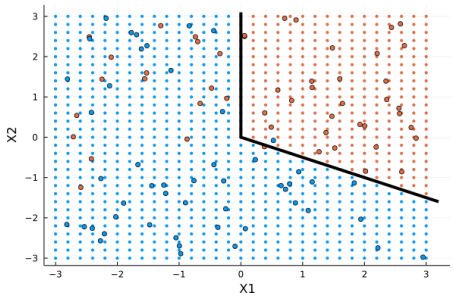
Question 9 The result of K-Means clustering applied to un-standardized data is the same as the result of K-Means clustering applied to standardized data.

☒ Wrong. ☐ Correct.

Question 10 Assume all points of a dataset with three predictors lie on a line. In this case, the proportion of variance explained (PVE) is zero for the second and third principal component.

☐ Wrong. ☒ Correct.

Question 11



The black decision boundary on the left could have been obtained by fitting the following method on the training data (shown as large blue and red dots).

☒ a neural network classifier with one hidden layer ☐ logistic regression
☐ 2 nearest neighbours classification ☐ none of these methods

Question 12 The largest variance in the penguin dataset in figure 1 on page 2 is parallel to the flipper_length_mm axis. $\phi_1 \approx (-0.55, -0.5, -0.4, 0.8)$

☒ Wrong. ☐ Correct.

Question 13 The task described in figure 2 on page 2 is a problem of type:

☒ regression ☐ classification

Question 14 For a given data generating process and some training data, we find for a quadratic fit $f_1(x) = 1 + 2x - 0.5x^2$ a training error of 0.5 and for a linear fit $f_2(x) = 0.5 + 1.8x$ a training error of 1.5. This implies that the test loss of f_1 for the given data generating process is equal or lower than the test loss of f_2. f_1 could be overfitting the training data.

☒ Wrong. ☐ Correct.

Question 15 Early stopping in gradient descent can be used to find the value of the regularization constant λ for the Lasso.

☒ Wrong. ☐ Correct.

Question 16 The data points in the penguin dataset in figure 1 on page 2 lie all very close (in the sense of more than 95% of the dataset's variability lies within this subspace) to a

☐ straight line ☐ 2D plane ☒ 3D linear subspace
☐ none of the other options

Question 17 In a classification task with 3 classes we know that the data generating process has probabilities $P(y = A|x = 1) = P(y = B|x = 1) = \frac{1}{4}$ and $P(y = C|x = 1) = \frac{1}{2}$. How large is the test log-likelihood at $x_0 = 1$ for a linear model with fitted parameters $\hat{\theta}_{10} = \hat{\theta}_{11} = \hat{\theta}_{20} = \hat{\theta}_{21} = \hat{\theta}_{30} = \hat{\theta}_{31} = 0$? $\frac{1}{4}\log\frac{1}{3} + \frac{1}{4}\log\frac{1}{3} + \frac{1}{2}\log\frac{1}{3} = \log\frac{1}{3} = -\log 3$

☐ $\frac{1}{4}\log(1)$ ☐ $\frac{1}{3}\log(\frac{1}{3})$ ☒ $-\log(3)$ ☐ 0

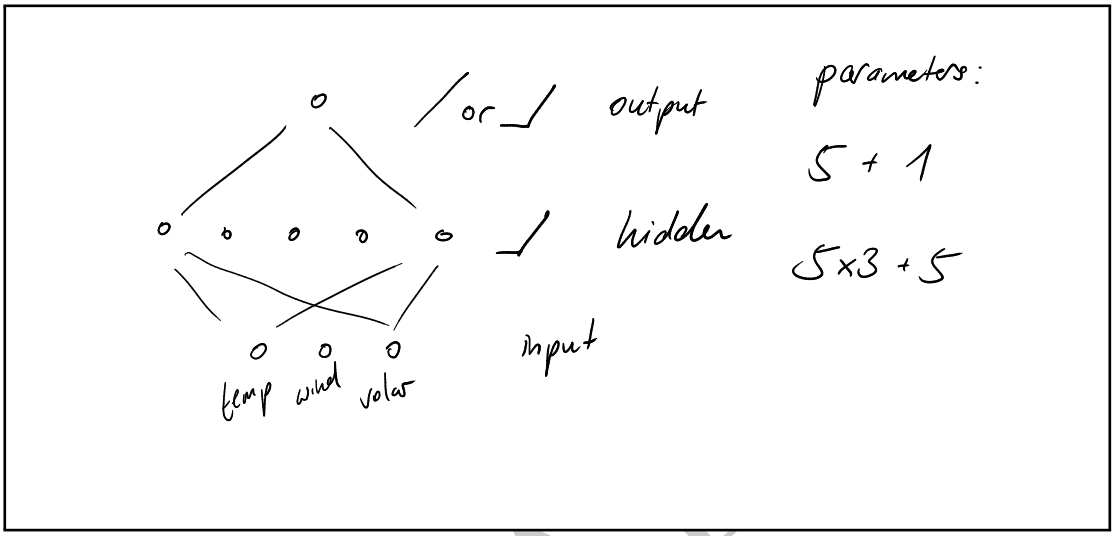
Question 18 The task described in figure 2 on page 2 is a problem of type:

☐ reinforcement learning ☒ supervised learning ☐ unsupervised learning

Open Questions (56 Points)

Please write your answers to the following questions in the designated boxes. Do not tick the checkboxes for grading.

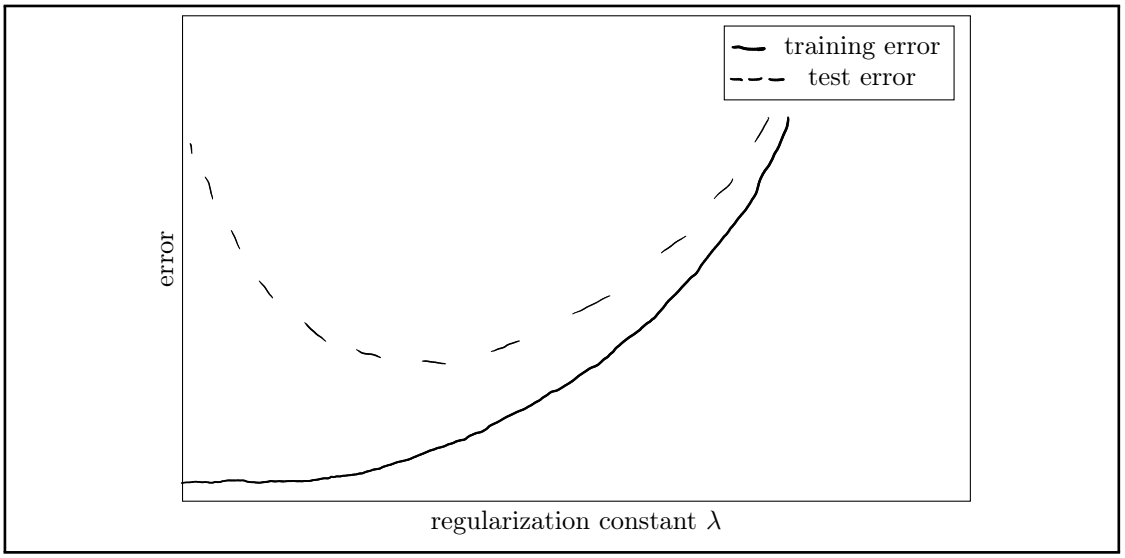
Question 19 Draw a sketch of a neural network with one hidden layer of 5 relu neurons to fit the data in figure 2 on page 2. Label the layers as "input", "hidden" and "output" layer and indicate the activation function of each layer. ☐0 ☐1 ☐2 ☐3 ☐4 *(for grading)*



Question 20 How many parameters (including biases) does the neural network of the previous question have? ☐0 ☐1 ☐2 ☐3 ☐4 *(for grading)*

$5 \times 3 + 5 + 5 + 1 = 26$

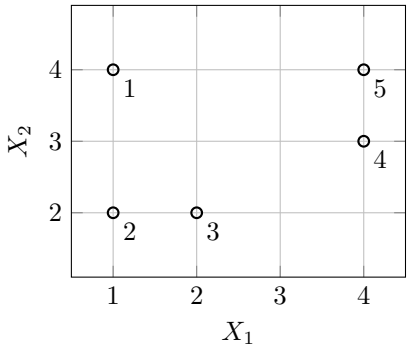
Question 21 Suppose you use L2 regularization in linear regression. Draw a sketch of the expected training and test error as a function of the regularization constant. ☐0 ☐1 ☐2 ☐3 ☐4 *(for grading)*



Question 22

dataset

id	X_1	X_2
1	1	4
2	1	2
3	2	2
4	4	3
5	4	4



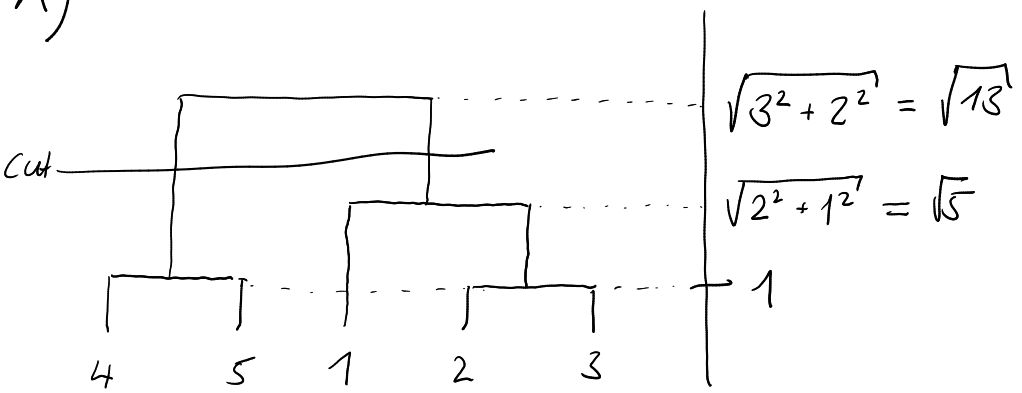
With hierarchical clustering we want to assign each data point in the above dataset to one of two clusters. In the figure the observation identity is given by the number next to the data point. We use the Euclidean distance and complete linkage.

A) Determine the dendrogram and compute the heights at which branches merge.

B) Assign each point to one of two clusters.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 *(for grading)*

A)



B) We cut the dendrogram at some height between $\sqrt{5}$ and $\sqrt{13}$

$\Rightarrow$ cluster 1: $\{4, 5\}$

cluster 2: $\{1, 2, 3\}$

Question 23 You are given the data set $((x_1 = 1, y_1 = 2), (x_2 = -1, y_2 = 4))$. Assume the data comes from a generator with normally distributed (Gaussian) noise and you would like to fit this data with the family of probability densities $p(y|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta_0-\theta_1 x)^2}{2}\right)$. Compute one update-step of gradient descent on the negative log-likelihood loss with learning rate $\eta = \frac{1}{4}$, assuming initial guess $\theta_0^{(0)} = 2, \theta_1^{(0)} = 1$.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 *(for grading)*

Negative log-likelihood loss:

$$L(\theta_0, \theta_1) = -\sum_{i=1}^{2} \log p(y_i | x_i)$$

$$= \frac{1}{2}(2 - \theta_0 - \theta_1)^2 + \frac{1}{2}(4 - \theta_0 + \theta_1)^2 + \text{const.}$$

(y_1, $x_1 = 1$, y_2, $x_2 = -1$)

$$\frac{\partial L}{\partial \theta_0} = -(2 - \theta_0 - \theta_1) - (4 - \theta_0 + \theta_1) = -6 + 2\theta_0$$

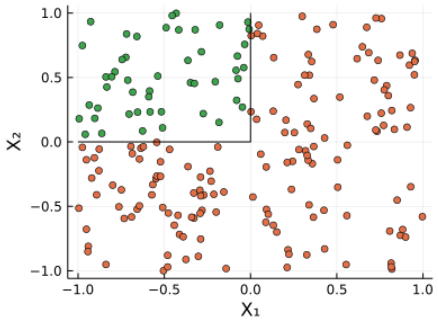
$$\frac{\partial L}{\partial \theta_1} = -(2 - \theta_0 - \theta_1) + (4 - \theta_0 + \theta_1) = 2 + 2\theta_1$$

$$\theta_0^{(1)} = \theta_0^{(0)} - \eta \frac{\partial L}{\partial \theta_0} = 2 - \frac{1}{4}(-6 + 2 \cdot 2) = \underline{2.5}$$

$$\theta_1^{(1)} = \theta_1^{(0)} - \eta \frac{\partial L}{\partial \theta_1} = 1 - \frac{1}{4}(2 + 2 \cdot 1) = \underline{0}$$

Question 24

On the right you see a decision problem with non-linear decision boundary (black line).



1. Determine a feature representation with 2 features $H_1 = \text{heaviside}(w_{11}X_1 + w_{12}X_2), H_2 = \text{heaviside}(w_{21}X_1 + w_{22}X_2)$ such that the decision boundary is linear in the feature representation. Give your answer in the form $w_{ij} = ...$ for all $i = 1, 2$, $j = 1, 2$. The heaviside function is given by $\text{heaviside}(x) = 1$ if $x > 0$, $\text{heaviside}(x) = 0$ otherwise.

2. Show that the decision boundary is indeed linear, i.e. show that there are regression coefficients $\beta_0, \beta_1, \beta_2$ such that $\beta_0 + \beta_1 H_1 + \beta_2 H_2 > 0$ for all points with $X_1 < 0$ and $X_2 > 0$ and $\beta_0 + \beta_1 H_1 + \beta_2 H_2 < 0$ for all other points.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 *(for grading)*

1, $w_{11} = -1$ $w_{12} = 0$ $\Rightarrow H_1 = 1$ if $X_1 < 0$

$w_{21} = 0$ $w_{22} = 1$ $\Rightarrow H_2 = 1$ if $X_2 > 0$

2. $\beta_0 = -1.5$, $\beta_1 = \beta_2 = 1$

$\Rightarrow -1.5 + 1 + 1 = 0.5 > 0$ for all points with $X_1 < 0$ and $X_2 > 0$.

For all other points either H_1 or H_2 or both are inactive and therefore $\beta_0 + \beta_1 H_1 + \beta_2 H_2 < 0$.

Question 25 Suppose you receive the following email from a colleague.

Dear Firstname,

For my current project I work with a database of 71 patients with different severity of a certain disease and 112 measured variables from the analysis of the blood etc. I would like to use machine learning to determine which of the 112 measured variables are unlikely to be related with the severity of the disease. Can you suggest one machine learning method you would use for this task? Please explain, how and why you would use this method.

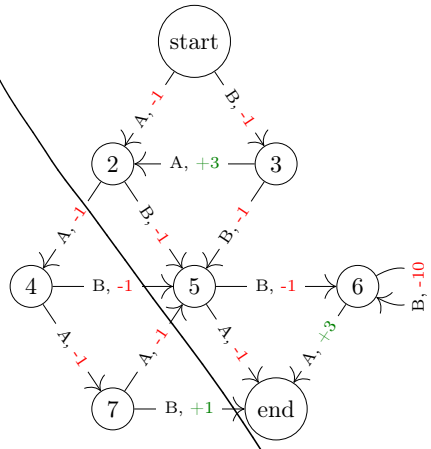
Cheers, B.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 *(for grading)*

Dear B.

I would suggest to run Lasso regression for different values of the regularization constant (this gives you the so-called Lasso path). With strong regularization, the weights of most predictors will be zero, leaving you with the predictors that matter most for linear prediction of the severity.

Your dataset contains fewer samples than predictors, which means you should be careful to avoid overfitting. Therfor I would be careful with interpreting the results of feature importance with a decision-tree-based method.



On the left there is a transition graph of a reinforcement learning task. The text on top of each arrow indicates the action (A or B) that has to be taken to transition from one state to the next and the reward (negative rewards in red, positive rewards in green) when doing this transition. For example, the arrow that starts at state 6 and ends at state 6 indicates that taking action B in state 6 costs -10 and leaves the agent in state 6. Each episode starts in state "start" and ends in state "end". The agent's goal is to find the actions that maximize the cumulative reward in each episode.

Question 26 Which action is optimal in state 5? A or B? ☐0 ☐1 *(for grading)*

Question 27 Assume a reinforcement learning agents learns for a very long time with Q learning: how large is $Q(\text{state } 3, \text{action B})$? ☐0 ☐1 ☐2 *(for grading)*

Question 28 Assume the agent does not know the transition graph. In the first episode it starts in the "start" state and performs the action sequence A, B, A. After this first episode, how large is $Q(\text{state } 2, \text{action B})$

1. for a Monte Carlo learner.
2. for a Q learner, with learning rate 0.5 and all Q values initialized to -1.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 *(for grading)*

Reinforcement learning will not be tested this year.

Question 29 Design a concrete example where we see that the variance of a flexible method is larger than the variance of an inflexible method and that the opposite holds for the bias.

Here is what we mean by a "concrete example": we should see with numbers what you want to show. For example, if the task were to "Show with a concrete example that predictions under linear regression are invariant under scaling of the training input data but predictions with ridge regression are not" a valid answer would be:

Assume data set $((x_1 = 0, y_1 = 1), (x_2 = 1, y_1 = 2)$. For linear regression the solution would be $\theta_0 = 1$ and $\theta_1 = (2-1)/(1-0) = 1$. Therefore, predictions for $x_0 = 2$ would be $\theta_1 x_0 + \theta_0 = 3$. If we scale the input data by $s = 2$, the solution would be $\tilde{\theta}_0 = 1$ and $\tilde{\theta}_1 = (2-1)/(2-0) = \frac{1}{2}$. Therefore, predictions for $\tilde{x}_0 = 2 \cdot 2 = 4$ would be $\tilde{\theta}_1 \tilde{x}_0 + \tilde{\theta}_0 = 3$, the same as before. On the other hand, for ridge regression with $\lambda = 1$ the solution for the unscaled case would be... etc.

☐0 ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 *(for grading)*

Assume a data generator produces datasets with

$\{(x_1 = 0,\ y_1 = \varepsilon_1),\ (x_2 = 1,\ y_2 = 2 + \varepsilon_2)\}$

where ε_1 and ε_2 have mean 0 and variance 1.

We use k-nearest neighbor regression with $k = 1$ for a flexible method and $k = 2$ for an inflexible method (in this context).

For $k = 1$ the bias^2 at $x_0 = 0$ is

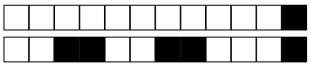
$(f(x_0) - E_{\hat{f}}(\hat{f}(x_0)))^2 = (0 - 0)^2 = \underline{0}$ and

$\mathrm{Var}\, \hat{f}(x_0) = \mathrm{Var}(\varepsilon_1) = \underline{1}$

For $k = 2$ the bias^2 at $x_0 = 0$ is

$\left(0 - \frac{f(0) + f(1)}{2}\right)^2 = \underline{1}$ and

$\mathrm{Var}\, \hat{f}(x_0) = \mathrm{Var}\left(\frac{\varepsilon_1 + \varepsilon_2}{2}\right) = \frac{1}{4}\mathrm{Var}(\varepsilon_1) + \frac{1}{4}\mathrm{Var}(\varepsilon_2) = \underline{\frac{1}{2}}$



This page can be used as a scratch pad. We won't look at it.
If you use this page for results that we should look at **write this explicitly and clearly**.

This page can be used as a scratch pad. We won't look at it.
If you use this page for results that we should look at **write this explicitly and clearly**.

This page can be used as a scratch pad. We won't look at it.
If you use this page for results that we should look at **write this explicitly and clearly**.