Exercise Set 2

- Submit your answer to Moodle by **Wednesday, 26 November 2025**, no later than 23:59.
- You may answer anonymously or include your name, as you prefer.
- Each assignment must be written independently. Collaborative discussion is encouraged, and using external resources—including web searches—is permitted, but all answers must be your own work and not copied.
- All work will be peer-reviewed (by yourself and randomly assigned peers). For expectations, grading rubric, and further details, see the Exercise and Peer Review Instructions. A good answer should:
  - **Clearly explain the reasoning and process**, not just present final results.
  - **Be well-organised and concise**, with logical structure, precise technical language, and properly labelled code snippets/figures/tables as needed.
  - **Demonstrate factual correctness** with relevant, sufficient evidence and well-justified claims.
  - **Show critical analysis**, including justification of methods, discussion of alternatives or limitations, and integration of course concepts to the extend applicable.
  - **Reflect master's-level understanding**: not just procedural work, but explanation, synthesis, and professional communication.
  - **Not look like a code dump**: unless explicitly requested, you do not need to include program code in your answer.
- The language of submission and review is English.
- Upload a single, well-formatted PDF. Double-check before submitting that all figures and sections are present; Jupyter Notebooks can produce incomplete PDFs. We recommend R Markdown (or Quarto). For help, visit https://rmarkdown.rstudio.com/lesson-1.html and https://bookdown.org/yihui/rmarkdown/. It is a good idea to learn a system for writing proper reports, also for future studies and life after graduation. Incomplete or unreadable submissions may result in no points awarded.
- Answer the problems in the order given.
- Consult Moodle's general instructions and grading criteria before starting.
- Main textbook: ISLR_v2 ("An Introduction to Statistical Learning with Applications in R", 2nd edition, James et al.) or ISLP (Python version)—either is acceptable. You may use other materials as well.
- Submit as early as possible: you can revise and resubmit until the deadline. Due to peer review, we cannot accept late submissions. If you miss the deadline (and don't even submit empty PDF) you will loose points from the exercise set, including peer review points. Check in advance, well before the deadline date, that you can produce legible PDF files (especially with Jupyter notebooks students have often had last-minute surprises as PDF generation has now worked as expected). You have been warned.
- Work at your own pace, but refer to the study schedule on Moodle to avoid last-day rush.
- If you require special arrangements (e.g., sudden illness), inform staff promptly. By default, we assume you have followed the study schedule up to the time of notification; if you contact us at the last moment, we assume you have had time to do all exercises before the force majeure reason.
- If you find the problems difficult:
  - Follow the study schedule. Solving tasks after corresponding lectures reduces difficulty and stress.
  - Attend lectures and read the recommended textbook chapters.
  - Participate in exercise workshops and ask for help in workshops or the Team.
  - Talk with fellow students.

Full details on peer review, grading expectations, and evaluation criteria for master's-level work are found in the separate Exercise and Peer Review Instructions file.

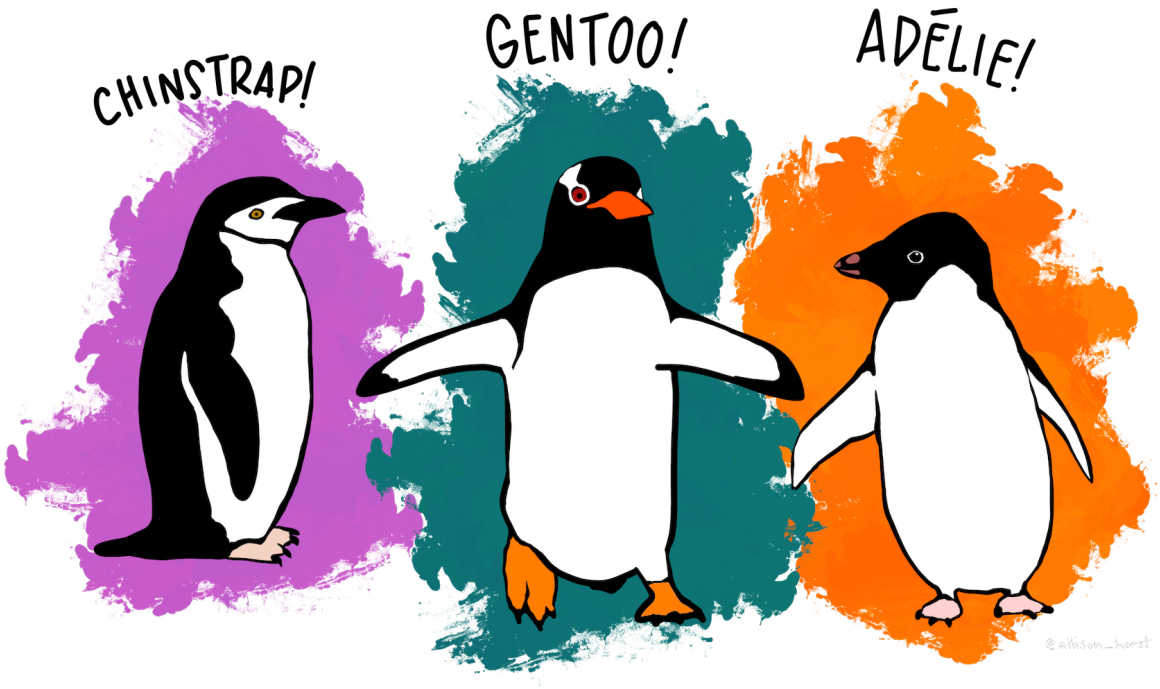
Problem 8

[5 points]

Topic: logistic regression, discriminative vs generative classifiers

In this problem, you should apply logistic regression (with an intercept term) to the Palmer Penguin dataset. In R, you can use `glm` to do the logistic regression; first, read through Sect. 4.7.2 of ISLR_v2. An alternative is `glmnet` from glmnetUtils library, which allows, e.g., for the Lasso and Ridge regularisation terms. In Python, I somehow prefer statsmodel implementation; see https://www.andrewvillazon.com/logistic-regression-python-statsmodels/ for a tutorial.

Palmer penguins dataset



Artwork by @allison_horst

- Dataset description.
- Use `penguins_train.csv` as your training data and `penguins_test.csv` as your testing data.
- Binary classification task: classify the penguin species as $y \in \{\text{Adelie}, \text{notAdelie}\}$ (`notAdelie` combining Gentoo and Chinstrap species) based on four morphological and weight measurements of the individual penguins, denoted by $\mathbf{x} = (x_1, x_2, x_3, x_4)^\top \in \mathbb{R}^4$

On performance measures

In this exercise set, you will train probabilistic classifiers which estimate $\hat{p}(y \mid \mathbf{x})$: the probability of class y given the covariate vector $\mathbf{x}$. A commonly used performance measure for classifiers is *accuracy*. We use the following notation:

- $y_i \in \{\text{Adelie}, \text{notAdelie}\}$: the true class of point i.
- $\hat{p}_i = \hat{p}(y = \text{Adelie} \mid \mathbf{x}_i)$: the estimated probability for the ith point in a dataset of size n being class Adelie.
- $\hat{y}_i$: the predicted class for point i which is $\hat{y}_i = \text{Adelie}$ if $\hat{p}_i \geq 0.5$ and $\hat{y}_i = \text{notAdelie}$ otherwise.

We define the *accuracy* on a (training or testing) dataset of n items as follows:

$$\text{accuracy} = \sum_{i=1}^{n} \mathrm{I}(y_i = \hat{y}_i)/n,$$

where $\mathrm{I}(z)$ is the indicator function which equals one if z is true and zero otherwise

$$\text{perplexity} = \exp\Big(-\sum_{i=1}^{n} \log \hat{p}(y = y_i \mid \mathbf{x}_i)/n\Big).$$

Perplexity is a transformation of the likelihood (perplexity $= \exp(-\text{loglikelihood}/n)$), which may be the most commonly used performance measure on probabilistic classifiers. Example values are perplexity $= 1$ for a perfect classifier, which always predicts the probability of one to an actual class, and perplexity $= 2$ for coin flipping, which has a predicted class probability $\hat{p} = 1/2$.

Task a

Fit logistic regression *without* regularisation on the training data and report the following:

- The model coefficients.
- The accuracy on the training and testing data.

Make a plot that shows the training dataset where the x-axis is the linear response $t_i = \beta^\top \mathbf{x}_i$ for the training data points and the y-axis is the probability $Pr(y_i = \text{Adelie} \mid \mathbf{x}_i)$ estimated by logistic regression. Plot the Adelie and notAdelie penguins differently, e.g., with a different glyph or colour (remember to explain which is which!).

Hint: In this task, with `sklearn.linear_model.LogisticRegression`, you should turn off regularisation (on by default) with the `penalty=None` option. You can also use R or `statsmodels`; see ISLR_v2 or ISLP, Section 4.7.2, for guidance.

Task b

Now fit the logistic regression model *with* Lasso regularisation and try to improve the results from task a. You can do this by trying various values; there is no need to be more sophisticated here. Similarly report

- The model coefficients
- The accuracy on the training and testing data

Make a plot similar to the one in task a for the regularised model.

Note that your solution must contain at least one zero coefficient.

Hint: For this task, you can use `sklearn.linear_model.LogisticRegression`, `statsmodels GLM.fit_regularised`, or the R `glmnet` and `glmnetUtils` libraries (the latter of which provides a formula interface).

Task c

In task a, the unregularised logistic regression may issue warnings (in R: `Warning: glm.fit: algorithm did not converge` or `Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred`). In Python, the warnings may look different depending on the implementation, or if you don't get warnings, you may have unexpectedly large regression coefficients. Why does R give warnings?

Problem 9

[4 points]

Objective: generative Bayes classifier

In this problem, you will study the quadratic discriminant analysis (QDA) classifier. Consider a simple case with two classes and only one feature ($K = 2$ and $p = 1$).

Prove that the QDA classifier is *not* linear if the class-specific variances differ ($\sigma_1^2 \neq \sigma_2^2$).

Hint: This problem is from the textbook (Problem 3, page 189). Please see the discussion in the textbook for hints and guidance. For this problem, you should follow the arguments laid out in Sect. 4.4.1 of the textbook, but without assuming that $\sigma_1^2 = \sigma_2^2$.

Problem 10

[6 points]

Objective: naive Bayes classifier

In this problem, you will study the Palmer penguins by building your Naive Bayes classifier. Please read the description of the dataset and performance measures from Problem 8.

Naive Bayes (NB) classifier

Your task is to build your own NB classifier; you should not use a ready-made classifier from a library. However, you do not need to create a generic classifier (such as `naiveBayes` in the R `e1071` library); it is enough that your classifier works for this particular task.

The idea of NB is that the dimensions are conditionally independent, given the class. Each class conditional feature distribution $p(x_i \mid y)$ is assumed to originate from an independent Gaussian distribution with its mean μ_{iy} and variance σ^2_{iy} for $i = 1, 2, 3, 4$.

Task a

Compute and report each attribute's means and standard deviations separately in the training set for both classes.

Estimate and report the class probabilities using Laplace smoothing with a *pseudocount* of 1 on the training set.

(You should produce a total of 18 numbers from this task.)

Task b

Now, you can find the class-specific expressions for $p(\mathbf{x} \mid y)$ needed by the NB classifier. Remember that according to NB assumption, the dimensions are independent, and hence, you can represent the class-specific $p(\mathbf{x} \mid y)$ likelihoods as products of 4 1-dimensional normal distributions.

Write down the formula needed to compute the posterior probability of the class being `Adelie` $\hat{p}(y = \text{Adelie} \mid \mathbf{x})$ as a function of the four measurements in $\mathbf{x}$ and the statistics (means, standard deviations, class probabilities) you computed in the task a above.

Task c

Using the formula you derived in Task b, compute and report your classifier's classification accuracy on the test set. Additionally, calculate and report the probabilities $\hat{p}(y = \text{Adelie} \mid \mathbf{x})$ for the three first penguins in the test set.

Hint: When computing classification accuracy, you can use the following rule to obtain "hard" classes:

$$\hat{y} = \begin{cases} \text{Adelie} & , \quad \hat{p}(y = \text{Adelie} \mid \mathbf{x}) \geq 0.5 \\ \text{notAdelie} & , \quad \hat{p}(y = \text{Adelie} \mid \mathbf{x}) < 0.5 \end{cases}$$

Problem 11

[4 points]

Objective: Understanding discriminative vs generative learning.

Download the reference below. You **do not need to read the full paper** or understand all the details! Instead, try to find the answers to the following questions.

Reference: Ng, Jordan (2001) On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. NIPS. http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf

Task a

Read the *Abstract* and *Introduction* (Sect. 1). According to the authors, is discriminative learning better than generative learning? Justify your answer.

Task b

By a "parametric family of probabilistic models", the authors mean a set of distributions where a group of parameters defines each distribution. An example of such a family is the family of normal distributions where the parameters are μ and Σ.

Ng and Jordan denote by h_{Gen} and h_{Dis} two models are chosen by optimising different objectives. Which two families do the authors discuss, and what are the (h_{Gen}, h_{Dis}) pairs for those models? What objectives are being optimised?

Task c

Study Figure 1 in the paper. Explain what it suggests (see the last paragraph of the Introduction). Reflect on what this means for the families in Task b.

Problem 12

[5 points]

Objective: comparing classifiers on synthetic data, application of different classifiers

In this problem, you will compare different classifiers using synthetic toy data sets.

Section 4.7.2 (Logistic regression) and 4.7.5 (NB) of ISLR_v2 (or ISLP) contain helpful information for solving this problem.

Toy data sets

We have generated ten training data sets of different sizes, `toy_train_<n>.csv` for $n \in \{2^3, 2^4, \ldots, 2^{12}\}$, and one test data set `toy_test.csv` with 10000 points.

Each toy data set has a binary class variable $y \in \{0, 1\}$ and two real-valued features $x_1, x_2 \in \mathbb{R}$.

The data are generated from the "true" model as follows:

- x_1 and x_2 are sampled from a normal distribution with zero mean and unit variance.
- The probability of y is given by:

$$p(y = 1 \mid x_1, x_2) = \sigma(0.1 - 2x_1 + x_2 + 0.2x_1x_2),$$

  where $\sigma(t) = 1/(1 + e^{-t})$ is the standard logistic function.

Task a

Is the Naive Bayes (NB) assumption valid for the toy data set? Explain why or why not.

Task b

For each training set, train several classifiers that output probabilities (described below) and then report their accuracy and perplexity on the test set (see Problem 8 for the description of perplexity - it is a transformed variant of log-loss).

Produce the following table (or make a plot) for accuracy and perplexity on the test set:

n	NB	LR	LRi	OptimalBayes	Dummy
8	?	?	?	?	?
16	?	?	?	?	?
32	?	?	?	?	?
64	?	?	?	?	?
128	?	?	?	?	?
256	?	?	?	?	?
512	?	?	?	?	?
1024	?	?	?	?	?
2048	?	?	?	?	?
4096	?	?	?	?	?

where the columns correspond to:

- Naive Bayes (NB) (e.g., `naiveBayes` from the library `e1071`)
- Logistic regression without an interaction term (e.g., `glm`)
- Logistic regression with an interaction term (e.g., `glm`)
- Optimal Bayes classifier that uses the actual class conditional probabilities (that you know in this case!) to compute $p(y \mid x_1, x_2)$ for a given (x_1, x_2) - no probabilistic classifier can do better than this

- "Dummy classifier" that does not depend on $\mathbf{x}$. It always outputs the probability $\hat{p}(y = 1 \mid x_1, x_2)$ as the fraction of $y = 1$ in the training data. "Dummy" means that the classifier output does not depend on the covariates. Including a dummy classifier in your comparison is always a good idea! One way to get a dummy classifier here is to train a logistic regression with only the intercept term.

Task c

Report the logistic regression coefficients with interaction terms for the largest training data set. How do they compare with the coefficients of the actual model that generated the data?

Discuss your observations and what you can conclude.

- Which of the models above are probabilistic, discriminative, and generative?
- How do accuracy and perplexity (log-likelihood) compare?
- Is there a relation to the insights from the previous problem?
- Why does logistic regression with the interaction term perform so well for larger datasets?
- Does your dummy classifier ever outperform other classifiers, or do different classifiers outperform the optimal Bayes classifier?

Instructions

Creating a function that takes the training and testing data as input and outputs the probabilities $p(y = 1 \mid x_1, x_2)$ for the testing data points is helpful. By using these probabilities, you can quickly compute accuracy and perplexity.

In R, you can include the interaction term in logistic regression by writing ("*" implies that interaction should be included in the model instead of "+", which assumes only additive effects):

```r
# R
model <- glm(y ~ x1 * x2, data[[4]], family = "binomial")
```

In Python, you can create interaction terms with, e.g., sklearn.preprocessing.PolynomialFeatures (set `interaction_only=True`).

```python
# Python
from sklearn.preprocessing import PolynomialFeatures

create_inter = PolynomialFeatures(degree = 2, interaction_only = True)
data3_inter = create_inter.fit_transform(data[3].iloc[:, :2])
GaussianNB().fit(data3_inter, data[3]["y"])
```

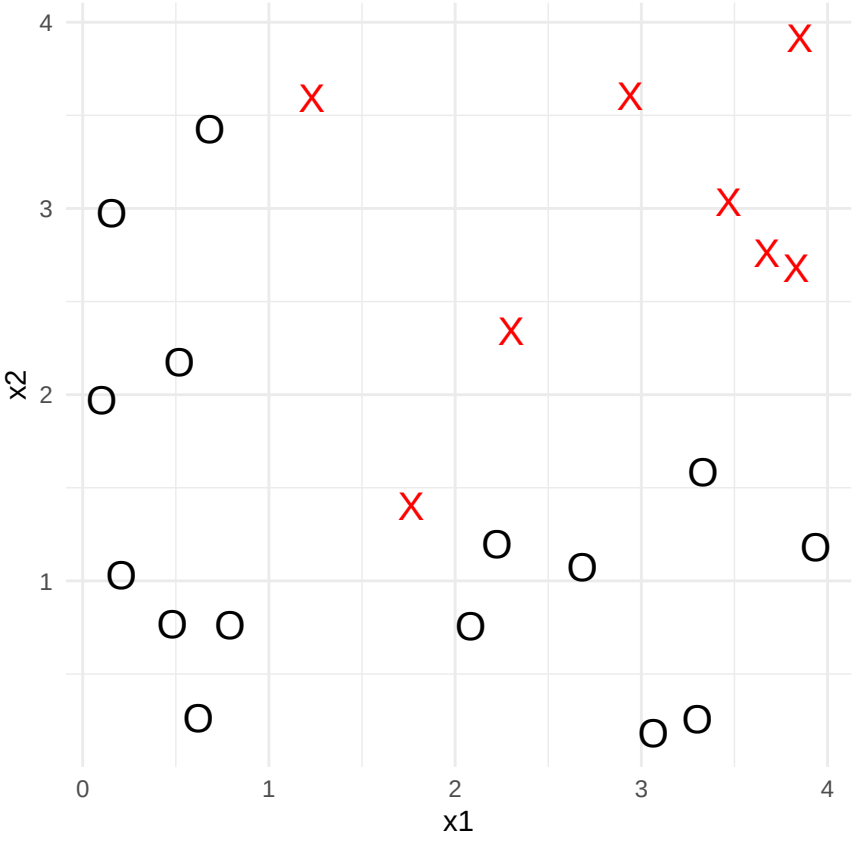
Problem 13

[5 points]

Objectives: basic principles of decision trees

In this task, you will simulate a decision tree algorithm by hand using the toy data shown in the figure.

Read Section 8.1 of ISLR_v2. Use the Gini index of Equation (8.6) as an impurity measure.



Task a

Sketch a run of the classification tree algorithm on the toy data and draw the resulting classification tree. For each split, report the Gini index value. Try to select the splits that obtain the best impurity measure.

(You do not need to worry about overfitting here: the resulting classification tree should have enough splits to fit the training data without error. Don't worry if your results are not optimal or super-accurate, as long as they are "in the ballpark".)

Problem 14

[3 points]

Learning objectives: basics of the k-NN method.

In this task, you will apply the *k-nearest neighbour* (k-NN) classifier by hand on a toy data set. You should be able to do this with pen and paper.

We will use the training dataset $D = \{(x_i, c_i)\}_{i=1}^{14}$, shown below, where $x_i \in \mathbb{R}$ are the covariates and $c_i \in \{-1, +1\}$ are the classes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	0.0	2.0	3.0	5.0	6.0	8.0	9.0	12.0	13.0	15.0	16.0	18.0	19.0	21.0
c_i	+1	+1	-1	+1	+1	+1	+1	-1	-1	+1	-1	-1	-1	-1

Task a

Where are the classification boundaries for the 1-NN and 3-NN classifiers? What are the respective classification errors on the training dataset?

Task b

How does the choice of k in k-NN affect the classification boundary (not in the above example but in general)? Give examples of the behaviour for extreme choices (very small or large k).

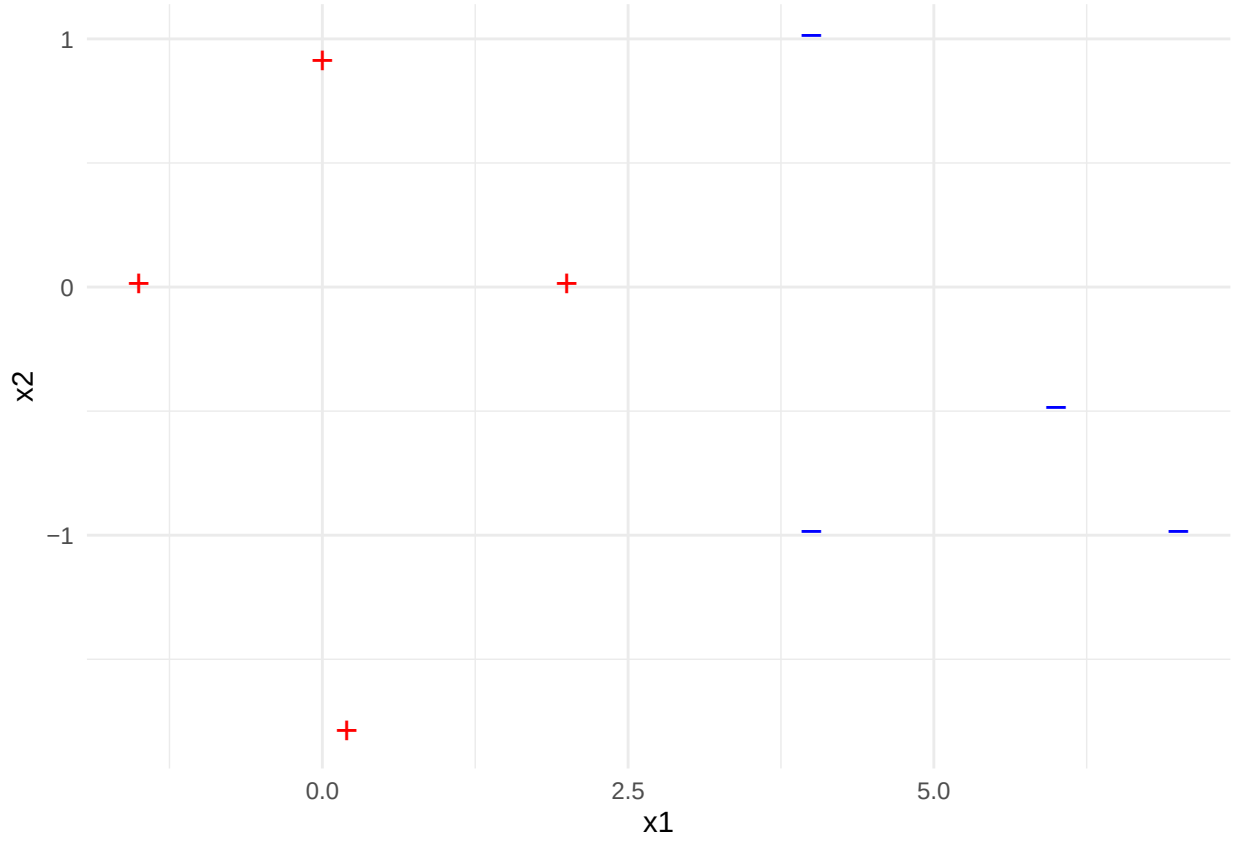
Problem 15

[6 points]

Topic: SVM

In this problem, you will study the support vector machine (SVM) classifier. Task a is on the toy data set below.

A linear decision boundary in 2 dimensions takes the form $\beta_0 + \beta_1 X + \beta_2 X_2 = 0$; we investigate a non-linear decision boundary in tasks b, c, and d.



	x1	x2	class
A	-1.5	0.0	1
B	0.0	0.9	1
C	0.2	-1.8	1
D	2.0	0.0	1
E	4.0	1.0	-1
F	4.0	-1.0	-1
G	6.0	-0.5	-1
H	7.0	-1.0	-1

Task a

Find a separating hyperplane with the largest margin on the data set above. Write down the equation for this hyperplane and report the margin size.

Which of the points (A–H) are support vectors?

Hint: You can answer without mathematical proof. You can do it simply by geometric intuition.

Task b

Sketch the curve $(1+X_1)^2+(2-X_2)^2=4$. On your sketch, indicate the set of points for which $(1+X_1)^2+(2-X_2)^2>4$, and the points for which $(1+X_1)^2+(2-X_2)^2\leq 4$.

Task c

Suppose a classifier assigns an observation to the blue class if $(1+X_1)^2+(2-X_2)^2>4$, and to the red otherwise. To what class are the observations $(0,0)$, $(-1,1)$, $(2,2)$, $(3,8)$ classified?

Task d

Argue that while the decision boundary of the above classifier is not linear in X_1 and X_2, it is linear in X_1, X_1^2, X_2, X_2^2.

Problem 16

[2 points]

Objectives: self-reflection, giving feedback on the course

Task a

- Write a learning diary of the topics of lectures 5-8 and this exercise set.

Instructions

Guiding questions: What did I learn? What did I not understand? Was there something relevant to other studies or (future) work? Your reply should be 1-3 paragraphs of text. You can also give feedback on the course.

Task b

Give an estimate of the hours used to solve the problems in this exercise set.