Statistical Machine Learning: Exercise 1

Refresher: Linear Algebra, Probabilities and Bayesian Decision Theory

Prof. Stefan Roth, Dr. Simone Schaub-Meyer

TECHNISCHE UNIVERSITÄT DARMSTADT

Summer Term 2021

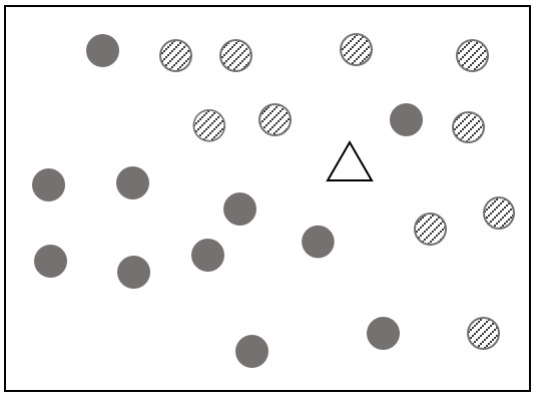
Due date: May 25th, 23:59 PM

Note: Many of the concepts required for solving this homework will be introduced in lectures 1, 2, 3, and 4. If you don't know how to do some of these problems by the time this is released, hold on a bit more for that specific problem. The recoding for lecture 4 will be released no later than May 10, 2021.

Task 1: Machine Learning Introduction (6 Points)

1a) Model Fitting (6 Points)

Assume you have labelled data points as shown in the Figure below. The circles are your training set and the triangle is your test set. Describe a model for both possible outcomes for the label of the triangle. Explain how these models differ in their training and testing accuracy, assuming that the correct label of the triangle is 'striped'.



Task 2: Linear Algebra Refresher (20 Points)

2a) Matrix Properties (5 Points)

A colleague of yours suggests that for matrices, addition and multiplication works similarly as for scalars, in particular, that the commutative, distributive and associative properties hold. Are these statements correct? Prove or disprove them analytically for both operations and all three properties considering three matrices A, B, C of size $n \times n$.

2b) Matrix Inversion (7 Points)

Given the following matrix

$$A = \begin{bmatrix} 1 & a & b \\ 1 & c & d \\ 0 & 0 & 1 \end{bmatrix},$$

analytically compute its inverse A^{-1} and illustrate the steps. What algorithm did you use? For which values of a, b, c, d is A invertible?

If we change the matrix to

$$A = \begin{bmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 8 & 3 & 12 \end{bmatrix},$$

is it still invertible? Why or why not?

2c) Matrix Pseudoinverse (3 Points)

Write the definition of the right and left Moore-Penrose pseudoinverse of a generic matrix $A \in \mathbb{R}^{n \times m}$.

Given $A \in \mathbb{R}^{2 \times 3}$, which one does exist? Write down the equation for computing it, specifying the dimensionality of the matrices in the intermediate steps.

2d) Basis Transformation (5 Points)

We are given the basis $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $b_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, b_2 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

1) Compute the transformation matrix T such that $Tv = w$ for a vector v in basis e and w in basis b.

1) What is the coordinate vector of $v = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$ in the basis b?

Task 3: Statistics Refresher (29 Points)

3a) Expectation and Variance (8 Points)

Let Ω be a finite set and $P : \Omega \to \mathbb{R}$ a probability measure that (by definition) satisfies $P(\omega) \geq 0$ for all $\omega \in \Omega$ and $\sum_{\omega \in \Omega} P(\omega) = 1$. Let $f : \Omega \to \mathbb{R}$ be an arbitrary function on Ω.

1) Write the definition of expectation and variance of f and discuss if they are linear operators.

2) You are given a set of three dice $\{A, B, C\}$. The following table contains results obtained from rolling these three standard, six-sided dice 18 times each. Each row corresponds to one die, each column shows how often a number was rolled on the respective die.

	1	2	3	4	5	6
A	1	5	6	3	2	1
B	6	1	1	4	1	5
C	3	2	3	3	4	3

Estimate the expectation and the variance for each die using unbiased estimators. (Show your computations.)

3) Estimate the KL-divergence between each die's distribution and the uniform distribution. According to your results, which one of them is closest to a fair, uniform die?

3b) It is a Cold World (7 Points)

Consider the following three statements:
a) A person with a cold has back pain 25% of the time.
b) 4% of the world population has a cold.
c) 10% of those who do not have a cold still have back pain.

1) Identify random variables from the statements above and define a symbol for each of them.
2) Define the domain of each random variable.
3) Represent the three statements above with your random variables.
4) If you suffer from back pain, what are the chances that you suffer from a cold? (Show all the intermediate steps.)

3c) Cure the Virus (14 Points)

You are a brilliant scientist trying to create a cure to stop the dangerous virus C. To produce the needed molecules, you need to use the bacterium FC with a specific mutation HC-SML. Sadly, the mutation is unstable and each day, the bacterium has a 42% chance of replicating with the mutation. Luckily, the bacterium also has a 2.6% chance to develop the mutation when it previously did not have it.

1) We describe the current state at time t with vector $\vec{s}_t = [m,\, \tilde{m}]^T$ where m is the chance your vial currently contains the bacterium with the mutation and $\tilde{m}$ describes the probability it doesn't. Formalize the problem of computing the next state $\vec{s}_{t+1}$ using matrix notation (this is called a *Markov chain*). To simplify the problem, we assume that the whole culture either has the mutation or it does not, percentages of mutated bacteria do not have to be modeled.

2) You have a small culture with the mutated bacterium and feel close to saving humanity. You decide to make a simulation to predict the outcome of breeding the bacterium for several days. What is the probability that the mutation will still be present in your culture after 18 generations? Code a simulation in Python, run it, include a snippet of your code, and attach a plot with the probability after each generation. Will your brilliant plan succeed?

3) After starting another project, you forget about your culture for a while and remember it weeks later. You wonder whether there is still a chance that the mutation is present.

Run your simulation for as long as it is necessary to obtain a stable probability. How likely is it that the culture contains the mutation in the long run and after how many timesteps do the ratios stop to change significantly? In addition, find a mathematical expression for a stable probability. Present the equation and the solution. Does the probability actually converge to this solution?

Task 4: Information Theory (5 Points)

4a) Entropy (5 Points)

You work for a telecommunication company that uses a system to transmit four different symbols S_1, S_2, S_3, S_4 over a channel. In the current system, each symbol has a probability to occur according to the following table:

	S_1	S_2	S_3	S_4
p_i	0.04	0.22	0.67	0.07

1) How many bits of information can be transmitted on average per symbol under this distribution? Compute the entropy. **2)** In general, what is the maximum number of bits per symbol that can be transmitted using a set of four symbols? Which distribution over the symbols is required to achieve this? [1]

[1] For a quick introduction to information theory, consider reading section 1.6 of Bishop's book.

Task 5: Bayesian Decision Theory (20 Points)

In this exercise, we consider data generated by a mixture of two Gaussian distributions with parameters $\{\mu_1, \sigma_1\}$ and $\{\mu_2, \sigma_2\}$. Each Gaussian represents a class labeled C_1 and C_2, respectively.

5a) Optimal Boundary (4 Points)

Explain in one short sentence what Bayesian Decision Theory is. What is its goal? Consider the case of two classes C_1 and C_2. Which condition holds at the optimal decision boundary? When do we decide for class C_1 over C_2?

5b) Decision Boundaries (8 Points)

If both classes have equal prior probabilities $p(C_1) = p(C_2)$ and the same variance $\sigma_1 = \sigma_2$, derive the decision boundary x^* analytically as a function of the two means μ_1 and μ_2.

5c) Different Misclassification Costs (8 Points)

Assume $\mu_1 > 0$, $\mu_1 = 2\mu_2$, $\sigma_1 = \sigma_2$ and $p(C_1) = p(C_2)$. If misclassifying sample $x \in C_2$ as class C_1 is four times more expensive than the opposite, how does the decision boundary change? Derive the boundary analytically. (There is no cost for correctly classifying samples.)