Prerequisite Knowledge Test

DATA11002 Introduction to Machine Learning (Autumn 2025)

Please solve all five problems and return your solution as a PDF file in the course Moodle area. Number your answers with the corresponding problem numbers and present the solutions in the same order.

We will grade this test *pass* or *fail*.

- You pass if you pass all problems.
- If you submit your answer by the deadline and fail, we will give you a new deadline to resubmit your answer. You **must** eventually pass to continue with the course.

This prerequisite knowledge test is a home examination you must do *individually*. It is *not allowed* to:

- Co-operate or discuss solutions with others.
- Copy ready-made answers.
- Use Generative AI, such as ChatGPT, to produce the answers.

It is allowed to use external sources, including web searches.

The purpose of this examination is to test and inform you about the prerequisite knowledge needed for the course:

- If you have the required prerequisite knowledge, these problems should be relatively easy and fast to complete.
- If you have substantial difficulties with some problems, you may need extra work and individual self-studies during this and subsequent machine learning courses.

The solutions should all be short. Don't be frightened by the notation or length of the problems. The problems are (unnecessarily) verbose because of lengthy hints and pointers to additional material.

How to prepare your solution

You must produce PDF solutions (such as for this test) during the course. I recommend using RStudio Desktop and R Markdown, designed to make, e.g., PDF documents, to prepare your solutions. R Markdown supports, e.g., LaTeX math notation and several output formats, including PDF, and languages including R and Python/SciPy. This file has been made using R Markdown! You can familiarise yourself with R Markdown by going through a brief tutorial course at https://rmarkdown.rstudio.com/lesson-1.html, after which you can use the book by Xie et al. for reference.

Problems

Problem 1

Topic: fundamental data analysis, software tools

A mystery data set in file `x.csv` has 2048 data items (rows), each having 32 real-valued variables (columns). The first row in the file gives the names of the variables.

Task a

Write a small program in R or Python that

1. loads the data set in `x.csv`,
2. finds the two variables having the largest variances and
3. makes a scatterplot of the data items using these two variables.

Your program must read the dataset file from the file and then produce the plot without user intervention.

Your program must work correctly for any dataset file of a similar format. For example, if you permute the rows or columns of the data file, you should get the same output because ordering rows or columns should not affect the result.

Attach a printout of your program code and the scatterplot it produces as an answer to this problem.

Hints

You should see two letters in the scatterplot from which it should be evident that you did okay. If you see something else, then something is wrong.

About the topic

The course will contain examples in Python and R. While you are not required to know any specific programming language beforehand, you should, in this course and your future career, have sufficient background to be able to learn to do basic independent data analysis operations in any of the commonly used programming languages. Please see the public course home page for the software needed during the course.

Reading material: OP, An introduction to R.

Prerequisite courses: TKT10002 Introduction to Programming or FYS1013 Scientific computing. Recommended: programming experience or TKT10003 Advanced course in programming or FYS2085 Scientific computing II.

Problem 2

Topic: matrix calculus

Let $\mathbf{A}$ be a 2×2 matrix given by $\mathbf{A}_{11} = 1$, $\mathbf{A}_{12} = \mathbf{A}_{21} = 2$, and $\mathbf{A}_{22} = 1.618$.

Task a

For the matrix $\mathbf{A}$:

1. Solve numerically and report the eigenvalues λ_i and column eigenvectors $\mathbf{x}_i$, where $i \in \{1, 2\}$. Normalise the eigenvectors to unit length (if necessary).
2. Verify that the eigenvectors are orthonormal.
3. Show, by performing the numerical matrix computation, that $\mathbf{A}$ satisfies the equation $\mathbf{A} = \sum_{i=1}^{2} \lambda_i \mathbf{x}_i \mathbf{x}_i^\intercal$.

Hints

You can use R or Python to find eigenvectors and eigenvalues and do matrix and vector multiplications.

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric square matrix. $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\mathbf{A}$ and a column vector $\mathbf{x} \in \mathbb{R}^n$ is the corresponding *eigenvector* if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Assume $\mathbf{A}$ has n orthonormal eigenvectors $\mathbf{x}_i \in \mathbb{R}^n$ and corresponding eigenvalues $\lambda_i \in \mathbb{R}$, where $i \in \{1, \ldots, n\}$. The fact that the eigenvectors are orthonormal means that $\mathbf{x}_i^\intercal \mathbf{x}_i = 1$ and $\mathbf{x}_i^\intercal \mathbf{x}_j = 0$ if $i \neq j$.

About the topic

Vector/matrix operations and eigenvalues are prevalent in machine learning. The course uses them extensively.

Reading material: Chapters 2, 3, and 4 of MML and FCLA.

Prerequisite courses: MAT11009 Basics of mathematics in machine learning I or MAT11002 Linear algebra and matrices I or FYS1012 Mathematics for physicists III.

Other helpful courses: MAT21001 Linear algebra and matrices II, MAT22011 Linear algebra and matrices III, MAT21019 Applications in matrices.

Problem 3

Topic: algebra, probabilities, random variables

Let Ω be a finite sample space, i.e., the set of all possible outcomes. Let $P(\omega) \geq 0$ be the probability of an outcome $\omega \in \Omega$. The probabilities are non-negative, and they sum up to unity, i.e., $\sum_{\omega \in \Omega} P(\omega) = 1$. Let X be a real-valued *random variable*, i.e., a function $X : \Omega \to \mathbb{R}$ which associates a real number $X(\omega)$ with each of the (random) outcomes $\omega \in \Omega$.

The *expectation* of X is defined by $E[X] = \sum_{\omega \in \Omega} P(\omega) X(\omega)$. The *variance* of X is defined by $\mathrm{Var}[X] = E\left[(X - \mu)^2\right]$, where $\mu = E[X]$.

Task a

Using the definitions above, prove that E is a linear operator.

Task b

Using the definitions above, prove that the variance can also be written as $\mathrm{Var}[X] = E\left[X^2\right] - E[X]^2$.

Hints

An operator L is said to be linear if for every pair of functions f and g and scalar $t \in \mathbb{R}$, (i) $L[f+g] = L[f] + L[g]$ and (ii) $L[tf] = tL[f]$. The proof in task b is short if you use linearity.

About the topic

Random variables and expectations are central concepts in machine learning. The course uses them extensively.

Reading material: Chapter 6 of MML and PI.

Prerequisite courses: MAT11015 Basics of mathematics in machine learning II or MAT12003 Probability I or FYS1014 Statistical analysis of observations.

Other useful courses: MAT22001 Probability IIa.

Problem 4

Topic: conditional probabilities, Bayes rule

The conditional probability ("X given Y'') is defined by $P(X \mid Y) = P(X \wedge Y)/P(Y)$, where $P(\square)$ is the probability that $\square$ is true and X and Y are Boolean random variables that can have values of true or false, respectively. The marginal probability $P(Y)$ can also be written as $P(Y) = P(X \wedge Y) + P(\neg X \wedge Y)$, where $\neg X$ denotes logical negation.

Task a

Derive Bayes' rule $P(X \mid Y) = \frac{P(Y|X)P(X)}{P(Y)}$ by using the definition of conditional probability.

Task b

A medical test for detection of pollen allergy behaves as follows (Nevis et al. 2016):

- (i) for persons that don't have the allergy, the test gives a (false) positive in 23% of the cases, and
- (ii) for persons with the allergy, the test gives a (false) negative in 15% of the cases.

According to statistics, 20% of the population in Finland suffers from pollen allergy.

Define suitable Boolean random variables, write down the equation (in terms of the three percentages mentioned above), and compute the value for the probability that a person is allergic to pollen if we have chosen the person to test at random and the test result is positive.

Hints

You can solve task b by using Bayes' rule, the definition of conditional probability, and the expression for marginal probability mentioned above.

About the topic

Bayes' rule and conditional probability provide the theoretical basis for probabilistic classifiers.

Reading material: Same as Problem 3.

Prerequisite courses: Same as Problem 3.

Problem 5

Topic: optimisation

Assume you are given constants $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$, where $i \in \{1, \ldots, n\}$, and a function $f(b) = \sum_{i=1}^{n} (bx_i - y_i)^2 / 2$.

Task a

By using derivatives, find the value of $b \in \mathbb{R}$ that minimises the value of $f(b)$.

Task b

What conditions must the constants x_i and y_i satisfy for the function to have a unique and finite minimum?

About the topic

Optimisation is an essential tool in machine learning. For example, we try to find model parameters that minimise a given loss in model selection. Differentiation is a common way to do optimisation.

Reading material: Chapter 7 of MML, Section 5 of MPK, and Section MAA6 of KO.

Prerequisite courses: high school mathematics.

Other useful courses: MAT11015 Basics of mathematics in machine learning II or FYS1010 Mathematics for physicists I.

References cited

- [MML] Deisenroth et al. (2020) Mathematics for Machine Learning. Cambridge University Press. https://mml-book.com
- [MPK] Häkkinen (2006) Matematiikan propedeuttinen kurssi. Jyväskylän yliopisto. http://www.math.jyu.fi/matyl/propedeuttinen/kirja/
- [KO] Kisallioppiminen.fi
- [FCLA] Beezer (2016) A First Course in Linear Algebra. http://linear.ups.edu/
- [PI] Grimmet et al. (2014) Probability: an introduction, 2nd Ed. Oxford University Press.
- [OP] Ohjelmoinnin perusteet ja jatkokurssi, syksy 2020. https://python-s20.mooc.fi/