Exercise Set 0: Prerequisite Knowledge

DATA11002 Introduction to Machine Learning (Autumn 2021)

Please solve all of the six problems and return your solution as a single pdf file in the course Moodle area as instructed preferably before the first lecture on **3 November 2021** but *at latest* by **7 November 2021**. Number your answers with the corresponding problem numbers and present the solutions in the same order.

This exercise set will be graded *pass* or *fail*. You will pass this exercise set if you pass all of the six problems. If you fail this exercise set you will have a chance to resubmit your answer. You must however eventually pass this exercise set to be able to do the other study attainments and to pass the course.

This exercise set must be completed individually by one person. It is not allowed to co-operate with the others or copy ready-made answers. It is allowed to use external sources, including web searches.

The purpose of this exercise set is to test and inform you about the your prerequisite knowledge needed during the course. Each of the problems are designed to cover separate areas of prerequisite knowledge that are needed during the course. If you have the required prerequisite knowledge the problems in this exercise set should be relatively easy and fast to complete. If you have substantial difficulties with some of the problems it means that you may need some extra work and individual self-studies during this and subsequent machine learning courses.

Notice that the solutions to the problems should all be quite short: don't be frightened by notation or length of the problems. The problems are (unnecessarily) verbose, because, e.g., of lengthy hints and pointers to additional material about the topics covered.

Problems

Problem 1

Topic: basic data analysis, software tools

A mystery data set at https://kaip.iki.fi/local/x.csv (and in Moodle) has 1000 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 that (i) loads the data set in `x.csv`, (ii) finds the two variables having the largest variances, and (iii) makes a scatterplot of the data items by using these two variables. Your program must read the dataset file from the net or from your drive and then produce the plot without user intervention. Your program must work correctly for any dataset file of similar format (for example, if the rows or columns of the data file were permuted you should get the same output, because the ordering of rows or columns should not affect the result). Attach a printout of your program code and the scatterplot produced by your program as an answer to this problem.

Hints

You should see two letters in the scatterplot from which it should be obvious that you did ok. If you see something else then something went wrong.

About the topic

The course will contain examples in R and some in Python. While you are not required to know any specific programming languages beforehand, you should have the sufficient background to be able to independently learn to do basic data analysis operations in any of the commonly used programming languages.

Before the course, I recommend that you install the RStudio Desktop and the Anaconda Individual Edition. I recommend that you start by writing your answers to this exercise set by using, e.g., RStudio Desktop and R Markdown. R Markdown supports, e.g., LaTeX math notation, and several output formats, including pdf, and languages including R and Python (and SciPy). This file has been compiled using R Markdown! You can familiarize yourself with R Markdown by going through a brief tutorial course at https://rmarkdown.rstudio.com/lesson-1.html, see the book by Xie et al. for further information.

Reading material: OP, An introduction to R.

Prerequisite courses: TKT10002 (or AYTKT10002) Introduction to Programming or FYS1013 Scientific computing. Useful: programming experience or TKT10003 (or AYTKT10003) Advanced course in programming or FYS2085 Scientific computing II.

Problem 2

Topic: matrix calculus

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 $\mathbf{x}_i^T \mathbf{x}_i = 1$ and $\mathbf{x}_i^T \mathbf{x}_j = 0$ if $i \neq j$. Let's define a new matrix $\mathbf{B} = \sum_{j=1}^{n} \lambda_j \mathbf{x}_j \mathbf{x}_j^T$ (this is known as "spectral decomposition").

Task a

Show that λ_i and $\mathbf{x}_i$, as defined above, are eigenvalues and eigenvectors of the matrix $\mathbf{B}$ as well.

Task b

Let $\mathbf{A}$ be a 2×2 matrix given by $\mathbf{A}_{11} = 1$, $\mathbf{A}_{12} = \mathbf{A}_{21} = 2$, and $\mathbf{A}_{22} = 3$. For this matrix, solve numerically and report the eigenvalues λ_i and eigenvectors $\mathbf{x}_i$, where $i \in \{1, 2\}$ (normalize the eigenvectors to unit length, if necessary). Check that the eigenvectors are orthonormal. Show by performing the numerical matrix computation that the equation $\mathbf{A} = \sum_{j=1}^{n} \lambda_j \mathbf{x}_j \mathbf{x}_j^T$ is satisfied at least for this particular matrix.

Hints

In task a, you can also show that $\mathbf{A} = \mathbf{B}$ in general case if you want; this is however not required. In task b, you can, e.g., use R, Python, or Matlab to find eigenvectors and eigenvalues and do matrix and vector multiplications.

About the topic

The vector and matrix operations and eigenvalues are prevalent machine learning and you need to know the basics. In the course, matrix and vector calculus and eigenvalues are used extensively.

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

Prerequisite courses: MAT11009 Basics of mathematics in machine learning I or MAT11002 (or AY-MAT11002) Linear algebra and matrices I or FYS1012 (or AYFYS1012) Mathematics for physicists III.

Other useful 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 $P(\omega)$ 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 $\mathrm{Var}[X] = E\left[(X - m)^2\right]$, where $m = E[X]$.

Task a

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

Task b

Using the definitions above, show 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. You need to understand the concept of expectation and random variables, these are applied extensively during the course.

Reading material: Chapter 6 of MML and PI.

Prerequisite courses: MAT11015 Basics of mathematics in machine learning II or MAT12003 (or AY-MAT12003) Probability I or FYS1014 (or AYFYS1014) 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 be also written as $P(Y) = P(X \wedge Y) + P(\neg X \wedge Y)$, where $\neg X$ denotes logical negation.

Task a

Derive the Bayes rule $P(X \mid Y) = P(Y \mid X)P(X)/P(Y)$ by using the definition of conditional probability.

Task b

Medical test for detection of pollen allergy behaves as follows (Nevis et al. 2016): (i) for persons not having the allergy the test gives a (false) positive in 23% of the cases and (ii) for persons having the allergy the test gives a (false) negative in 15% of the cases. According to statistics, 20% of the population in Finland suffer 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 really allergic to pollen, if the test result is positive.

Hints

Task b can be solved by using the Bayes rule, the definition of conditional probability, and the expression for marginal probability mentioned above.

About the topic

The Bayes rule and conditional probability provides the theoretical basis, e.g., for probabilistic classifiers. The reading materials and the courses for the topics of this problem are the same as for the Problem 3.

Problem 5

Topic: optimization

Assume that you are given three constants $a \in \mathbb{R}$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$ and a function $f(x) = ax^4 + bx + c$.

Task a

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

Task b

What conditions must a, b, and c satisfy in order for the function to have a unique and finite minimum?

About the topic

One of the basic tools in machine learning is optimization, for example, in model selection we try to find model parameters which minimize a given loss. Differentiation is one of the most common ways to do the optimization.

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 (or AYFYS1010) Mathematics for physicists I .

Problem 6

Topic: algorithms

The Fibonacci numbers $F(i)$ are defined for $i \in \mathbb{N}$ recursively as $F(i+2) = F(i+1) + F(i)$, with $F(1) = F(2) = 1$.

Task a

Using pseudo-code, write down an algorithm that takes $n \in \mathbb{N}$ as an input and outputs the Fibonacci numbers from $F(1)$ to $F(n)$.

Task b

What is the time complexity of your algorithm, expressed by using the O-notation?

About the topic

The mathematics of machine learning (Problems 2-5) are usually in practice applied with programs that implement various algorithms and therefore understanding the basics of algorithmics is essential. You should at least be able to understand and write pseudocode and analyse runtime and memory usage of algorithms using the big O notation.

Reading material: TIRA (Secs. 1-2) and IA (Secs. 2-3).

Useful courses: TKT20001 Data Structures and Algorithms or AYTKT200011 Data structures and algorithms I. (Full content of these courses is not needed! Especially useful for this course is the analysis of time and space complexity with the big O notation, see, e.g., Sec. 2 of TIRA.)

Prerequisite knowledge

The following studies should give you the necessary prerequisite knowledge to participate to this course:

- Generic skills learned during BSc studies (including scientific writing addressed, e.g., in the BSc thesis).
- Basic mathematics skills (Problems 2-5) which are covered in:
  - high school mathematics *or* AYMFK-M101A Lukiomatematiikan kertaus; and
  - MAT11001 Introduction to university mathematics *or* AYMAT11001 Johdatus yliopistomatematiikkaan *or* FYS1010 Mathematics for physicists I *or* AYFYS1010 Matemaattiset apuneuvot I; and
  - MAT11009 Basics of mathematics in machine learning I *or* MAT11002 Linear algebra and matrices I *or* AYMAT11002 Lineaarialgebra ja matriisilaskenta I *or* FYS1012 Mathematics for physicists III *or* AYFYS1012 Matemaattiset apuneuvot III; and
  - MAT11015 Basics of mathematics in machine learning II *or* MAT12003 Probability I *or* AYMAT12003 Todennäköisyyslaskenta I *or* FYS1014 Statistical analysis of observations *or* AYFYS1014 Havaintojen tilastollinen käsittely.
- Basic programming skills (Problem 1) which are covered in:
  - TKT10002 Introduction to programming *or* AYTKT10002 Ohjelmoinnin perusteet *or* FYS1013 Scientific computing I.

Programming experience will always be useful, especially if you want to continue to apply and/or study machine learning. Good courses to prep your programming skills include TKT10003 Advanced course in programming *or* AYTKT10003 Ohjelmoinnin jatkokurssi *or* FYS2085 Scientific computing II.

Additionally, it is useful to know the basic ideas of pseudocode and the analysis of time and space complexity with big O notation (Problem 6) to the level discussed, e.g., in Secs. 1-2 of TIRA or Secs. 2-3 of IA. These topics (plus lots of other topics not needed in this course!) are covered in:

- TKT20001 Data structures and algorithms *or* AYTKT200011 Data structures and algorithms I.

Notice that there are many other ways (in addition to the courses listed above) to obtain the necessary prerequisite knowledge. If you are comfortable doing the six problems of this exercise set then you probably have the necessary prerequisite knowledge. The prerequisite knowledge that is relevant for each of the problems is described more in detail below in conjunction with the individual problems, under the subtitles "about the topic". 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/
- [TIRA] Laaksonen (2020) Tietorakenteet ja algorithmit. https://www.cs.helsinki.fi/u/ahslaaks/tirakirja/
- [IA] Cormen et al. (2009) Introduction to Algorithms, 3rd Ed.. MIT Press.