ATM4171 Introduction to Machine Learning for Atmospheric and Earth System Research
Spring 2020
Exercise Set 0: Prerequisite Knowledge

Please solve all of the six problems and return your solution as a single pdf file in the course Moodle area (`https://moodle.helsinki.fi/course/view.php?id=37817`) on **1 May 2020**, at latest. Number your answers with the corresponding problem numbers and present the solutions in the same order. If it is impossible for you to use the Moodle for some reason: please *by the same deadline* (i) attach the answer pdf file to an email to the course mailbox (`ml-inar2020@helsinki.fi`), (ii) include string "`Exercise Set 0`" to the subject line of the email, and to the body of the email (iii) your name and (iv) student number.

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 pass the course. This exercise set must be completed individually by one person.

The purpose of this exercise set is to test your prerequisite knowledge 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.

If you want to study further the mathematics that underlies machine learning: a good source to study the topic—which goes beyond the level required in this course—is the new book "Mathematics for Machine Learning" (MML), which is available as a pdf file from `https://mml-book.com`. A possible online source for basic linear algebra (Prob. 3) is the book "Introduction to Applied Linear Algebra" (IALA), available at `http://vmls-book.stanford.edu/`, or any other linear algebra text book. Finnish speakers may find "Matematiikan propedeuttinen kurssi" (MPK) about high school mathematics useful to prep some basic math skills, available at `http://www.math.jyu.fi/matyl/propedeuttinen/kirja/`.

We plan to provide those who need it a brief tutorial of the prerequisite topics during May, before the course starts.

1. (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 $m = E[X]$. Using these definitions, show that:

   (a) E is a linear operator.

(b) The variance can also be written as $\mathrm{Var}[X] = E\left[X^2\right] - E\left[X\right]^2$.

(Hint: 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 of 1b is short if you use linearity. Further reading: You can, e.g., browse the first sections of Chapter 6 of MML for an introduction about probabilities.)

2. (CONDITIONAL PROBABILITIES, BAYES RULE) The conditional probability ("X given Y") is defined by $P(X \mid Y) = P(X \wedge Y)/P(Y)$, where $P(\Box)$ is the probability that $\Box$ 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.

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

   (b) Medical test for detection of pollen allergy behaves as follows: (i) for persons not having the allergy the test gives a (false) positive in 1 % of the cases and (ii) for persons having the allergy the test gives a (false) negative in 10 % of the cases. According to statistics, 15 % 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.

(Hint: Item 2b can be solved by using the Bayes rule, the definition of conditional probability, and the expression for marginal probability mentioned above.)

3. (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 $x \in \mathbb{R}^n$ is the corresponding *eigenvector* if $\mathbf{A}x = \lambda x$. Assume $\mathbf{A}$ has n orthonormal eigenvectors $x_i \in \mathbb{R}^n$ and corresponding eigenvalues $\lambda_i \in \mathbb{R}$, where $i \in \{1, \ldots, n\}$. The fact that the eigenvectors are orthonormal means that $x_i^T x_i = 1$ and $x_i^T x_j = 0$ if $i \neq j$. Let's define a new matrix $\mathbf{B} = \sum_{j=1}^n \lambda_j x_j x_j^T$ (this is known as "spectral decomposition").

   (a) Show that λ_i and x_i, as defined above, are eigenvalues and eigenvectors of the matrix $\mathbf{B}$ as well.

(Hint: If you want, you can also show that $\mathbf{A} = \mathbf{B}$; this is however not required here. Further reading: You can read, e.g., Chapter 2 of MML or Chapter 6 of IALA.)

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

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

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

(Further reading: See Sec. 5.1 of MML for introduction to differentiation of univariate functions. MPK, Secs. 5.4–5.5, also covers the math needed here.)

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

   (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 $F(n)$.

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

6. (BASIC DATA ANALYSIS, SOFTWARE TOOLS) A mystery data set at `http://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.

   (a) Write a small program in Python or 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. Attach a printout of your program code and the scatterplot produced by your program as an answer to this problem.

   (Hint: 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. Other comments: The course will contain examples in R and 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.)