Fall 2022 *Instructor: Ashwin Pananjady*

7750: Mathematical Foundations of Machine Learning

Linear algebra and probability for data analysis

Homework 0

Released: Aug 21 *Due: Aug 26, 11:59pm ET*

Note: This homework will not be graded, but you are expected to submit solutions in the format detailed in the syllabus. Don't worry about "complete" solutions per se; you will receive 2/50 points alloted to your HW score as bonus just for a submission that follows all the instructions.

Objective. To solve some problems in linear algebra, calculus, and probability, and to set up and use Jupyter notebooks. The problems are deliberately dry with keywords, so that you can quickly find resources to help fill any gaps in your knowledge. Do not be worried if you cannot solve some problems; the primary point is to give you the opportunity to go back and refresh. We will post detailed solutions to these problems after the deadline, and also discuss them briefly during lecture.

Resources. The following resources may be helpful to get up to speed if you feel yourself struggling with portions of this:

1. "Essence of linear algebra" playlist on Youtube by 3Blue1Brown.
2. A handbook of mathematics for ML, by Garrett Thomas `https://gwthomas.github.io/docs/math4ml.pdf`. Not all of this is really required, but it has a nice exposition of basics in calculus, linear algebra, and probability, and can be used as a handbook over the course of the semester.
3. Matrix cookbook `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`. Again, most things here will not be used, but this is a useful handbook.
4. Trefethen and Bau `http://people.maths.ox.ac.uk/~trefethen/text.html`. These contain some linear algebra notes that are relatively advanced. We will use it later on to discuss a nice geometric exposition of the singular value decomposition.
5. For Python, there are a variety of blogs that will be thrown up just by a Google search "Python for machine learning beginner". You should feel free to use whatever you find convenient to learn basic syntax. Use Problem 3 for a recommended installation of Jupyter notebooks (if you don't have a working version already).

Notation: Capital boldface letters will be matrices, and small boldface letter will be vectors. Capital letters (not boldface) will be random variables (and sometimes random vectors), and small letters will typically be scalars. Dimensions of matrices and vectors will be specified when needed, but for the most part, you should be able to intuit these yourself (and doing this is a useful exercise). The set of real numbers is represented by the set $\mathbb{R}$, d-dimensional vectors by $\mathbb{R}^d$ and $n \times d$ matrices by $\mathbb{R}^{n \times d}$. The symbol $:=$ denotes a definition; the left-hand-size is defined to be the right-hand-side.

Problem 1 (Calculus and linear algebra).

(a) For a vector $\mathbf{w} \in \mathbb{R}^n$ and another vector $\mathbf{a} \in \mathbb{R}^n$, consider the function $f(\mathbf{w}) := f(w_1, \ldots, w_n) = \langle \mathbf{a}, \mathbf{w} \rangle := \sum_{i=1}^n a_i w_i$. What is the partial derivative of f with respect to w_i? The *gradient* of a function is the collection of its partial derivatives when viewed as a vector. What is the gradient ∇f? Recall that the gradient is (usually) a function of the $\mathbf{w}$ variable, but what do you observe in this case?

(b) Compute the gradient of the function $f(w_1, w_2, w_3) = w_1 w_2 + w_2 w_3 + w_1 w_3$.

(c) Suppose we have the linear equation $\mathbf{y} = \mathbf{A}\mathbf{v}$, where $\mathbf{y}$ and $\mathbf{v}$ are vectors and $\mathbf{A}$ is a matrix. Argue that $\mathbf{y}$ can be written as a linear combination of the columns of $\mathbf{A}$.

(d) Given any vector $\mathbf{v}$ with real entries, show that $\mathbf{v}^\top \mathbf{v}$ is always non-negative.

(e) A square and symmetric matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$ is one that satisfies $\mathbf{X}^\top = \mathbf{X}$. In addition, $\mathbf{X}$ is said to be positive semidefinite (PSD) (written $\mathbf{X} \succeq 0$) if, for all vectors $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^T \mathbf{X} \mathbf{v} = \sum_{i=1}^d \sum_{j=1}^d \mathbf{X}_{i,j} v_i v_j \geq 0$.

Suppose a matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$ for another matrix $\mathbf{B}$ (not necessarily square). Argue that $\mathbf{A}$ is square, symmetric, and PSD.

Hint: You may want to try this problem for when $\mathbf{B}$ is just a $1 \times d$ vector. For the general case, think about whether you can express $\mathbf{v}^\top \mathbf{A} \mathbf{v}$ in terms of $\mathbf{B}\mathbf{v}$.

(f) Recall that a matrix $\mathbf{A}$ has (scalar) eigenvalue λ associated with an eigenvector $\mathbf{v}$ if $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. To avoid identifiability issues, we also assume that the ℓ_2 norm of $\mathbf{v}$ is equal to 1, i.e., $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2} = 1$. Compute eigenvectors and eigenvalues for the following matrix by hand:

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

(g) Let $\mathbf{A}$ be an invertible matrix. Show that if $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, then it is also an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue λ^{-1}.

Problem 2 (Probability).

(a) The conditional probability of an event $\mathcal{E}$ given another event $\mathcal{F}$ is given by $\Pr(\mathcal{E}|\mathcal{F}) = \Pr(\mathcal{E} \cap \mathcal{F}) / \Pr(\mathcal{F})$. Use this fact to derive Bayes' rule:

$$\Pr(\mathcal{F}|\mathcal{E}) = \Pr(\mathcal{E}|\mathcal{F}) \cdot \Pr(\mathcal{F}) / \Pr(\mathcal{E}),$$

which makes sense provided we are not dividing by zero.

Hint: Use the fact that intersections commute, i.e., the event $\mathcal{E} \cap \mathcal{F}$ is identical to the event $\mathcal{F} \cap \mathcal{E}$.

(b) Suppose we have two random variables X and Y taking real values (i.e., in the set $\mathbb{R}$) and having finite expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, respectively. Can you write $\mathbb{E}[X+Y]$ in terms of expectations of the individual random variables? Can you write $\mathbb{E}[XY]$ in terms of expectations of the individual random variables?

(c) The variance of a random variable is given by $\mathbb{E}[(X - \mathbb{E}[X])^2]$. Argue that this is always non-negative. Suppose we have two independent random variables X and Y with variances σ_1^2 and σ_2^2, respectively. What is the variance of $X + Y$? What is the variance of $X - Y$?

(d) Suppose X is a random variable with expectation 0 and variance 1. What are the expectation and variance of $Y = aX + b$ (here, a and b are some fixed real numbers, and your answer will be in terms of (a, b)).

(e) Now consider an n-dimensional vector of random variables (i.e. each entry of the vector is itself a random variable) $Z \in \mathbb{R}^n$. Suppose Z has zero expectation, i.e., $\mathbb{E}[Z_i] = \mu_i = 0$ for all $1 \leq i \leq n$ and that its *covariance matrix* has entries $\mathbf{\Sigma}_{i,j} = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)]$. What is the expectation and covariance matrix of $W = \mathbf{A}Z + \mathbf{b}$, where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{b}$ is an n-dimensional vector?

Hint: Use the fact that the expectation is *linear*: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Problem 3 (Python installation and basic simulation). It is strongly recommended that you use Jupyter notebooks for your coding assignments. This exercise guides you through an installation through Anaconda (please do this only if you do not have a working version of Jupyter already), and asks some basic questions about generating data and plotting. The latter questions will only make use of the numpy package.

(a) Please install Python using Anaconda if you haven't already, documentation can be found here: `https://docs.anaconda.com/anaconda/`. Review the following tutorial on using Python through Jupyter notebooks: `https://cs4540-f18.github.io/notes/python-basics` If you really need to refresh, here is a more elementary tutorial on Python programming: `https://www.learnpython.org/`

(b) Sample 100 points from a 2-dimensional multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, where $\mathbf{I}$ is the 2×2 identity matrix. Store these points in a 100×2 dimensional matrix.

(c) Produce a 2-dimensional scatter plot of all the points you just sampled.

(d) Suppose for the moment that you didn't know the mean or covariance matrix of the Gaussian distribution from which the points were sampled. What is a reasonable *estimate* of the 2-dimensional mean vector from the data that you sampled? Compute this estimate and report it. Analogously, what is a reasonable *estimate* of the 2×2 covariance matrix from the data that you sampled? Compute this estimate and report it.

(e) Compute the ℓ_2 distance between your mean estimate and the true mean $(0, 0)$ and report it.

(f) Intuitively, what do expect will happen to this error if you drew 1000 samples instead?