Machine Learning Exercise Sheet 1

Math Refresher

The machine learning lecture relies heavily on your knowledge of undergraduate mathematics, especially linear algebra and probability theory. You should think of this exercise sheet as a test to see if you meet the prerequisites for taking this course. If you struggle with a large fraction of the exercises you should reconsider taking this lecture at this point and instead first prepare by taking a course that reinforces your mathematical foundations (e.g. "Basic Mathematical Tools for Imaging and Visualization" (IN2124)).

Homework

Reading

We strongly recommend that you review the following documents to refresh your knowledge. You should already be familiar with most of their content from your previous studies.

- Linear algebra `http://cs229.stanford.edu/section/cs229-linalg.pdf` (except sections 4.4, 4.5, 4.6), and `http://ee263.stanford.edu/notes/matrix_crimes.pdf` (common linear algebra mistakes)
- Probability theory `http://cs229.stanford.edu/summer2020/cs229-prob.pdf`

Additionally, we recommend this refresher of matrix derivatives and we will be referring to the Matrix Cookbook several times in this lecture.

Linear Algebra

Notation. We use the following notation in this lecture:

- Scalars are denoted with lowercase letters, e.g. a, x, μ.
- Vectors are denoted with bold lowercase letters, e.g. $\boldsymbol{a}$, $\boldsymbol{x}$, $\boldsymbol{\mu}$.
- Matrices are denoted with bold uppercase letters, e.g. $\boldsymbol{A}$, $\boldsymbol{X}$, $\boldsymbol{\Sigma}$.
- $\mathbb{R}^N$ denotes N-dimensional Euclidean space, i.e. the set of N-dimensional vectors with real-valued entries. For example, $\boldsymbol{x} = (2, \sqrt{2}, 6.5, -7)^T$ is an element of $\mathbb{R}^4$, which we denote as $\boldsymbol{x} \in \mathbb{R}^4$.
- $\mathbb{R}^{M\times N}$ is the set of matrices with M rows and N columns. For example, the matrix $\boldsymbol{A} = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 4 & 5 \end{pmatrix}$ is an element of $\mathbb{R}^{2\times 3}$, which we denote as $\boldsymbol{A} \in \mathbb{R}^{2\times 3}$.
- A function $f : \mathcal{X} \to \mathcal{Y}$ maps elements of the set $\mathcal{X}$ into the set $\mathcal{Y}$. An example would be a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined as $f(x, y) = 2x^2 + xy - 4$.

Problem 1: Let $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{Z} \in \mathbb{R}^{P\times Q}$. The function $f : \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^{P\times Q} \to \mathbb{R}$ is defined as

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{Z}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{y} + \boldsymbol{B}\boldsymbol{x} - \boldsymbol{y}^T \boldsymbol{C}\boldsymbol{Z}\boldsymbol{D} - \boldsymbol{y}^T \boldsymbol{E}^T \boldsymbol{y} + \boldsymbol{F}.$$

What should be the dimensions (shapes) of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}$ for the expression above to be a valid mathematical expression?

Problem 2: Let $\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{M} \in \mathbb{R}^{N\times N}$. Express the function $f(\boldsymbol{x}) = \sum_{i=1}^N \sum_{j=1}^N x_i x_j M_{ij}$ using **only** matrix-vector multiplications.

Problem 3: Let $\boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{b} \in \mathbb{R}^M$. We are interested in solving the following system of linear equations for $\boldsymbol{x}$

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

a) Under what conditions does the system of linear equations have a **unique** solution $\boldsymbol{x}$ for **any** choice of $\boldsymbol{b}$?

b) Assume that $M = N = 5$ and that $\boldsymbol{A}$ has the following eigenvalues: $\{-5, 0, 1, 1, 3\}$. Does Equation 1 have a unique solution $\boldsymbol{x}$ for any choice of $\boldsymbol{b}$? Justify your answer.

Problem 4: Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$. Assume that there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ such that $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$. What can you say about the eigenvalues of $\boldsymbol{A}$? Justify your answer.

Problem 5: A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is positive semi-definite (PSD) if and only if for any $\boldsymbol{x} \in \mathbb{R}^N$ it holds that $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} \geq 0$. Prove that a symmetric matrix $\boldsymbol{A}$ is PSD **if and only if** it has no negative eigenvalues.

Problem 6: Let $\boldsymbol{A} \in \mathbb{R}^{M\times N}$. Prove that the matrix $\boldsymbol{B} = \boldsymbol{A}^T \boldsymbol{A}$ is positive semi-definite for any choice of $\boldsymbol{A}$.

Calculus

Problem 7: Consider the following function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{2} a x^2 + bx + c$$

We are interested in solving the following optimization problem

$$\min_{x\in\mathbb{R}} f(x)$$

a) Under what conditions does this optimization problem have (i) a unique solution, (ii) infinitely many solutions or (iii) no solution? Justify your answer.

b) Assume that the optimization problem has a unique solution. Write down the closed-form expression for $x^\star$ that minimizes the objective function, i.e. find $x^\star = \arg\min_{x\in\mathbb{R}} f(x)$.

Problem 8: Consider the following function $g : \mathbb{R}^N \to \mathbb{R}$

$$g(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a symmetric, PSD matrix, $\boldsymbol{b} \in \mathbb{R}^N$ and $c \in \mathbb{R}$.

We are interested in solving the following optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \ g(\boldsymbol{x})$$

a) Compute the Hessian $\nabla^2 g(\boldsymbol{x})$ of the objective function. Under what conditions does this optimization problem have a unique solution?

b) Why is it necessary for a matrix $\boldsymbol{A}$ to be PSD for the optimization problem to be well-defined? *Hint: What happens if $\boldsymbol{A}$ has a negative eigenvalue?*

c) Assume that the matrix $\boldsymbol{A}$ is positive definite (PD). Write down the closed-form expression for $\boldsymbol{x}^\star$ that minimizes the objective function, i.e. find $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} g(\boldsymbol{x})$.

Probability Theory

Notation. We use the following notation in our lecture

- For conciseness and to avoid clutter, we use $p(x)$ to denote multiple things
  1. If X is a discrete random variable, $p(x)$ denotes the probability mass function (PMF) of X at point x (usually denoted as $p_X(x)$ or $p(X = x)$ in the statistics literature).
  2. If X is a continuous random variable, $p(x)$ denotes the probability density function (PDF) of X at point x (usually denoted as $f_X(x)$ in the statistics literature).
  3. If $A \in \Omega$ is an event, $p(A)$ denotes the probability of this event (usually denoted as $\Pr(\{A\})$ or $\mathbb{P}(\{A\})$ in the statistics literature)

  You will mostly encounter (1) and (2) throughout the lecture. Usually, the meaning is clear from the context.
- Given the distribution $p(x)$, we may be interested in computing the expected value $\mathbb{E}_{p(x)}[f(x)]$ or, equivalently, $\mathbb{E}_X[f(x)]$. Usually, it is clear with respect to which distribution we are computing the expectation, so we omit the subscript and simply write $\mathbb{E}[f(x)]$.
- $x \sim p$ means that x is distributed (sampled) according to the distribution p. For example, $x \sim \mathcal{N}(\mu, \sigma^2)$ (or equivalently $p(x) = \mathcal{N}(x|\mu, \sigma^2)$) means that x is distributed according to the normal distribution with mean μ and variance σ^2.

Problem 9: Prove or disprove the following statement

$$p(a|b, c) = p(a|c) \Rightarrow p(a|b) = p(a)$$

Problem 10: Prove or disprove the following statement

$$p(a|b) = p(a) \Rightarrow p(a|b, c) = p(a|c)$$

Problem 11: You are given the joint PDF $p(a, b, c)$ of three continuous random variables. Show how the following expressions can be obtained using the rules of probability

1. $p(a)$
2. $p(c|a, b)$
3. $p(b|c)$

Problem 12: Researchers have developed a test which determines whether a person has a rare disease. The test is fairly reliable: if a person is sick, the test will be positive with 95% probability, if a person is healthy, the test will be negative with 95% probability. It is known that $\frac{1}{1000}$ of the population have this rare disease. A person (chosen uniformly at random from the population) takes the test and obtains a positive result. What is the probability that the person has the disease?

Problem 13: Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and $f(x) = ax + bx^2 + c$. What is $\mathbb{E}[f(x)]$?

Problem 14: Let $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ (where $\boldsymbol{A} \in \mathbb{R}^{N \times N}$). What are the values of the following expressions:

- $\mathbb{E}[g(\boldsymbol{x})]$,
- $\mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T]$,
- $\mathbb{E}[g(\boldsymbol{x})^T g(\boldsymbol{x})]$,
- the covariance matrix $\mathrm{Cov}[g(\boldsymbol{x})]$.