

## 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](http://ee263.stanford.edu/notes/matrix_crimes.pdf) (common linear algebra mistakes)
- Probability theory <http://cs229.stanford.edu/summer2020/cs229-prob.pdf>

#### Linear Algebra

**Notation.** We use the following notation in this lecture:

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

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

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

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

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

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

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (1)$$

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

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

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

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

## Calculus

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

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

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  $x^*$  that minimizes the objective function, i.e. find  $x^* = \arg \min_{x \in \mathbb{R}} f(x)$ .

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

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

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

We are interested in solving the following optimization problem

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


---

- a) Compute the Hessian  $\nabla^2 g(\mathbf{x})$  of the objective function. Under what conditions does this optimization problem have a unique solution?
- b) Why is it necessary for a matrix  $\mathbf{A}$  to be PSD for the optimization problem to be well-defined?  
*Hint: What happens if  $\mathbf{A}$  has a negative eigenvalue?*
- c) Assume that the matrix  $\mathbf{A}$  is positive definite (PD). Write down the closed-form expression for  $\mathbf{x}^*$  that minimizes the objective function, i.e. find  $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathbb{R}^N} g(\mathbf{x})$ .

## 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  $\mu$  and variance  $\sigma^2$ .

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

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