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`

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 $f(x, y) = 2x^2 + xy - 4$.

$M \times 1$ $N \times 1$ $P \times Q$

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?

$A \in R^{M\times N}$ $B \in R^{1\times M}$ $C \in R^{N\times P}$ $D \in R^{Q\times 1}$ $E \in R^{N\times N}$ $F \in R$

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. $f(x) = x \cdot x^T \cdot M$ $\quad x^T M \cdot x$

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 $\boldsymbol{b}$? $r = m = n$

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. $|A| = \prod_{i=1}^{n} \lambda_i = 0$ ⇒有自由 ⇒满秩冲突 A是奇异矩阵不可逆.

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. $A^{-1}A = AA^{-1} = I$ $\quad B = A^{-1}$

A是可逆矩阵 ⇒ 满秩 ⇒ n个特征向量,特征值全不为0.

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. $A = U \cdot \Lambda U^T$ $\quad x^T \cdot A \cdot x = \underbrace{x^T \cdot U}_{y^T} \cdot \Lambda U^T \cdot x = y^T \cdot \Lambda \cdot y = \sum_{i=1}^{n} \lambda_i \cdot y_i^2$ $\quad \because y_i^2 \geq 0$ $\therefore \lambda_i \geq 0$ 时 A is PSD.

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 $\boldsymbol{A}$. $x^T B x \geq 0$ $\quad = x^T A^T \cdot A \cdot x = (Ax)^T(Ax) = \|Ax\|_2^2 \geq 0.$

If PSD, then $\lambda \geq 0$: $x^T A \cdot x \geq 0 \Leftarrow Av = \lambda v$ ← eigenvector; $v^T \lambda \cdot v \geq 0$, $\lambda\, v^T \cdot v \geq 0$, $\lambda \cdot \|v\|_2^2 \geq 0$

If $\lambda \geq 0$, then PSD ≥ 0: $x^T A x = \sum_{i=1}^{n} w_i v_i A \sum_{j=1}^{n} w_j v_j$, $x = \sum_{i=1}^{n} w_i v_i$; $= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \cdot v^T A v$ ⇐ $Av = \lambda v$; $= \sum_{i=1}^{n}\sum_{j=1}^{n} w^2 \cdot v^T \lambda v$ ⇒ $\lambda \|v\|_2^2 \geq 0$.

Calculus

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

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

We are interested in solving the following optimization problem

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

i $a > 0$ $\quad$ ii ~~$a < 0$~~ $a = 0$ $\quad$ iii $a = 0$, 且 $b \neq 0$ or $a < 0$

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 $x^\star = \arg\min_{x\in\mathbb{R}} f(x)$. $f'(x) = ax + b = 0$ $\quad \boxed{x = -\frac{b}{a}}$

Problem 8: Consider the following function $g : \mathbb{R}^N \to \mathbb{R}$ $\quad x \in \mathbb{R}^{N\times 1}$

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

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

$1\times n \quad n\times n: \quad (x_1 \cdots x_n) \begin{pmatrix} A_{11} & \cdots & \\ \vdots & & \\ & & A_{nn} \end{pmatrix}$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a symmetric, PSD matrix, $\boldsymbol{b} \in \mathbb{R}^{N\times 1}$ and $c \in \mathbb{R}$. $= (A_{11} \cdot x_1 + \cdots + A_{n1} \cdot x_n \;\cdots\; A_{1n} \cdot x_1 + \cdots + A_{nn} \cdot x_n) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $\quad 1\times n, \; n\times 1$

We are interested in solving the following optimization problem $\quad = A_{11} \cdot x_1 \cdot x_1 + A_{12} \cdot x_1 \cdot x_2 + \cdots$

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

$g(x) = \frac{1}{2}\sum_{i}^{n}\sum_{j}^{n} x_i A_{ij} x_j + \sum_{i}^{n} b_i x_i + c$ $\quad = \frac{1}{2}\sum_{i}^{n}\sum_{j}^{n} x_i \cdot A_{ij} \cdot x_j$

$\frac{\partial g(x)}{\partial x_1 \partial x_2} = A_{kl}$

$1\times n \quad n\times 1$

$\Rightarrow \nabla^2 g(x) = \begin{bmatrix} \frac{\partial g(x)}{\partial x_1 \partial x_1} & \cdots \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \cdots \\ \vdots & & \vdots \end{bmatrix} = A$ $\quad (b_1 \cdots b_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \Rightarrow b_1 x_1 + b_2 x_2 \cdots b_n \cdot x_n = \sum_{i}^{n} b_i \cdot x_i$

$x = a\vec{v}$ v → negative λ.

$x^T A x + b x + c$

$= a v^T \cdot A \cdot a v + b a v + c$

$= a^2 v^T \lambda v + b a v + c$ A is pd.

A has negative eigen value

$= a^2 \lambda \|v\|_2^2$

$a \to \infty$. $\lambda < 0$ $\|v\|_2^2 > 0$. 当A为正定矩阵时，有极小值

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?* 不存在最小值，why?

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 $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} g(\boldsymbol{x})$.

$g(x) = \frac{1}{2}\sum_i^n \sum_j^n x_i A_{ij} x_j + \sum_i^n b_i x_i + c$ $\quad \frac{\partial g(x)}{\partial x_k} = \frac{1}{2}\sum_i^n A_{ij} x_i + \frac{1}{2}\sum_i^n A_{ij} x_j + b_k$

$= \sum_i^n A_{ij} x_i + b_k$

$\nabla g(x) = Ax + b = 0$ ∵A is PD $\lambda > 0$ ∴A 可逆 $x = -A^{-1}b$

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.

c = F or c = U

$p(a=T) = p(a=T, c=F) + p(a=T, c=U)$
$= p(T|F)\cdot p(F) + p(T|U)\cdot p(U) = \frac{3}{4}$

Problem 9: Prove or disprove the following statement

a, b are dice roll c = a+b.

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

$p(a=T|b=T) = \frac{p(a=T, b=T)}{p(b=T)}$

a, b is independent ⇒ $p(a|b) = p(a)$.

Problem 10: Prove or disprove the following statement

one it we know the sum.

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

$= \frac{p(a=T, b=T | c=F)\cdot p(c=F) + p(a=T, b=T | c=U)\cdot p(c=U)}{p(b=T)}$ = different ...

概率密度函数

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

p(c)

1. $p(a)$ $\quad p(a) = \iint p(a,b,c)\, db\, dc$ $\quad \frac{p(b,c)}{p(c)} = \frac{\int p(a,b,c)\cdot da}{\iint p(a,b,c)\cdot db\, da}$
2. $p(c|a, b)$ $\quad \frac{p(a,b,c)}{p(a,b)} = \frac{p(a,b,c)}{\int p(a,b,c)\cdot dc}$
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?

a = 健康 b = 不健康 $p(c|b) = \frac{95}{100}$ c = 阳 d = 阴. $p(b|c)$ $p(c)$

$p(a) = \frac{999}{1000}$ $p(b) = \frac{1}{1000}$ $p(d|a) = \frac{95}{100}$ $p(c,b) = p(c|b)\cdot p(b) = p(b,c) = p(b|c)\cdot p(c)$.

$p(c|a) = \frac{5}{100}$ $= \frac{95}{100}\cdot\frac{1}{1000}\cdot\frac{1}{p(c)} = \frac{95}{5090} = \boxed{0.0187}$

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 $\mathrm{Cov}[g(\boldsymbol{x})]$.

$$E[ax+bx^2+c] = E[ax] + E[bx^2] + E[c]$$
$$= aE[x] + bE[x^2] + c$$
$$= a\mu + b(D[x] - E[x]^2) + c$$
$$= a\mu + b(\sigma^2 - \mu^2) + c$$

书 P82

$$E[gx] = E[Ax] = AE[x] = A\mu$$
$$E[g(x)\cdot g(x)^T] = E[Ax\cdot x^T\cdot A^T]$$
$$= A\cdot(\mu\cdot\mu^T + \Sigma)\cdot A^T$$

$E[tr(x)] = tr(E(x))$

$tr(AB) = tr(BA)$

$$E(g(x)^T\cdot g(x)] = E[x^TA^TAx]$$

cycle permutation

$E[tr(ABC)] = E[tr(CAB)]$

$$= E[tr(Axx^TA^T)]$$
$$= tr\,E(Axx^TA^T)$$
$$= tr[A\cdot(\mu\cdot\mu^T+\Sigma)\cdot A^T]$$

$$p(c) = p(c,a) + p(c,b)$$
$$= \frac{5}{100}\cdot\frac{999}{1000} + \frac{95}{100}\cdot\frac{1}{1000}$$
$$= \frac{5090}{100\cdot 1000}$$

$$cov(g(x)) = E[g(x)^2] - E[g(x)]^2$$
$$= E(g(x)\cdot g(x)^T] - E[g(x)]\cdot E[g(x)]^T$$
$$= A\cdot(\mu\cdot\mu^T+\Sigma)\cdot A^T - A\cdot\mu\cdot\mu^T\cdot A$$
$$= A\Sigma\cdot A^T$$