CQF Module 4.3: Mathematics Toolbox for Machine Learning Exercises

Panos Parpas

Exercise 1(Dot products) Compute $x^\top y$ where $x = (1, -2, 5, -1)^\top$ and $y = (0, 4, -3, 7)^\top$.

Exercise 2(Matrix product) Compute $y = Ax$ as well as the ℓ_2 norm of x and y, where

$$A = \begin{pmatrix} -1 & 4 & 7 & 2 \\ 3 & -2 & -1 & 0 \\ 5 & 3 & 0 & -1 \end{pmatrix}, \quad x = (-3, 2, 1, 3)^\top.$$

Exercise 3(Basis) A collection of n linearly independent vectors of dimension n is called a basis. An important property of a basis is that if the n dimensional vectors $a_1, \dots, a_n$ with $a_i \in \mathbb{R}^n$ are a basis then any vector $b \in \mathbb{R}^n$ can be written as,

$$b = \sum_{i=1}^{n} \beta_i a_i$$

where $\beta_i \in \mathbb{R}$.
Which of the following set of vectors form a basis for $\mathbb{R}^2$?

1. $\{(1, 1), (1, 0)\}$
2. $\{(2, 4), (3, -1)\}$
3. $\{(1, -1), (0, 2), (2, 1)\}$
4. $\{(2, -1), (-2, 1)\}$
5. $\{(0, 3)\}$

Exercise 4(Span of vectors) The set of all linear combinations of vectors $v_1, \ldots, v_n$ is called the span of these vectors. Sometimes the span is written as $\mathrm{Span}[v_1, \ldots, v_n]$. Which of the following points are within the span of $\{(-1, 0, 2), (3, 1, 0)\}$?

1. $(0, 1, 1)$
2. $(1, 1, 4)$
3. $(2, 1, 1)$
4. $(-3, 4, 2)$
5. $(0, 0, 0)$

Exercise 5(Eigen decomposition). Consider a matrix $A \in \mathbb{R}^{d\times s}$ and assume it has an eigen decomposition of $A = Q\Lambda Q^{-1}$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. When A is symmetric we also have $Q^{-1} = Q^\top$. Answer the following questions.

a. If A is symmetric, show that $x^\top Ax \geq 0$ for any $x \in \mathbb{R}^{d\times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b. Show that $\mathrm{Tr}(A) = \sum_{i=1}^d \lambda_i$ where Tr is the trace of A, i.e.$\mathrm{Tr}(A) = \sum\limits_{i=1}^{d} A_{ii}$.

c. Show that $\det(A) = \prod_{i=1}^d \lambda_i$ where $\det(A)$ is the determinant of A.

Hint: The following property of the trace (known as the permutation invariance property) may be useful: $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. You may also need to use the fact that if A and B are $d \times d$ matrices, then $\det(AB) = (\det A)(\det B)$ and that $\det(A^{-1}) = \det(A)^{-1}$.

Exercise 6(Vector notation) We define the probability density on the vector $\mathbf{x} \in \mathbb{R}^3$ with all elements $0 \leq x_k \leq 1$ as

$$p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1x_2 + x_2^2 + 2x_2x_3)\,. \tag{1}$$

Where C is the normalizing constant. Put this into notation that only uses $\mathbf{x} \in \mathbb{R}^3$ as a single whole vector.

Exercise 7(Index notation). Turn the following matrix-vector expressions into index notation:

1. $\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\mathbf{x}$
2. $\mathrm{Tr}(\boldsymbol{A})$
3. $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B})$
4. $\mathbf{y}^\intercal \boldsymbol{A}^\intercal \mathbf{x}$

Exercise 8 (Jacobian). First find the dimensions, then the Jacobian. It's probably easiest here to use index notation.

1. $f(\mathbf{x}) = \sin(x_1)\cos(x_2)$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.
2. $f(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{y}$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.
3. $f(\mathbf{x}) = \mathbf{x}\mathbf{x}^\intercal$, find $\mathrm{d}f/\mathrm{d}\mathbf{x}$.
4. $f(\mathbf{t}) = \sin(\log(\mathbf{t}^\intercal \mathbf{t}))$, find $\mathrm{d}f/\mathrm{d}\mathbf{t}$.
5. $f(\boldsymbol{X}) = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})$ for $\boldsymbol{A} \in \mathbb{R}^{D\times E}$, $\boldsymbol{X} \in \mathbb{R}^{E\times F}$, $\boldsymbol{B} \in \mathbb{R}^{F\times D}$, find $\mathrm{d}f/\mathrm{d}\boldsymbol{X}$.

Exercise 9 (Chain rule) Compute the derivatives $\mathrm{d}f/\mathrm{d}\mathbf{x}$ of the following functions.

- First, write out the chain rule for the given decomposition.
- Give the shapes of intermediate results, and make clear which dimension(s) will be summed over.
- Provide expressions for the derivatives, and describe your steps in detail. Providing an expression means specifying everything up to the point where you could implement it.
- Give the results in vector notation if you can.

1. $f(z) = \log(1+z), \qquad z = \mathbf{x}^\intercal \mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^D.$
2. $f(z) = \exp(-\frac{1}{2}z), \qquad z = \mathbf{y}^\intercal \boldsymbol{S}^{-1}\mathbf{y}, \qquad \mathbf{y} = \mathbf{x} - \boldsymbol{\mu}.$
3. $f(\boldsymbol{A}) = \mathrm{Tr}(\boldsymbol{A}), \qquad \boldsymbol{A} = \mathbf{x}\mathbf{x}^\intercal + \sigma^2 \boldsymbol{I}.$
4. $\mathbf{f}(\mathbf{z}) = \tanh(\mathbf{z}), \qquad \mathbf{z} = \boldsymbol{A}\mathbf{x} + \mathbf{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{M\times N}.$

Remember: Generally, scalar functions are applied elementwise to vectors/matrices.

Exercise 10(Hessian of Linear Regression). For the stationary point of linear regression, find the Hessian, and prove that it is positive definite. If you need to make any additional assumptions state them explicitly.

Exercise 11[SVD and PCA] Assume we are given a dataset $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ such that $\text{mean}(\boldsymbol{x}_n) = \mathbf{0}$. Write $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]^\top \in \mathbb{R}^{N\times D}$, demonstrate how to use singular value decomposition (SVD) of $\mathbf{X}$ to obtain PCA solutions of M principal components.