Self-Assessment Assignment for the Machine Learning course

Yevgeny Seldin, Christian Igel
Department of Computer Science, University of Copenhagen

Math and ML

You need basic linear algebra, calculus, and probability theory for understanding machine learning. To recall some of your math knowledge, answer the following questions. Do the calculations by hand.

If you feel unsure about what a question means and how to answer it, this indicates that you are not fully comfortable with mathematical skills that are assumed in this course. No worries. In this case, just take your time and to go back to your notes from school or your first study years – or grab one of the numerous textbooks that are around.

Feel free to ask questions about the mathematical background in the exercise classes!

1 Vectors and Matrices

Consider the two vectors

$$\boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \text{ and } \boldsymbol{b} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

and the matrix

$$\boldsymbol{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad .$$

1. Calculate the *inner product* (also known as *scalar product* or *dot product*) denoted by $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$, or $\boldsymbol{a} \cdot \boldsymbol{b}$.

2. Calculate the length (also known as Euclidean norm) $\|\boldsymbol{a}\|$ of the vector $\boldsymbol{a}$.

3. Calculate the *outer product* $\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$. Is it equal to the inner product $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$ you computed in Question 1?

4. Calculate the four quantities, $\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}$, $\boldsymbol{b}\boldsymbol{a}^{\mathrm{T}}$, $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{a}$. Which of them are equal and which are not? (Test yourself: there is one pair that is equal and all the remaining quantities are different.)

5. Calculate the inverse of matrix $\boldsymbol{M}$, denoted by $\boldsymbol{M}^{-1}$. We remind that you should get that $\boldsymbol{M}\boldsymbol{M}^{-1} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

6. Calculate the matrix-vector product $\boldsymbol{M}\boldsymbol{a}$.

7. Let $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$. Calculate the transpose of $\boldsymbol{A}$, denoted by $\boldsymbol{A}^{\mathrm{T}}$. Is $\boldsymbol{A}$ symmetric? (A matrix is called symmetric if $\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}}$.)

8. What is the rank of $\boldsymbol{A}$? (The rank is the number of linearly independent columns.) Give a short explanation.

9. What should be the relation between the number of columns and the rank of a square matrix in order for it to be invertible? Is $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$ invertible?

10. Calculate the projection of vector $\boldsymbol{a}$ onto vector $\boldsymbol{b}$. (The result should be a vector pointing in the same direction as $\boldsymbol{b}$!)

2 Derivatives

We denote the derivative of a univariate function $f(x)$ with respect to the variable x by $\frac{df(x)}{dx}$. We denote the partial derivative of a multivariate function $f(x_1, \ldots, x_n)$ with respect to the variable x_i, where $1 \leq i \leq n$, by $\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}$. The partial derivative $\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}$ is the derivative of $f(x_1, \ldots, x_n)$ with respect to x_i when we treat all other variables x_j for $j \neq i$ in f as constants.

Please recall the basic rules for derivatives, namely the sum rule, the chain rule, and the product rule.

1. What is the derivative of $f(x) = \frac{1}{1+\exp(-x)}$ with respect to x?

2. What is the partial derivative of $f(w, x) = 2(wx + 5)^2$ with respect to w?

3 Probability Theory: Sample Space

An urn contains five red, three orange, and one blue ball. Two balls are randomly selected (without replacement).

1. What is the sample space of this experiment?
2. What is the probability of each point in the sample space?
3. Let X represent the number of orange balls selected. What are the possible values of X?
4. Calculate the probability that X is equal to zero, $\mathbb{P}\{X = 0\}$.
5. Calculate the expectation of X, denoted by $\mathbb{E}[X]$.
6. Calculate the variance of X, denoted by $\mathbb{V}[X]$.

4 Probability Theory: Properties of Expectation

Let X and Y be two discrete random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Starting from the definitions, prove the following identities:

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
2. If X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. (Mark the step where you are using the independence assumption. Note that this assumption was not required in point 1.)
3. Provide an example of two random variables X and Y for which $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$. (Describe how you define the random variables, provide a joint probability distribution table [see comment below], and calculate $\mathbb{E}[XY]$ and $\mathbb{E}[X]\mathbb{E}[Y]$.)
4. $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$.
5. Variance of a random variable is defined as $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. Show that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Comment: A convenient way to represent a joint probability distribution of two discrete random variables is a table. For example, if X and Y are Bernoulli random variables with bias $\frac{1}{2}$ (fair coins), then the joint distribution table looks like this:

$X\backslash Y$	0	1
0	1/4	1/4
1	1/4	1/4

.

And if Z_1 and Z_2 are Bernoulli random variables with bias $\frac{1}{2}$ and we define $X = Z_1 + Z_2$ and $Y = Z_1 \times Z_2$, then the joint distribution of X and Y is:

$X \backslash Y$	0	1
0	1/4	0
1	1/2	0
2	0	1/4

.

5 Probability Theory: Complements of Events

1. The complement of event A is denoted by $\bar{A}$ and defined by $\bar{A} = \Omega \setminus A$. Starting from probability axioms prove that $\mathbb{P}\{A\} = 1 - \mathbb{P}\{\bar{A}\}$.
2. In many cases it is easier to calculate the probability of a complement of an event than to calculate the probability of the event itself. Use this to solve the following question. We flip a fair coin 10 times.
   - What is the probability to observe at least one tail?
   - What is the probability to observe at least two tails?

6 Induction

Prove by induction that for all integer n and d, such that $n \geq d \geq 1$:

$$\sum_{i=0}^{d} \binom{n}{i} \leq n^d + 1.$$

7 Programming

Implement the following task in Python.

1. Download the Iris dataset from UCI Machine Learning repository, `https://archive.ics.uci.edu/ml/datasets/iris`. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
2. Visualize the data set in the 2D space of the first two features (sepal length and sepal width). Points from different classes should be visualized with different colors.
3. Calculate the average sepal length for each of the three classes.

4. Calculate the variance of sepal length for each of the three classes. (It is advised to implement the variance calculation yourself rather than use built-in functions: this will give you some hands-on experience with the concept.)

5. Python is an interpreter language (as opposed to a compiled language) and, therefore, in terms of run-time efficiency it is highly recommended to use precompiled functions instead of for-loops whenever possible. Use matrix-vector multiplications and other built-in functions to compute the Euclidean distance between the first data point and all the remaining data points *without using any for-loops*! Report the average distance and its variance.

   The Euclidean distance between two points $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$ is defined by $\mathrm{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$. If we denote the data points by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, then you should compute $\mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2), \ldots, \mathrm{dist}(\mathbf{x}_1, \mathbf{x}_n)$ and then the average and the variance. Practice writing efficient code — do not use any for-loops!