Machine Learning
Michaelmas Term 2021
Week 1

Problem Sheet 0 Solutions

1 Linear Algebra Background

Consider the following matrix $\mathbf{X}$ and vectors $\mathbf{y}$ and $\mathbf{z}$

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Answer the following questions:

1. The dot (or inner) product of $\mathbf{y}$ and $\mathbf{z}$ is denoted by $\mathbf{y} \cdot \mathbf{z}$ (or sometimes $\langle \mathbf{y}, \mathbf{z} \rangle$, $\mathbf{y}^\mathsf{T}\mathbf{z}$ or $\mathbf{z}^\mathsf{T}\mathbf{y}$). What is $\mathbf{y} \cdot \mathbf{z}$?

   Solution: $\mathbf{y} \cdot \mathbf{z} = 3 \cdot 2 + 1 \cdot 3 = 9$.

2. What are the products $\mathbf{X}\mathbf{y}$ and $\mathbf{z}^\mathsf{T}\mathbf{X}$?

   Solution:

$$\mathbf{X}\mathbf{y} = \begin{pmatrix} 1 \cdot 3 + 2 \cdot 1 \\ 3 \cdot 3 + 4 \cdot 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 13 \end{pmatrix}$$
$$\mathbf{z}^\mathsf{T}\mathbf{X} = \begin{pmatrix} 2 \cdot 1 + 3 \cdot 3 & 2 \cdot 2 + 3 \cdot 4 \end{pmatrix} = \begin{pmatrix} 11 & 16 \end{pmatrix}$$

   Note that $\mathbf{X}\mathbf{y}$ is 2×1 and $\mathbf{z}^\mathsf{T}\mathbf{X}$ is 1×2.

3. Does the inverse of matrix $\mathbf{X}$ exist? If so what is it?

   Solution: Yes, the inverse of $\mathbf{X}$ exists. It can be checked that

$$\mathbf{X}^{-1} = \begin{pmatrix} -2 & 1 \\ {}^{3}/_{2} & {}^{-1}/_{2} \end{pmatrix}.$$

4. What is the determinant of matrix $\mathbf{X}$?

   Solution: $\det(\mathbf{X}) = 1 \cdot 4 - 2 \cdot 3 = -2$.

5. What is the rank of matrix $\mathbf{X}$?

   Solution: Since $\det(\mathbf{X})$ is not zero, the matrix is *full rank*. In this case the rank is 2, since the matrix is 2×2.

2 Calculus Background

Consider the function $f(x) = x^3 - 3x + 7$ and answer the following questions.

1. What is the derivative $\frac{\mathrm{d}f}{\mathrm{d}x}$?

   Solution: $\frac{\mathrm{d}f}{\mathrm{d}x} = 3x^2 - 3$

2. What is the maximum value of f on the interval $[0, 2]$? How about the minimum value?

   Solution: We look for points where $\frac{\mathrm{d}f}{\mathrm{d}x} = 0$; in this case, solutions to $3x^2 - 3 = 0$ are $x = \pm 1$. The second derivative, $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = 6x$. Since $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}|_{x=1} > 0$, in the interval $[0, 2]$, the minimum value of 5 is attained at $x = 1$. The maximum value must occur at one of the endpoints. As $f(0) = 7$ and $f(2) = 9$, the maximum value of 9 is attained at $x = 2$.

3. What are the minimum and maximum values of f on the interaval $[-2, 0]$? At what points are they attained?

   Solution: In this case, we observe that as $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}|_{x=-1} < 0$, the maximum value of 9 is attained at $x = -1$ in the interval $[-2, 0]$. The minimum value must occur at one of the endpoints. As $f(-2) = 5$ and $f(0) = 7$, the minimum value of 5 is attained at $x = -2$.

Now consider the function $f(x, y) = x^3 + y^2 + xy$ and answer the following questions.

1. What is the gradient of f at the point $(2, 3)$?

   Solution:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 3x^2 + y \\ 2y + x \end{pmatrix}$$

   At $(2, 3)$, we have $\nabla f|_{(2,3)} = \begin{pmatrix} 15 & 8 \end{pmatrix}^\mathsf{T}$.

2. What is the Hessian of f at the point $(0, 0)$?

   Solution: Let $\mathbf{H}$ denote the Hessian of f.

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 6x & 1 \\ 1 & 2 \end{pmatrix}$$

   At $(0, 0)$ the Hessian is $\begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$.

3. List all the critical points of f. What can you say about each of them?

   Solution: We look for solutions to $\nabla f = \begin{pmatrix} 0 & 0 \end{pmatrix}^\mathsf{T}$. These are given at points where $6x^2 = x$ and $y = -3x^2$. Thus, we get the points $(0, 0)$ and $({}^1/_6, {}^{-1}/_{12})$.

We now look at the Hessians at these two points. First at $(0,0)$, the Hessian as calculated in the previous part is $\begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$. The eigenvalues are solutions of

$$\det\left(\begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix} - \lambda \mathbf{I}\right) = 0,$$

where $\mathbf{I}$ is the 2×2 identity matrix. It can be checked that one eigenvalue is positive and the other negative. Thus, $(0,0)$ is a saddle point.

At the point $({}^1/_6, {}^{-1}/_{12})$, the Hessian is given by $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. The eigenvalues are solutions of

$$\det\left(\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} - \lambda \mathbf{I}\right) = 0.$$

In this case, both eigenvalues are positive. Thus, the Hessian at this point is positive definite and hence $({}^1/_6, {}^{-1}/_{12})$ is a local minimum.

It can be easily checked that when $x \to \infty$ and $y \to \infty$, $f \to \infty$, and for $y = 0$ and $x \to -\infty$, $f \to -\infty$. Thus, global optima are not attained at any point in $\mathbb{R}^2$.

Let

$$f(x,y,z) = (x+1)^2 + (y+2)^2 + (z-2)^2$$
$$g = x^2 + y^2 + z^2 - 36.$$

Using the method of Lagrange multipliers, find the critical points of $f(x,y,z)$ subject to $g(x,y,z) = 0$.
Solution: The corresponding Lagrangian is

$$\Lambda(x,y,z,\lambda) = f(x,y,z) + \lambda g(x,y,z).$$

Now

$$\frac{\partial \Lambda}{\partial x} = 2(x+1) + 2\lambda x \qquad \frac{\partial \Lambda}{\partial y} = 2(y+2) + 2\lambda y$$
$$\frac{\partial \Lambda}{\partial z} = 2(z-2) + 2\lambda z \qquad \frac{\partial \Lambda}{\partial \lambda} = x^2 + y^2 + z^2 - 36$$

Finding the critical points of Λ reduces to solving the following system of equations:

$$2\,(x+1) + 2\lambda x = 0 \qquad 2\,(y+2) + 2\lambda y = 0$$
$$2\,(z-2) + 2\lambda z = 0 \qquad x^2 + y^2 + z^2 - 36 = 0$$

Solving for x, y, z, we obtain $x = -1/(\lambda+1)$, $y = -2/(\lambda+1)$ and $z = 2/(\lambda+1)$. (This requires $\lambda \neq -1$, which is the case since otherwise $2(x+1) + 2 \cdot (-1)x = 2 = 0$.) Substituting x, y, z, we obtain

$$\left(\frac{-1}{\lambda+1}\right)^2 + \left(\frac{-2}{\lambda+1}\right)^2 + \left(\frac{2}{\lambda+1}\right)^2 = 36.$$

Simplifying, we get $9/(\lambda+1)^2 = 36$, i.e., $(\lambda+1)^2 = 1/4$, and henc $\lambda + 1 = \pm 1/2$. So there are two values for λ to consider, $\lambda_1 = -1/2$ and $\lambda_2 = -3/2$. Consequently, critical points of the original problem are $\mathbf{p}_1 = (-2, -4, 4)$ and $\mathbf{p}_2 = (2, 4, -4)$. It is now easy to check that $f(\mathbf{p}_1) = 9$ and $f(\mathbf{p}_2) = 81$, so f has a minimum at $\mathbf{p}_1$ and a maximum at $\mathbf{p}_2$.

3 Probability Background

Probability Distributions

Write down the probability density/mass functions for the following distributions: Uniform over $[-1, 1]$, Univariate Gaussian, Laplace, Bernoulli, Binomial, Multivariate Gaussian. If necessary, please consult Murphy (2012, Chap 2) (or Wikipedia) and familiarize yourself with these distributions. If you are already familiar with these, go ahead and look for new interesting ones in Murphy (2012)!

Bayes' Rule

Write down Bayes' rule (look it up if necessary). Then answer the following questions.

1. Suppose an unbiased coin is tossed 5 times. Let A denote the event that the first toss resulted in heads and B the event that overall there were 3 heads and 2 tails. What is $\mathbb{P}\left[A|B\right]$?

   Solution: Using Bayes' rule we write,

   $$\mathbb{P}\left[A|B\right] = \frac{\mathbb{P}\left[B|A\right] \cdot \mathbb{P}\left[A\right]}{\mathbb{P}\left[B\right]}$$

   Clearly $\mathbb{P}\left[A\right] = {}^1/_2$. Also, $\mathbb{P}\left[B\right] = \binom{5}{3} \cdot \frac{1}{2^5}$ (there must be three heads and two tails, thus the term $\binom{5}{3}$). Finally, $\mathbb{P}\left[B|A\right] = \binom{4}{2} \cdot \frac{1}{2^4}$ (conditioned on the first toss being heads, we need 2 out of the remaining 4 to be heads). Performing the calculation, we get $\mathbb{P}\left[A|B\right] = {}^3/_5$.

2. Suppose your friend throws a fair (unbiased) coin and a coin with both sides having heads into a hat. She then pulls one out (assume that both have an equal chance of being chosen) tosses it and reports the outcome as heads. Conditioned on this event what is the probability that she pulled out the biased coin? What if the outcome had been tails?

   Solution: Let A be the event that the biased coin was chosen and B be the event that the outcome of the coin toss was heads. Thus, what is being asked is $\mathbb{P}\left[A|B\right]$. Using Bayes' rule we write,

   $$\mathbb{P}\left[A|B\right] = \frac{\mathbb{P}\left[B|A\right] \cdot \mathbb{P}\left[A\right]}{\mathbb{P}\left[B\right]}$$

Now $\mathbb{P}[A] = {}^1/_2$ (both coins are equally likely to be chosen) and $\mathbb{P}[B] = {}^1/_2 + ({}^1/_2)^2$. $\mathbb{P}\left[B|A\right] = 1$ since if the biased coin is chosen then we always get heads. Performing the calculations, we get $\mathbb{P}\left[A|B\right] = {}^2/_3$. If we denote by $\neg B$ the event that the outcome was tails, we have $\mathbb{P}\left[\neg B|A\right] = 0$, thus, $\mathbb{P}\left[A|\neg B\right] = 0$.

4 Statistics Background

The average scores on the questions on this sheet of a group of 5 students (obviously not from this class!) were $\{6, 7, 5, 3, 8\}$ on a scale of 0 to 10. Answer the questions below. It would be a good idea to try to do the calculations in python to get used to the language if you aren't familiar with it.

1. What is the mean and median score? What is the variance and standard deviation?

   Solution: The mean is 5.8, median 6, variance approximately 2.96 and standard deviation approximately 1.72.

2. Suppose you were asked to standardize these so that the mean is 0 and variance is 1, what transformation would you apply?

   Solution: Let $\{x_1, \ldots, x_5\}$ be the set of observations. Let $\bar{x} = (x_1 + \cdots + x_5)/5$ and $\overline{V} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2$. Then $y_i = (x_i - \bar{x})/\overline{V}$ is the transformation yielding the required property.

3. Is the estimator for the mean you used unbiased? How about the one for the variance?

   Solution: The estimator for the mean is indeed unbiased. If x_i are drawn independently from some distribution D with mean μ, then $\mathbb{E}[\bar{x}] = \mu$. On the other hand, the estimator for variance is not unbiased, calculations show that $\mathbb{E}[\overline{V}] = \frac{n-1}{n}\sigma^2$, where σ^2 is the true variance.

References

Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective.* MIT Press, 2012.