CS M146: Introduction to Machine Learning
Prof. Aditya Grover

UCLA Fall Quarter 2023
Release: 28 Sep 2023

Homework 0

Due: 5 Oct 2023, Thursday, before 11:59 pm

Except for the programming questions, you should be able to solve these problems by hand.

Problem 1 (Multivariate Calculus)

Consider $y = x\sin(z)e^{-x}$. What is the partial derivative of y with respect to x? **Solution:**

$$\begin{aligned}\frac{\partial y}{\partial x} &= \sin(z)\left(-xe^{-x} + e^{-x}\right) \\ &= (1-x)\sin(z)e^{-x}\end{aligned}$$

Problem 2 (Linear Algebra)

(a) Consider the matrix $\boldsymbol{X}$ and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

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

i. What is the inner product $\mathbf{y}^T\mathbf{z}$? **Solution:**

$$\mathbf{y}^T\mathbf{z} = 1 \cdot 2 + 3 \cdot 3 = 11$$

ii. What is the product $\boldsymbol{X}\mathbf{y}$? **Solution:**

$$\boldsymbol{X}\mathbf{y} = \begin{pmatrix} 2 \cdot 1 + 4 \cdot 3 \\ 1 \cdot 1 + 3 \cdot 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 10 \end{pmatrix}$$

iii. Is $\boldsymbol{X}$ invertible? If so, give the inverse; if not, explain why not. **Solution:** $\det(\boldsymbol{X}) = 2 \cdot 3 - 4 \cdot 1 = 2 \neq 0$ so $\boldsymbol{X}$ is invertible.
For a 2×2 matrix,

$$\boldsymbol{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{\det(\boldsymbol{A})}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Thus,

$$\boldsymbol{X}^{-1} = \frac{1}{2}\begin{pmatrix} 3 & -4 \\ -1 & 2 \end{pmatrix}.$$

iv. What is the rank of $\boldsymbol{X}$? **Solution:** The rows (or equivalently, columns) of $\boldsymbol{X}$ are linearly independent, so $\text{rank}(\boldsymbol{X}) = 2$.

(b) **Vector Norms** [4 pts]
Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with following norms (you can hand draw or use software for this question): **Solution:** Each l_p-norm corresponds to a unit ball. In $\mathbb{R}^2$, this corresponds to a disc of diameter 2 centered at the origin.

i. $||\mathbf{x}||_2 \leq 1$ (Recall $||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$.) **Solution:** the unit circle (origin-centered circle with diameter of 2), and its interior

ii. $||\mathbf{x}||_0 \leq 1$ (Recall $||\mathbf{x}||_0 = \sum_{i:x_i \neq 0} 1$.) **Solution:** the axes

iii. $||\mathbf{x}||_1 \leq 1$ (Recall $||\mathbf{x}||_1 = \sum_i |x_i|$.) **Solution:** origin-centered diamond with diameter of 2, and its interior

iv. $||\mathbf{x}||_\infty \leq 1$ (Recall $||\mathbf{x}||_\infty = \max_i |x_i|$.) **Solution:** origin-centered square with side length of 2, and its interior

(c) **Matrix Decompositions [6 pts]**

i. Give the definition of the eigenvalues and the eigenvectors of a square matrix. **Solution:** An *eigenvector* of a square matrix is a vector that points in a direction that is invariant under the associated linear transformation. In other words—, if $\mathbf{v}$ is a vector that is not zero, then it is an eigenvector of a square matrix $\boldsymbol{A}$ if $\boldsymbol{A}\mathbf{v}$ is a scalar multiple of $\mathbf{v}$. This condition could be written as the equation

$$\boldsymbol{A}\mathbf{v} = \lambda\mathbf{v},$$

where λ is a number (scalar) known as the *eigenvalue* associated with the eigenvector $\mathbf{v}$.

ii. Find the eigenvalues and eigenvectors of

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Solution: The eigenvectors $\mathbf{v}$ of this transformation satisfy the equation $\boldsymbol{A}\mathbf{v} = \lambda\mathbf{v}$. Rearrange this equation to obtain $(\boldsymbol{A} - \lambda\boldsymbol{I})\mathbf{v} = 0$, which has a solution only when its determinant $|\boldsymbol{A} - \lambda\boldsymbol{I}|$ equals zero. Set the determinant to zero to obtain the polynomial equation,

$$p(\lambda) = |\boldsymbol{A} - \lambda\boldsymbol{I}| = (2-\lambda)^2 - 1 = (\lambda^2 - 4\lambda + 4) - 1 = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1) = 0,$$

known as the characteristic polynomial of the matrix $\boldsymbol{A}$. In this case, it has the roots $\lambda = 1$ and $\lambda = 3$.
For $\lambda = 1$, the equation becomes,

$$(\boldsymbol{A} - \boldsymbol{I})\mathbf{v} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which has the solution,

$$\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

For $\lambda = 3$, the equation becomes,

$$(\boldsymbol{A} - 3\boldsymbol{I})\mathbf{w} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which has the solution,

$$\mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Thus, the vectors $\mathbf{v}$ and $\mathbf{w}$ are eigenvectors of $\boldsymbol{A}$ associated with the eigenvalues $\lambda = 1$ and $\lambda = 3$, respectively.

iii. For any positive integer k, show that the eigenvalues of $\boldsymbol{A}^k$ are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$, the k^{th} powers of the eigenvalues of matrix $\boldsymbol{A}$, and that each eigenvector of $\boldsymbol{A}$ is still an eigenvector of $\boldsymbol{A}^k$. **Solution:** We will prove by induction.
Base case (given): $\boldsymbol{A}\mathbf{v} = \lambda\mathbf{v}$
Inductive step: Prove that if $\boldsymbol{A}^k\mathbf{v} = \lambda^k\mathbf{v}$, then $\boldsymbol{A}^{k+1}\mathbf{v} = \lambda^{k+1}\mathbf{v}$.
This holds because $\boldsymbol{A}^{k+1}\mathbf{v} = \boldsymbol{A}\left(\boldsymbol{A}^k\mathbf{v}\right) = \boldsymbol{A}\left(\lambda^k\mathbf{v}\right) = \lambda^k\left(\boldsymbol{A}\mathbf{v}\right) = \lambda^{k+1}\mathbf{v}$.
Since both the basis and the inductive step have been performed, by mathematical induction, the statement $\boldsymbol{A}^k\mathbf{v} = \lambda^k\mathbf{v}$ holds for all positive integer k. Q.E.D.

(d) **Vector and Matrix Calculus [5 pts]**
Consider the vectors $\mathbf{x}$ and $\mathbf{a}$ and the symmetric matrix $\boldsymbol{A}$.

i. What is the first derivative of $\mathbf{a}^T\mathbf{x}$ with respect to $\mathbf{x}$? **Solution:** Let a_i and x_i denote the elements of $\mathbf{a}$ and $\mathbf{x}$, respectively. Then $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = \sum_{i=1}^n a_i x_i$ so $\frac{\partial f(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^n a_i x_i = a_k$ and $\nabla_{\mathbf{x}}\mathbf{a}^T\mathbf{x} = \mathbf{a}$.

ii. What is the first derivative of $\mathbf{x}^T\boldsymbol{A}\mathbf{x}$ with respect to $\mathbf{x}$? What is the second derivative? **Solution:** Let a_{ij} denote the element in row i, column j of $\boldsymbol{A}$. Then $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{A}\mathbf{x} = \sum_{i=1}^n\sum_{j=1}^n x_i a_{ij} x_j$ so

$$\begin{aligned}
\frac{\partial f(\mathbf{x})}{\partial x_k} &= \frac{\partial}{\partial x_k}\sum_{i=1}^n\sum_{j=1}^n a_{ij}x_i x_j \\
&= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j\neq k} a_{ij}x_i x_j + \sum_{i\neq k} a_{ik}x_i x_k + \sum_{j\neq k} a_{kj}x_k x_j + a_{kk}x_k^2\right] \\
&= 0 + \sum_{i\neq k} a_{ik}x_i + \sum_{j\neq k} a_{kj}x_j + 2a_{kk}x_k \\
&= \sum_{i=1}^n a_{ik}x_i + \sum_{j=1}^n a_{kj}x_j = 2\sum_{i=1}^n a_{ki}x_i
\end{aligned}$$

where the last equality follows from $a_{ij} = a_{ji}$. Note that the k^{th} entry of $\nabla_{\mathbf{x}} f(\mathbf{x})$ is just the inner product of the k^{th} row of $\boldsymbol{A}$ and $\mathbf{x}$; therefore, $\nabla_{\mathbf{x}}\mathbf{x}^T\boldsymbol{A}\mathbf{x} = 2\boldsymbol{A}\mathbf{x}$.
To find the second derivative,

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_k x_\ell} = \frac{\partial}{\partial x_k}\left[\frac{\partial f(\mathbf{x})}{\partial x_\ell}\right] = \frac{\partial}{\partial x_k}\left[2\sum_{i=1}^n a_{\ell i}x_i\right] = 2a_{\ell k} = 2a_{k\ell}$$

Therefore, $\nabla_{\mathbf{x}}^2\mathbf{x}^T\boldsymbol{A}\mathbf{x} = 2\boldsymbol{A}$.

(e) **Geometry [5 pts]**

i. Show that the vector $\mathbf{w}$ is orthogonal to the line $\mathbf{w}^T\mathbf{x} + b = 0$. (Hint: Consider two points $\mathbf{x}_1, \mathbf{x}_2$ that lie on the line. What is the inner product $\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2)$?) **Solution:** This line is all $\mathbf{x}$ such that $\mathbf{w}^T\mathbf{x} + b = 0$. Consider two such points, called $\mathbf{x}_1$ and $\mathbf{x}_2$. Note that $\mathbf{x}_1 - \mathbf{x}_2$ is a vector parallel to our line. Also note that

$$\mathbf{w}^T\mathbf{x}_1 + b = 0 = \mathbf{w}^T\mathbf{x}_2 + b \implies \mathbf{w}^T\mathbf{x}_1 = \mathbf{w}^T\mathbf{x}_2 \implies \mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0,$$

which shows that the vector $\mathbf{w}$ is orthogonal to our line.

ii. Argue that the distance from the origin to the line $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{b}{||\mathbf{w}||_2}$. **Solution:** We can show this by first finding the closest point to the origin that lies on this line, and then finding the distance to this point. Let $\mathbf{a}^*$ be the closest point to the origin the lies on the line. We can write $\mathbf{a}^*$ as

$$\begin{aligned} \mathbf{a}^* = \quad & \min \quad \mathbf{a}^T\mathbf{a} \\ & \text{s.t.} \quad \mathbf{w}^T\mathbf{a} + b \;=\; 0 \end{aligned}$$

So we first solve this constrained optimization problem and find $\mathbf{a}^*$. We start by taking the derivative of the objective, setting it to zero, and using Lagrange multipliers (i.e. setting the derivative of the Lagrangian $\mathbf{a}^T\mathbf{a} - \lambda(\mathbf{w}^T\mathbf{a} + b)$ to zero). We can write

$$\nabla_{\mathbf{a}} \left[\mathbf{a}^T\mathbf{a} - \lambda(\mathbf{w}^T\mathbf{a} + b)\right] = 2\mathbf{a} - \lambda\mathbf{w} = 0$$

to find that $\mathbf{a}^* = \frac{\lambda}{2}\mathbf{w}$. Hence, plugging this value for $\mathbf{a}$ into the constraint, we can write

$$\mathbf{w}^T\mathbf{a} + b = \mathbf{w}^T\left(\frac{\lambda}{2}\mathbf{w}\right) + b = \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + b = 0$$

Thus,

$$\lambda = \frac{-2b}{\mathbf{w}^T\mathbf{w}}$$
$$\mathbf{a}^* = \frac{-b}{\mathbf{w}^T\mathbf{w}}\mathbf{w}$$

Once we have $\mathbf{a}^*$, we can compute the distance between $\mathbf{a}^*$ and the origin to get

$$\textit{distance} = ||\mathbf{a}|| = \sqrt{(\mathbf{a}^*)^T\mathbf{a}^*} = \sqrt{\left(\frac{-b}{\mathbf{w}^T\mathbf{w}}\right)^2 \mathbf{w}^T\mathbf{w}} = \frac{b}{\mathbf{w}^T\mathbf{w}}\sqrt{\mathbf{w}^T\mathbf{w}} = \frac{b}{\sqrt{\mathbf{w}^T\mathbf{w}}} = \frac{b}{||\mathbf{w}||}$$

Problem 3 (Probability and Statistics)

(a) Consider a sample of data S obtained by flipping a coin five times. $X_i, i \in \{1, \ldots, 5\}$ is a random variable that takes a value 0 when the outcome of coin flip i turned up heads, and 1 when it turned up tails. Assume that the outcome of each of the flips does not depend on the outcomes of any of the other flips. The sample obtained $S = (X_1, X_2, X_3, X_4, X_5) = (1, 1, 0, 1, 0)$.

i. What is the sample mean for this data? **Solution:** $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{3}{5}$

ii. What is the unbiased sample variance ? **Solution:** $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{4}\left(3\left(1-\frac{3}{5}\right)^2 + 2\left(0-\frac{3}{5}\right)^2\right) = \frac{1}{4}\left(3\cdot\frac{4}{25} + 2\cdot\frac{9}{25}\right) = \frac{3}{10}$
Give partial credit if student uses the biased estimator – $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$

iii. What is the probability of observing this data assuming that a coin with an equal probability of heads and tails was used? (*i.e.*, The probability distribution of X_i is $P(X_i = 1) = 0.5$, $P(X_i = 0) = 0.5$.) **Solution:** $P(S) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$

iv. Note the probability of this data sample would be greater if the value of the probability of heads $P(X_i = 1)$ was not 0.5 but some other value. What is the value that maximizes the probability of the sample S? [Optional: Can you prove your answer is correct?]
Solution: To maximize the probability of sample S, p should take on the sample mean: $p^* = \frac{3}{5}$.
Optional: Let $p = P(X = 1)$. We wish to find the value of p, $0 \le p \le 1$ that maximizes the likelihood $L(p) = p^3(1-p)^2$. To find the critical points, find p^* such that $\frac{dL}{dp}\Big|_{p^*} = 0$.

$$\frac{dL}{dp} = -2p^3(1-p) + 3p^2(1-p)^2 = p^2(1-p)(-2p + 3(1-p)) = p^2(1-p)(3-5p)$$
$$p^* = 0, \frac{3}{5}, 1$$

Thus, L attains a minimum of 0 at $p = 0, 1$, and L attains a maximum of $\left(\frac{3}{5}\right)^3\left(\frac{2}{5}\right)^2 = \frac{108}{3125}$ at $p = \frac{3}{5}$. (We could also have computed $\frac{d^2L}{dp^2}$ to determine the minimum and maximum.)

v. Given the following joint distribution between X and Y, what is $P(X = T|Y = b)$?

$P(X,Y)$		Y		
		a	b	c
X	T	0.2	0.1	0.2
	F	0.05	0.15	0.3

Solution: $P(X = T|Y = b) = \frac{P(X=T,Y=b)}{P(Y=b)} = \frac{0.1}{0.1+0.15} = 0.4$.

(b) Match the distribution name to its formula.

(a) Gaussian	(i) $p^x(1-p)^{1-x}$, when $x \in \{0,1\}$; 0 otherwise
(b) Exponential	(ii) $\frac{1}{b-a}$ when $a \le x \le b$; 0 otherwise
(c) Uniform	(iii) $\binom{n}{x}p^x(1-p)^{n-x}$
(d) Bernoulli	(iv) $\lambda e^{-\lambda x}$ when $x \ge 0$; 0 otherwise
(e) Binomial	(v) $\frac{1}{\sqrt{(2\pi)\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

Solution: a-v, b-iv, c-ii, d-i, e-iii

(c) What is the mean and variance of a $Bernoulli(p)$ random variable? **Solution:** The PMF for the Bernouilli distribution is

$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$
$$= p^k(1-p)^{1-k} \text{ for } k \in \{0,1\}$$

Let $q = 1 - p$. Then

$$\mu = \mathbb{E}[X] = \sum_{k \in \{0,1\}} k f(k;p) = (0)(q) + (1)(p) = p$$
$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = (0)^2(q) + (1)^2(p) - (p)^2 = p(1-p) = pq$$
$$H = \mathbb{E}[-\ln(f(k;p))] = (-\ln q)(q) + (-\ln p)(p) = -q\ln q - p\ln p$$

(d) If the variance of a zero-mean random variable X is σ^2, what is the variance of $2X$? What about the variance of $X+2$? **Solution:** If all values are scaled by a constant, the variance is scaled by the square of that constant: $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$. Thus, $\mathrm{Var}(2X) = 4\mathrm{Var}(X) = 4\sigma^2$.

Variance is invariant with respect to changes in a location parameter; that is, if a constant is added to all values of the variable, the variance is unchanged: $\mathrm{Var}(X+a) = \mathrm{Var}(X)$. Thus, $\mathrm{Var}(X+2) = \mathrm{Var}(X) = \sigma^2$.

Problem 4 (Algorithms)

Big-O notation For each pair (f, g) of functions below, list which of the following are true: $f(n) = O(g(n))$, $g(n) = O(f(n))$, or both. Briefly justify your answers.

(a) $f(n) = ln(n), g(n) = lg(n)$. Note that ln denotes log to the base e and lg denotes log to the base 2. **Solution:** Both since both functions are equivalent upto a multiplicative constant

(b) $f(n) = 3^n, g(n) = n^{10}$ **Solution:** $g(n) = O(f(n))$ since $f(n)$ grows much more rapidly as n becomes large.

(c) $f(n) = 3^n, g(n) = 2^n$ **Solution:** $g(n) = O(f(n))$ since $f(n)$ grows much more rapidly as n becomes large.

Problem 5 (Programming Skills)

Start familiarizing yourself with the Python libraries `numpy` and `matplotlib` by completing the following exercises. You may find the following references helpful:

- https://numpy.org/doc/stable/reference/random/generated/numpy.random.multivariate_normal.html
- http://docs.scipy.org/doc/numpy/reference/generated/numpy.linalg.eig.html

You do not have to submit code, simply paste screenshots of the code and the plots in your pdf.

(a) Sampling from multivariate probability distributions

i. Draw 1000 samples $\mathbf{x} = \binom{x_1}{x_2}$ from a 2-dimensional Gaussian distribution with mean $\binom{0}{0}$ and identity covariance matrix, i.e. $p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2}} \exp\left(-\frac{||\mathbf{x}||^2}{2}\right)$, and make a scatter plot (x_1 vs x_2). **Solution:** See attached code.

ii. How does the scatter plot change if the mean is $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$? **Solution:** See solution code. The data moves up and to the left. Namely, the center of the data moves from roughly $[0, 0]$ to roughly $[-1, 1]$.

iii. How does the (original) scatter plot change if you double the variance of each component? **Solution:** See solution code. The data become more "spread out".

iv. How does the (original) scatter plot change if the covariance matrix is changed to $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$? **Solution:** See solution code. The data become skewed so that they stretch from the lower left to the upper right.

v. How does the (original) scatter plot change if the covariance matrix is changed to $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$? **Solution:** See solution code. The data become skewed so that they stretch from the upper left to the bottom right.

(b) There are now lots of really interesting data sets publicly available to play with. They range in size, quality and the type of features and have resulted in many new machine learning techniques being developed. Find a public, free, supervised (i.e. it must have features *and* labels), machine learning dataset. For example, you can use one of the open datasets released by the US government here: https://catalog.data.gov/dataset Once you have found the data set, provide the following information:

i. The name of the data set and its link.

ii. A brief (i.e. 2-3 sentences) description of the data set including what the features are and what is being predicted.

iii. The number of examples in the data set.

iv. The number of features for each example.

For this question, do not just copy and paste the description from the website or the paper; reference it, but use your own words. Your goal here is to understand the data set, where it came from, and potential issues involved.