CS 189:289A Introduction to Machine Learning
Spring 2020 Jonathan Shewchuk

HW2: I ♥ Math

Due Wednesday, February 12 at 11:59 pm

- Homework 2 is an entirely written assignment; no coding involved.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted.
- In all of the questions, **show your work**, not just the final answer.
- **Start early. This is a long assignment. Most of the material is prerequisite material not covered in lecture; you are responsible for finding resources to understand it.**

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word (submit PDF format, **not** .doc/.docx format) or submit neatly handwritten and scanned solutions. **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
   - In your write-up, please state with whom you worked on the homework.
   - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats.
     "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

1 Identities with Expectation

For this exercise, recall the following useful identity: for a probability event A, $\mathbb{P}(A) = \mathbb{E}[\mathbf{1}\{A\}]$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

1. Let X be a random variable with pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ (and zero everywhere else). Use induction on k to show that for any nonnegative integer $k \geq 0$, $\mathbb{E}[X^k] = \frac{k!}{\lambda^k}$.
   Hint: use integration by parts.

2. Assume that X is a non-negative real-valued random variable. Prove the following identity:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)dt.$$

   If you prefer, assume that X has a density $f(x)$ and a CDF $F(x)$; this might simplify notation.

3. Again assume $X \geq 0$, but now additionally let $\mathbb{E}[X^2] < \infty$. Prove the following:

$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}.$$

   Note that by assumption we know $\mathbb{P}(X \geq 0) = 1$, so this inequality is indeed quite powerful.
   Hint: Use the Cauchy–Schwarz inequality: $|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle$. You have most likely seen it applied when the inner product is the real dot product, however it holds for arbitrary inner products; without proof, use the fact that a valid inner product on the set of random variables is given by $\mathbb{E}(UV)$, for random variables U and V.

4. Now assume $\mathbb{E}[X^2] < \infty$, and additionally assume $\mathbb{E}X = 0$ (X no longer has to be non-negative). Prove the following inequality:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + t^2}, \text{ for any } t \geq 0$$

   There is no typo — compared to the previous part, the inequality is flipped.
   Hint: Use similar logic as in the previous part, and think of how to apply Cauchy–Schwarz. Use the fact that $t - X \leq (t - X)\mathbf{1}\{t - X > 0\}$.

2 Probability Potpourri

1. Recall the covariance of two random variables X and Y is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. For a multivariate random variable Z (i.e., each index of Z is a random variable), we define the covariance matrix Σ such that $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$, where μ is the mean value of the random column vector Z. Prove that the covariance matrix is always positive semidefinite (PSD).
   Hint: Use linearity of expectation.

2. The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

   (i) on a given shot there is a gust of wind and she hits her target.

   (ii) she hits the target with her first shot.

   (iii) she hits the target exactly once in two shots.

   (iv) there was no gust of wind on an occasion when she missed.

3. An archery target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Arrows striking within the inner circle are awarded 4 points, arrows within the middle ring are awarded 3 points, and arrows within the outer ring are awarded 2 points. Shots outside the target are awarded 0 points.

   Consider a random variable X, the distance of the strike from the center (in feet), and let the probability density function of X be

   $$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   What is the expected value of the score of a single strike?

4. Let $X \sim \text{Pois}(\lambda)$, $Y \sim \text{Pois}(\mu)$. given that $X \perp\!\!\!\perp Y$, derive an expression for $\mathbb{P}(X \mid X + Y = n)$. What well-known probability distribution is this? What are its parameters?

3 Properties of Gaussians

1. Prove that $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$, where $\lambda \in \mathbb{R}$ is a fixed constant, and $X \sim N(0, \sigma^2)$. As a function of λ, $\mathbb{E}[e^{\lambda X}]$ is also known as the *moment-generating function*.

2. For $t > 0$ prove that $\mathbb{P}(X \geq t) \leq \exp\left(-t^2/2\sigma^2\right)$, then show that $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-t^2/2\sigma^2\right)$.
   Hint: Consider using Markov's inequality in combination with the result of the previous part.

3. Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$ be iid (independent and identically distributed). Can you prove a similar concentration result for the average of n Gaussians: $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$? What happens as $n \to \infty$?

Hint: Without proof, use the fact that linear combinations of iid Gaussian-distributed variables are also Gaussian-distributed. Be warned that summing two Gaussian variables does **not** mean that you can sum their probability density functions.

4. Give an example of two Gaussian-distributed random variables X and Y, such that there exists a linear combination $\alpha X+\beta Y$, for some $\alpha, \beta \in \mathbb{R}$, which is *not* Gaussian-distributed. Note that examples of the kind $X \sim N(0,1), Y = -X$ and their linear combination $X + Y = 0$ *will not* be valid solutions; we will consider constant random variables as Gaussians with variance equal to 0.

5. Take two orthogonal vectors $u, v \in \mathbb{R}^n$, $u \perp v$, and let $X = (X_1, \ldots, X_n)$ be a vector of n iid standard Gaussians, $X_i \sim N(0,1), \forall i \in [n]$. Let $u_x = \langle u, X \rangle$ and $v_x = \langle v, X \rangle$. Are u_x and v_x independent?
Hint: First try to see if they are correlated; you may use the fact that jointly normal random variables are independent iff they are uncorrelated.

6. Prove that $\mathbb{E}\left[\max_{1 \le i \le n} |X_i|\right] \le C\sqrt{\log(2n)}\sigma$ for some constant $C \in \mathbb{R}$, where $X_1, \ldots, X_n \sim N(0, \sigma^2)$ are iid. (Interestingly, a similar lower bound holds: $\mathbb{E}\left[\max_{1 \le i \le n} |X_i|\right] \ge C'\sqrt{\log(2n)}\sigma$ for some C'; but you don't need to prove the lower bound).
Hint: Use Jensen's inequality: $f(\mathbb{E}[Y]) \le \mathbb{E}[f(Y)]$ for any convex function f.

4 Linear Algebra Review

1. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Prove equivalence between these three different definitions of positive semidefiniteness (PSD).

   (a) For all $x \in \mathbb{R}^n$, $x^\top A x \ge 0$.

   (b) All the eigenvalues of A are nonnegative.

   (c) There exists a matrix $U \in \mathbb{R}^{n \times n}$ such that $A = UU^\top$.

   Mathematically, we write positive semidefiniteness as $A \succeq 0$.

2. Now that we're equipped with different definitions of positive semidefiniteness, use them to prove the following properties of PSD matrices.

   (a) If A and B are PSD, then $2A + 3B$ is PSD.

   (b) If A is PSD, all diagonal entries of A are nonnegative: $A_{ii} \ge 0, \forall i \in [n]$.

   (c) If A is PSD, the sum of all entries of A is nonnegative: $\sum_{j=1}^n \sum_{i=1}^n A_{ij} \ge 0$.

   (d) If A and B are PSD, then $\mathrm{Tr}(AB) \ge 0$, where $\mathrm{Tr}(M)$ denotes the *trace* of M.

   (e) If A and B are PSD, then $\mathrm{Tr}(AB) = 0$ if and only if $AB = 0$.

3. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric, PSD matrix. Write $\|A\|_F$ as a function of the eigenvalues of A.

4. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Prove that the largest eigenvalue of A is

$$\lambda_{\max}(A) = \max_{\|x\|_2 = 1} x^\top A x.$$

5 Gradients and Norms

1. Define the ℓ_p-norm as $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, where $x \in \mathbb{R}^n$. Prove that the ℓ_1, ℓ_2, ℓ_∞ norms are all within a constant factor of one another. The Cauchy–Schwarz inequality is helpful here.

2. (a) Let $\alpha = \sum_{i=1}^n y_i \ln \beta_i$ for $y, \beta \in \mathbb{R}^n$. What are the partial derivatives $\frac{\partial \alpha}{\partial \beta_i}$?

   (b) Let $\beta = \sinh \gamma$ for $\gamma \in \mathbb{R}^n$ (treat the sinh as an element-wise operation; i.e. $\beta_i = \sinh \gamma_i$). What are the partial derivatives $\frac{\partial \beta_i}{\partial \gamma_j}$?

   (c) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^n$, $\rho \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$. What are the the partial derivatives $\frac{\partial \gamma_i}{\partial \rho_j}$?

   (d) Let $f(x) = \sum_{i=1}^n y_i \ln(\sinh(Ax + b)_i)$; $A \in \mathbb{R}^{n \times m}$, $y \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ are given. What are the partial derivatives $\frac{\partial f}{\partial x_j}$?
   Hint: Use the chain rule.

3. Consider a linear decision function $f(x) = w \cdot x + \alpha$ and the hyperplane decision boundary $H = \{x : w \cdot x = -\alpha\}$. Prove that if w is a unit vector, then the *signed distance* (the ℓ_2-norm distance with an appropriate sign) from x to the closest point on H is $w \cdot x + \alpha$.

4. Consider the function which maps a vector to its maximum entry, $x \mapsto \max_i x_i$. While this function is non-smooth, a common trick in machine learning is to use a smooth approximation, *LogSumExp*, defined as follows.

$$\text{LSE} : \mathbb{R}^n \to \mathbb{R}, \text{LSE}(x) = \ln\left(\sum_{i=1} e^{x_i}\right).$$

   One of the nice properties of this function is that it is convex, which can be proved by showing its Hessian matrix is positive semidefinite. To that end, compute its gradient and Hessian.

5. Let $X \in \mathbb{R}^{n \times d}$ be a data matrix, consisting of n samples, each of which has d features, and let $y \in \mathbb{R}^n$ be a vector of outcomes. We wish to find the *best linear approximation*, i.e. we want to find the θ that minimizes the loss $L(\theta) = \|y - X\theta\|_2^2$. Assuming X has full column rank, compute $\theta^* = \text{argmin}_\theta L(\theta)$ in terms of X and y.

6 Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x$, where $A \in \mathbb{R}^{n \times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$.

1. Find the optimizer x^*.
2. Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1.
3. Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.
4. Using exercise 4 in Problem 4, prove $\|Ax\|_2 \leq \lambda_{\max(A)} \|x\|_2$.
   Hint: Use the fact that, if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.
5. Using the previous two parts, show that for some $0 < \rho < 1$,
$$\|x^{(k)} - x^*\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$
6. Let $x^{(0)} \in \mathbb{R}^n$ be the starting value for our gradient descent iterations. If we want a solution $x^{(k)}$ that is $\epsilon > 0$ close to x^*, i.e. $\|x^{(k)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of ρ, $\|x^{(0)} - x^*\|_2$, and ϵ.