CS 189 Spring 2019 | Introduction to Machine Learning Jonathan Shewchuk

HW2

Due: Wednesday, February 13 at 11:59 pm

Deliverables:

1. Submit a PDF of your homework, **with an appendix listing all your code**, 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 inadverdently 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 $\mathbb{E}X^k = \frac{k!}{\lambda^k}$.
   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 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. Prove that $\mathbb{P}(X \geq t) \leq \exp\left(-t^2/2\sigma^2\right)$, and conclude 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. 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 (under some regularity, which is satisfied for iid Gaussians) linear combinations of Gaussians are also Gaussian.

4. Give an example of two Gaussian 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. 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 $\mathbb{E}\left[\max_{1\le i\le n}|X_i|\right] \le C\sqrt{\log(2n)}\sigma$, where $X_1, \ldots, X_n \sim N(0, \sigma^2)$ are iid. In fact, a stronger version of this claim holds - $\mathbb{E}\left[\max_{1\le i\le n}|X_i|\right] \ge C'\sqrt{\log(2n)}\sigma$ for some C' (you don't need to prove the lower bound).
   Hint: Use Jensen's inequality, which says that $f(\mathbb{E}[Y]) \le \mathbb{E}[f(Y)]$, for any convex function f. Take $f(Y) = e^Y$, and use exercise 1 of this Problem.

3 Linear Algebra Review

1. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Prove equivalence between the following different definitions of positive semi-definiteness (PSD):

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

   (b) All eigenvalues of A are non-negative.

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

   Mathematically, we write positive semi-definiteness as $A \succeq 0$.

2. Now that we're equipped with different definitions of positive semi-definiteness, 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 non-negative, $A_{ii} \ge 0, \forall i \in [n]$.

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

   (d) If A and B are PSD, then $\mathrm{Tr}(AB) \ge 0$.

   (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 matrix. Prove that the largest eigenvalue of A is

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

4 Gradients and Norms

1. Define ℓ_p norms as $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, where $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$.
   Hint: For the second inequality, consider applying the Cauchy-Schwarz inequality.

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

3. Let $X, A \in \mathbb{R}^{n\times n}$ (not necessarily symmetric). Compute $\nabla_X \operatorname{Tr}(A^\top X)$.

4. Consider the optimization problem $\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Ax - b^\top x$, where $A \in \mathbb{R}^{n\times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \le \lambda_{\max}(A) < 1$.
   (a) Find the optimizer x^*.
   (b) 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.
   (c) Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.
   (d) Using exercise 3 in Problem 3, prove $\|Ax\|_2 \le \lambda_{\max(A)}\|x\|_2$.
   Hint: Use the fact that, if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.
   (e) Using the previous two parts, show that for some $0 < \rho < 1$,
   $$\|x^{(k)} - x^*\|_2 \le \rho\|x^{(k-1)} - x^*\|_2.$$
   (f) 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. $\|x^{(k)} - x^*\|_2 \le \epsilon$, 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.

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. For example, each row of X could have information about a house on the market, like its area, number of floors, number of bathrooms/bedrooms, etc., and each entry of y could be the price of that house. We are interested in building a model that predicts house prices from the set of its features, as listed above. Suppose that domain knowledge tells us that the relationship between the features and outcomes is linear; ideally, there exists a set of parameters $\theta \in \mathbb{R}^d$ such that $X\theta = y$. However, n is large and there is noise

in the acquisition of X and y, so this system is overdetermined. Still, 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 $\theta^* = \arg\min_\theta L(\theta)$ in terms of X and y.

5 Covariance Practice

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

2. Let X be a multivariate random variable (recall, this means it is a vector of random variables) with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Let Σ have one zero eigenvalue. Prove the space where X takes values with non-zero probability (this space is called the support of X) has dimension $n-1$. How could you construct a new $\tilde{X}$ so that no information is lost from the original distribution but the covariance matrix of $\tilde{X}$ has no zero eigenvalues? What would $\tilde{X}$ look like if Σ has $m \le n$ zero eigenvalues?
   Hint: use the identity $\mathrm{Var}(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(Y_i, Y_j)$.