CS 189:289A Spring 2017 Introduction to Machine Learning

Homework 2: I ♥ Math Due Mon., February 13

Instructions

- We prefer that you typeset your answers using LaTeX or other word processing software. Neatly handwritten and scanned solutions will also be accepted.
- Please make sure to start **each question on a new page**, as grading (with Gradescope) is much easier that way! Write your name and student ID near the top of each page.
- Submit a **PDF of your writeup** to the Homework 2 assignment on Gradescope.
- You should be able to see CS 189/289A on Gradescope when you login with your bCourses email address. Please make a Piazza post if you have any problems accessing Gradescope.
- The assignment covers concepts in probability, linear algebra, matrix calculus, optimization, and decision theory.
- **Start early. This is a long assignment. Some of the material may not have been covered in lecture; you are responsible for finding resources to understand it.**
- Due **Monday, February 13, 2017 at 11:59 PM.**

Q1. Conditional Probability

In the following questions, **show your work**, not just the final answer.

(a) 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.

(b) Let A, B, C be events. Show that if

$$P(A \mid B, C) > P(A \mid B)$$

then

$$P(A \mid B, C^c) < P(A \mid B),$$

where C^c denotes the complement of C. Assume that each event on which we are conditioning has positive probability.

Q2. Positive Definiteness

Definition. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix.

- We say that A is **positive definite** if $\forall x \in \mathbb{R}^n - \{0\},\ x^\top A x > 0$. We denote this with $A \succ 0$.
- Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n,\ x^\top A x \geq 0$. We denote this with $A \succeq 0$.

(a) For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, prove that all of the following are equivalent.

(i) $A \succeq 0$.

(ii) $B^\top A B \succeq 0$, for some invertible matrix $B \in \mathbb{R}^{n \times n}$.

(iii) All the eigenvalues of A are nonnegative.

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

(Suggested road map: (i) $\Leftrightarrow$ (ii), (i) $\Rightarrow$ (iii) $\Rightarrow$ (iv) $\Rightarrow$ (i). For the implication (iii) $\Rightarrow$ (iv) use the *Spectral Theorem for Symmetric Matrices*.)

(b) For a symmetric positive definite matrix $A \succ 0 \in \mathbb{R}^{n \times n}$, prove the following.

(i) For every $\lambda > 0$, we have that $A + \lambda I \succ 0$.

(ii) There exists a $\gamma > 0$ such that $A - \gamma I \succ 0$.

(iii) All the diagonal entries of A are positive; i.e., $A_{ii} > 0$ for $i = 1, \ldots, n$.

(iv) $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$, where A_{ij} is the element at the i-th row and j-th column of A.

Q3. Derivatives and Norms

In the following questions, **show your work**, not just the final answer.

(a) Let $x, a \in \mathbb{R}^n$. Compute $\nabla_x \left(a^\top x\right)$.

(b) Let $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$. Compute $\nabla_x \left(x^\top A x\right)$.
How does the expression you derived simplify in the case that A is symmetric?

(Hint: to get a feeling for the problem, explicitly write down a 2×2 or 3×3 matrix A with components A_{11}, A_{12}, etc., explicitly expand $x^\top A x$ as a polynomial without matrix notation, calculate the gradient in the usual way, and put the result back into matrix form. Then generalize the result to the $n \times n$ case.)

(c) Let $A, X \in \mathbb{R}^{n \times n}$. Compute $\nabla_X \left(\text{trace}\left(A^\top X\right)\right)$.

(d) For a function $f : \mathbb{R}^d \to \mathbb{R}$ to be a norm, the distance metric $\delta(x, y) = f(x - y)$ must satisfy the triangle inequality. Is the function $f(x) = (\sqrt{|x_1|} + \sqrt{|x_2|})^2$ a norm for vectors $x \in \mathbb{R}^2$? Prove it or give a counterexample.

(e) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty$.

(f) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$.
(Hint: The Cauchy–Schwarz inequality may come in handy.)

Q4. Eigenvalues

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $A \succeq 0$.

(a) Prove that the largest eigenvalue of A is

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

(Hint: Use the *Spectral Theorem for Symmetric Matrices* to reduce the problem to the diagonal case.)

(b) Similarly, prove that the smallest eigenvalue of A is

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

(c) Is either of the optimization problems described in parts (a) and (b) a convex program? Justify your answer.

(d) Show that if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2, and deduce that

$$\lambda_{\max}(A^2) = \lambda_{\max}(A)^2 \text{ and } \lambda_{\min}(A^2) = \lambda_{\min}(A)^2.$$

(e) From parts (a), (b), and (d), show that for any vector $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$,

$$\lambda_{\min}(A) \leq \|Ax\|_2 \leq \lambda_{\max}(A).$$

(f) From part (e), deduce that for any vector $x \in \mathbb{R}^n$,

$$\lambda_{\min}(A)\,\|x\|_2 \leq \|Ax\|_2 \leq \lambda_{\max}(A)\,\|x\|_2.$$

Q5. Gradient Descent

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

(a) Using the first-order optimality conditions, derive a closed-form solution for the minimum possible value of x, which we denote 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) Show that for some $0 < \rho < 1$,

$$\|x^{(k)} - x^*\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$

(e) Let $x^{(0)} \in \mathbb{R}^n$ be a starting value for our gradient descent iterations. If we want our solution $x^{(k)}$ to be $\epsilon > 0$ 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 $\rho, \|x^{(0)} - x^*\|_2$, and ϵ. Note that $0 < \rho < 1$, so $\log \rho < 0$.

(f) Observe that the running time of each iteration of gradient descent is dominated by a matrix-vector product. What is the overall running time of gradient descent to achieve a solution $x^{(k)}$ which is ϵ-close to x^*? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2, \epsilon$, and n.

Q6. Classification

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional "doubt" category labeled $c + 1$. Let $f : \mathbb{R}^d \to \{1, \ldots, c + 1\}$ be a decision rule. Define the loss function

$$L(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \ldots, c\}, \\ \lambda_r & \text{if } i = c + 1, \\ \lambda_s & \text{otherwise,} \end{cases}$$

where $\lambda_r \geq 0$ is the loss incurred for choosing doubt and $\lambda_s \geq 0$ is the loss incurred for making a misclassification. Hence the risk of classifying a new data point x as class $i \in \{1, 2, \ldots, c + 1\}$ is

$$R(f(x) = i|x) = \sum_{j=1}^{c} L(f(x) = i, y = j)\, P(Y = j|x).$$

(a) Show that the following policy obtains the minimum risk. (1) Choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and $P(Y = i|x) \geq 1 - \lambda_r/\lambda_s$; (2) choose doubt otherwise.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Explain why this is consistent with what one would expect intuitively.

Q7. Gaussian Classification

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with classes ω_1 and ω_2, $P(\omega_1) = P(\omega_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

(b) The Bayes error is the probability of misclassification,

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2)\, P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1)\, P(\omega_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.

Q8. Maximum Likelihood Estimation

Let X be a discrete random variable which takes values in $\{1, 2, 3\}$ with probabilities $P(X = 1) = p_1$, $P(X = 2) = p_2$, and $P(X = 3) = p_3$, where $p_1 + p_2 + p_3 = 1$. Show how to use the method of maximum likelihood to estimate p_1, p_2 and p_3 from n observations of X: $x_1, \ldots, x_n$. Express your answer in terms of the counts $k_1 = \sum_{i=1}^{n} \mathbb{1}(x_i = 1)$, $k_2 = \sum_{i=1}^{n} \mathbb{1}(x_i = 2)$, and $k_3 = \sum_{i=1}^{n} \mathbb{1}(x_i = 3)$, where $\mathbb{1}(x = a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a. \end{cases}$