CS 189 Introduction to Machine Learning
Fall 2019 Jennifer Listgarten & Stella Yu

HW01

Due: 9/10 at 10 p.m.

0 Getting Started

Read through this page carefully. You may typeset your homework in latex or submit neatly handwritten/scanned solutions. Please start each question on a new page. Deliverables:

1. Submit a PDF of your writeup, **with an appendix for your code**, to assignment on Gradescope, "HW[X] Write-Up". If there are graphs, include those graphs in the correct sections. Do not simply reference your appendix.
2. If there is code, submit all code needed to reproduce your results, "HW[X] Code".
3. If there is a test set, submit your test set evaluation results, "HW[X] Test Set".
4. In all cases, replace "[X]" with the number of the assignment you are submitting.

1 Properties of Gaussians

Let $N(\mu,\ \sigma^2)$ denote the normal (Gaussian) distribution with mean μ and variance σ^2.

(a) Recall that $\mathbb{E}[X]$ denotes the expected value of a random variable X. Prove that if $X \sim N(0,\ \sigma^2)$, then $\mathbb{E}[e^{\lambda X}]$ satisfies $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$, where $\lambda \in \mathbb{R}$ is a fixed constant. As a function of λ, $\mathbb{E}[e^{\lambda X}]$ is known as the *moment-generating function*.

(b) Recall that $\mathbb{P}(X \geq t)$ denotes the probability that the inequality $X \geq t$ holds. Prove that if $X \sim N(0,\ \sigma^2)$ then $\mathbb{P}(X \geq t) \leq \exp\left(-t^2/2\sigma^2\right)$, 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.

(c) Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$ be independent and identically distributed (iid). Recall that this means that they are mutually independent and all share the same distribution. Can you prove a similar concentration result for the average of the X_i: $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$? 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.

(d) Take two orthogonal vectors $u, v \in \mathbb{R}^d$, $u \perp v$, and let $X = (X_1, \ldots, X_d)$ be a vector of d iid standard Gaussians, $X_j \sim N(0,1), \forall\, j \in \{1, 2, \ldots d\}$. Further, let $\langle \cdot,\ \cdot \rangle$ denote the standard inner product in $\mathbb{R}^d$. If we define $u_x = \langle u, X\rangle$ 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.

2 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. You should assume unless otherwise stated that X is a different random variable in each part.

(a) Let X be a random variable with pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and zero everywhere else (note this an exponential distribution). Use induction on k to show that for $k \in \mathbb{Z}_{\geq 0}$, $\mathbb{E}[X^k] = \frac{k!}{\lambda^k}$. Do not use the moment generating function of X without deriving it first (you should not need to use it however).
Hint: use integration by parts.

(b) 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 PDF $f(x)$ and a CDF $F(x)$; this might simplify notation.

(c) 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, $|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle$, and the fact that a valid inner product on the set of random variables U and V is given by $\mathbb{E}[UV]$ (no proof necessary).

3 Gradients and Norms

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

(b) (i) Let $\alpha = \sum_{j=1}^d y_j \ln \beta_j$ for $y, \beta \in \mathbb{R}^d$. What are the partial derivatives $\frac{\partial \alpha}{\partial \beta_j}$?

(ii) Let $\beta = \sinh(\gamma)$ for $\gamma \in \mathbb{R}^d$, where we treat the sinh as an element-wise operation: i.e. $\beta_j = \sinh\left(\gamma_j\right)$. What are the partial derivatives $\frac{\partial \beta_j}{\partial \gamma_k}$ for all $1 \leq j,\ k \leq d$?

(iii) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^d$, $\rho \in \mathbb{R}^m$, $A \in \mathbb{R}^{d \times m}$. What are the the partial derivatives $\frac{\partial \gamma_j}{\partial \rho_\ell}$ for $1 \leq j \leq d$ and $1 \leq \ell \leq m$?

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

(d) 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 the samples' outcome values. The relationship between the features

and outcomes is linear; ideally, there exists a set of parameters $w \in \mathbb{R}^d$ such that $Xw = y$. However, n is large and there is noise in the acquisition of X and y, so we wish to find the *best linear approximation* for this data. Assuming X has full column rank, compute $w^* = \arg\min_w \|y - Xw\|_2^2$ in terms of X and y.

4 Linear Algebra Review

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

(i) For all $x \in \mathbb{R}^d$, $x^\top A x \geq 0$.

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

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

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

(b) Now that we're equipped with different definitions of positive semi-definiteness, prove the following properties of PSD matrices:

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

(ii) If A is PSD, all diagonal entries of A are non-negative, $A_{jj} \geq 0, \forall\, j \in \{1, 2, \ldots d\}$.

(iii) If A is PSD, the sum of all entries of A is non-negative, $\sum_{k=1}^{d} \sum_{j=1}^{d} A_{jk} \geq 0$.

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

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

(c) Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix. Prove that the largest eigenvalue of A is

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

5 Covariance Practice

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 C such that $C_{ij} = \mathrm{Cov}(Z_i, Z_j)$. Concisely, $C = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$ where $\mu = \mathbb{E}[Z]$. Prove that the covariance matrix is always positive semi-definite.
Hint: Use linearity of expectation.

6 A Simple Classification Approach

Make sure to submit the code you write in this problem to "HW1 Code" on Gradescope.

Classification is an important problem in applied machine learning and is used in many different applications like image classification, object detection, speech recognition, machine translation and others.

In *classification*, we assign each datapoint a class from a finite set (for example the image of a digit could be assigned the value $0, 1, \ldots, 9$ of that digit). This is different from *regression*, where each datapoint is assigned a value from a continuous domain like $\mathbb{R}$ (for example features of a house like location, number of bedrooms, age of the house, etc. could be assigned the price of the house).

In this problem we consider the simplified setting of classification where we want to classify data points from $\mathbb{R}^d$ into *two* classes. For a linear classifier, the space $\mathbb{R}^d$ is split into two parts by a hyperplane: All points on one side of the hyperplane are classified as one class and all points on the other side of the hyperplane are classified as the other class.

The goal of this problem is to show that even a regression technique like linear regression can be used to solve a classification problem. This can be achieved by regressing the data points in the training set against -1 or 1 depending on their class and then using the level set of 0 of the regression function as the classification hyperplane (i.e. we use 0 as a threshold on the output to decide between the classes).

Later in lecture we will learn why linear regression is not the optimal approach for classification and we will study better approaches like logistic regression, SVMs and neural networks.

(a) The dataset used in this exercise is a subset of the MNIST dataset. The MNIST dataset assigns each image of a handwritten digit their value from 0 to 9 as a class. For this problem we only keep digits that are assigned a 0 or 1, so we simplify the problem to a two-class classification problem. You should have access to the following files: `starter.py`, `test_features.npy`, `test_labels.npy`, `train_features.npy`, `train_labels.npy`.

Download and visualize the dataset (example code included). Include three images that are labeled as 0 and three images that are labeled as 1 in your submission.

(b) We now want to use linear regression for the problem, treating class labels as real values $y = -1$ for class "zero" and $y = 1$ for class "one". In the dataset we provide, the images have already been flattened into one dimensional vectors (by concatenating all pixel values of the two dimensional image into a vector) and stacked as rows into a feature matrix $\mathbf{X}$. We want to set up the regression problem $\min_{\mathbf{w}} \left\|\mathbf{X}\mathbf{w} - \mathbf{y}\right\|_2^2$ where the entry y_i is the value of the class (-1 or 1) corresponding to the image in row $\mathbf{x}_i^\top$ of the feature matrix. **Solve this regression problem for the training set and report the value of $\left\|\mathbf{X}\mathbf{w} - \mathbf{y}\right\|_2^2$ as well as the weights $\mathbf{w}$ (include at least the first 20 entries of $\mathbf{w}$; given its length you do not need to include the entire vector).** For this problem you may only use pure Python and numpy (no machine learning libraries!).

(c) Given a new flattened image $\mathbf{x}$, one natural rule to classify it is the following one: It is a zero if $\mathbf{x}^\top\mathbf{w} \leq 0$ and a one if $\mathbf{x}^\top\mathbf{w} > 0$. **Report what percentage of the digits in the training set are correctly classified by this rule. Report what percentage of the digits in the test set are correctly classified by this rule.**

(d) **Why is the performance typically evaluated on a separate test set (instead of the training set) and why is the performance on the training and test set similar in our case?** We will cover these questions in more detail later in the class.

(e) Somebody suggests to use 0 (for class 0) and 1 (for class 1) as the entries for the target vector $\mathbf{y}$. Try out how well this is doing (make sure to adapt the classification rule, i.e. the threshold set for the outputs). **Report what percentage of digits are correctly classified using this approach on the training set and test set.** How are the performances of the two approaches if you add a bias column to the feature matrix $\mathbf{X}$? A bias column is a column of all ones, i.e. the new feature matrix $\mathbf{X}'$ is

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\top & 1 \end{bmatrix}$$

Report what percentage of digits are correctly classified using regression targets $0/1$ **and** $-1/1$ **with bias on the training set and test set. Try to explain the results!**