Homework #0

Spring 2021, CSE 446/546: Machine Learning
Prof. Simon Du and Prof. Sewoong Oh
Due: 04/05/21 Monday 11:59 PM Pacific Time
A: 38 points. B: 4 points

Please review all homework guidance posted on the website before submitting to Gradescope. Reminders:

- Please provide succinct answers along with succinct reasoning for all your answers. Points may be deducted if long answers demonstrate a lack of clarity. Similarly, when discussing the experimental results, concisely create tables and/or figures when appropriate to organize the experimental results. In other words, all your explanations, tables, and figures for any particular part of a question must be grouped together.
- For every problem involving generating plots, please include the plots as part of your PDF submission.
- When submitting to Gradescope, please link each question from the homework in Gradescope to the location of its answer in your homework PDF. Failure to do so may result in point deductions. For instructions, see https://www.gradescope.com/get_started#student-submission.
- Please recall that B problems, indicated in boxed text are only graded for 546 students, and that they will be weighted at most 0.2 of your final GPA (see website for details). In Gradescope there is a place to submit solutions to A and B problems separately. You are welcome to create just a single PDF that contains answers to both, submit the same PDF twice, but associate the answers with the individual questions in Gradescope.
- If you collaborate on this homework with others, you must indicate who you worked with on your homework. Failure to do so may result in accusations of plagiarism and point deductions. Please review website for collaboration policy.
- For every problem which requires coding, please provide the code files on Gradescope in a separate assignment created for code. Submitting all code files rewards *[1 point]* .

Probability and Statistics

A.1 *[2 points]* (Bayes Rule, from Murphy exercise 2.4.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

A.2 For any two random variables X, Y the *covariance* is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. You may assume X and Y take on a discrete values if you find that is easier to work with.

a. *[1 point]* If $\mathbb{E}[Y \mid X = x] = x$ show that $\text{Cov}(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big]$.

b. *[1 point]* If X, Y are independent show that $\text{Cov}(X, Y) = 0$.

A.3 Let X and Y be independent random variables with PDFs given by f and g, respectively. Let h be the PDF of the random variable $Z = X + Y$.

a. *[2 points]* Show that $h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x$. (If you are more comfortable with discrete probabilities, you can instead derive an analogous expression for the discrete case, and then you should give a one sentence explanation as to why your expression is analogous to the continuous case.).

b. *[1 point]* If X and Y are both independent and uniformly distributed on $[0,1]$ (i.e. $f(x) = g(x) = 1$ for $x \in [0,1]$ and 0 otherwise) what is h, the PDF of $Z = X + Y$?

A.4 *[1 point]* A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is Gaussian distributed with mean μ and variance σ^2. Given that for any $a, b \in \mathbb{R}$, we have that $Y = aX + b$ is also Gaussian, find a, b such that $Y \sim \mathcal{N}(0,1)$.

A.5 *[2 points]* For a random variable Z, its mean and variance are defined as $\mathbb{E}[Z]$ and $\mathbb{E}\big[(Z - \mathbb{E}[Z])^2\big]$, respectively. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and variance σ^2. If we define $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, what is the mean and variance of $\sqrt{n}(\widehat{\mu}_n - \mu)$?

A.6 If $f(x)$ is a PDF, the cumulative distribution function (CDF) is defined as $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$. For any function $g\colon \mathbb{R} \to \mathbb{R}$ and random variable X with PDF $f(x)$, recall that the expected value of $g(X)$ is defined as $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(y)f(y)\,\mathrm{d}y$. For a boolean event A, define $\mathbf{1}\{A\}$ as 1 if A is true, and 0 otherwise. Thus, $\mathbf{1}\{x \le a\}$ is 1 whenever $x \le a$ and 0 whenever $x > a$. Note that $F(x) = \mathbb{E}[\mathbf{1}\{X \le x\}]$. Let $X_1, \ldots, X_n$ be *independent and identically distributed* random variables with CDF $F(x)$. Define $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le x\}$. Note, for every x, that $\widehat{F}_n(x)$ is an *empirical estimate* of $F(x)$. You may use your answers to the previous problem.

a. *[1 point]* For any x, what is $\mathbb{E}\Big[\widehat{F}_n(x)\Big]$?

b. *[1 point]* For any x, the variance of $\widehat{F}_n(x)$ is $\mathbb{E}\left[\Big(\widehat{F}_n(x) - F(x)\Big)^2\right]$. Show that $\mathrm{Var}\Big[\widehat{F}_n(x)\Big] = \frac{F(x)(1-F(x))}{n}$.

c. *[1 point]* Using your answer to b, show that for all $x \in \mathbb{R}$, we have $\mathbb{E}\left[\Big(\widehat{F}_n(x) - F(x)\Big)^2\right] \le \frac{1}{4n}$.

B.1 *[1 point]* Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables drawn uniformly at random from $[0,1]$. If $Y = \max\{X_1, \ldots, X_n\}$ then find $\mathbb{E}[Y]$.

B.2 *[1 point]* Let X be random variable with $\mathbb{E}[X] = \mu$ and $\mathbb{E}\big[(X-\mu)^2\big] = \sigma^2$. For any $x > 0$, use Markov's inequality to show that $\mathbb{P}\{X \ge \mu + \sigma x\} \le 1/x^2$.

Linear Algebra and Vector Calculus

A.7 (Rank) Let $A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$. For each matrix A and B,

a. *[2 points]* What is its rank?

b. *[2 points]* What is a (minimal size) basis for its column span?

A.8 (Linear equations) Let $A = \begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix}$, $b = \begin{bmatrix} -2 & -2 & -4 \end{bmatrix}^\top$, and $c = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$.

a. *[1 point]* What is Ac?

b. *[2 points]* What is the solution to the linear system $Ax = b$? (Show your work).

A.9 (Hyperplanes) Assume w is an n-dimensional vector and b is a scalar. A hyperplane in $\mathbb{R}^n$ is the set $\{x \in \mathbb{R}^n \mid w^\top x + b = 0\}$.

a. *[1 point]* ($n = 2$ example) Draw the hyperplane for $w = \begin{bmatrix}-1 & 2\end{bmatrix}^\top$, $b = 2$? Label your axes.

b. *[1 point]* ($n = 3$ example) Draw the hyperplane for $w = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^\top$, $b = 0$? Label your axes.

c. *[2 points]* Given some $x_0 \in \mathbb{R}^n$, find the *squared distance* to the hyperplane defined by $w^\top x + b = 0$. In other words, solve the following optimization problem:

$$\begin{aligned} &\min_x \|x_0 - x\|_2^2 \\ &\text{subject to } : w^\top x + b = 0\,. \end{aligned}$$

Here $\|\cdot\|_2 \colon \mathbb{R}^n \to [0, \infty)$ is the standard ℓ_2-norm on $\mathbb{R}^n$, defined as $\|v\|_2 \coloneqq \left(\sum_{i=1}^n v_i^2\right)^{1/2}$ for $v \in \mathbb{R}^n$. (Hint: If $\widetilde{x}_0$ is the minimizer of the above problem, show that $\|x_0 - \widetilde{x}_0\|_2 = \left|\frac{w^\top(x_0 - \widetilde{x}_0)}{\|w\|_2}\right|$. What is $w^\top \widetilde{x}_0$?)

A.10 For possibly non-symmetric $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and $c \in \mathbb{R}$, let $f(x, y) = x^\top \mathbf{A} x + y^\top \mathbf{B} x + c$. Define

$$\nabla_z f(x, y) = \left[\frac{\partial f}{\partial z_1}(x, y) \quad \frac{\partial f}{\partial z_2}(x, y) \quad \ldots \quad \frac{\partial f}{\partial z_n}(x, y)\right]^\top.$$

a. *[2 points]* Explicitly write out the function $f(x, y)$ in terms of the components $A_{i,j}$ and $B_{i,j}$ using appropriate summations over the indices.

b. *[2 points]* What is $\nabla_x f(x, y)$ in terms of the summations over indices *and* vector notation?

c. *[2 points]* What is $\nabla_y f(x, y)$ in terms of the summations over indices *and* vector notation?

B.3 *[1 point]* The *trace* of a square matrix $X \in \mathbb{R}^{n \times n}$ is the sum of the diagonal entries; $\mathrm{Tr}[X] = \sum_{i=1}^n X_{i,i}$. If $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$, show that $\mathrm{Tr}[AB] = \mathrm{Tr}[BA]$.

B.4 *[1 point]* Let $v_1, \ldots, v_n$ be a set of non-zero vectors in $\mathbb{R}^d$. Let $V = \begin{bmatrix}v_1 & v_2 & \ldots & v_n\end{bmatrix}$ be the vectors concatenated.

a. What is the minimum and maximum rank of $\sum_{i=1}^n v_i v_i^\top$?

b. What is the minimum and maximum rank of V?

c. Let $A \in \mathbb{R}^{D \times d}$ for $D > d$. What is the minimum and maximum rank of $\sum_{i=1}^n (Av_i)(Av_i)^\top$?

d. What is the minimum and maximum rank of AV? What if V is rank d?

Programming

A.11 For the A, b, c as defined in Problem 8, use NumPy to compute (take a screen shot of your answer):

a. *[2 points]* What is A^{-1}?

b. *[1 point]* What is $A^{-1}b$? What is Ac?

A.12 *[4 points]* Two random variables X and Y have equal distributions if their CDFs, F_X and F_Y, respectively, are equal, i.e. for all x, $|F_X(x) - F_Y(x)| = 0$. The central limit theorem says that the sum of k independent, zero-mean, variance $1/k$ random variables converges to a (standard) Normal distribution as k tends to infinity. We will study this phenomenon empirically (you will use the Python packages Numpy and Matplotlib). Define $Y^{(k)} = \frac{1}{\sqrt{k}} \sum_{i=1}^k B_i$ where each B_i is equal to -1 and 1 with equal probability. From your solution to problem A.5, we know that $\frac{1}{\sqrt{k}} B_i$ is zero-mean and has variance $1/k$.

a. For $i = 1, \ldots, n$ let $Z_i \sim \mathcal{N}(0,1)$. If $F(x)$ is the true CDF from which each Z_i is drawn (i.e., Gaussian) and $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_i \leq x\}$, use the answer to problem A.6 above to choose n large enough such that, for all $x \in \mathbb{R}$,

$$\sqrt{\mathbb{E}\left[\left(\widehat{F}_n(x) - F(x)\right)^2\right]} \leq 0.0025 \ ,$$

and plot $\widehat{F}_n(x)$ from -3 to 3.
(Hint: Use `Z=numpy.random.randn(n)` to generate the random variables, and
`import matplotlib.pyplot as plt`;
`plt.step(sorted(Z), np.arange(1,n+1)/float(n))` to plot).

b. For each $k \in \{1, 8, 64, 512\}$ generate n independent copies $Y^{(k)}$ and plot their empirical CDF on the same plot as part a.
(Hint: `np.sum(np.sign(np.random.randn(n,k))*np.sqrt(1./k), axis=1)` generates n of the $Y^{(k)}$ random variables.)

Be sure to always label your axes. Your plot should look something like the following (Tip: checkout `seaborn` for instantly better looking plots.)