CS 4780/5780: Introduction to Machine Learning **(Due: 09/08/2020)**

CS4780/5780 Mathematical Background Exam

Instructors: Nika Haghtalab, Thorsten Joachims

Note: The purpose of this prerequisite exam is to ensure your preparedness for succeeding in this class. The exam must be completed individually. You may not seek help from staff or fellow students. You are allowed to use the provided references (click-able points in underline and red color) or any textbooks, but it is an academic integrity violation if you search for answers.

You may either print out the exam and write your answers into the space provided, annotate the exam pdf digitally, or type the answers into the provided Latex file. If you choose the Latex option, make sure you **do not resize the solution boxes.**

Submit your solution via Gradescope, which is also linked on Canvas.

Solutions are due on Tuesday Sept. 8 at 5pm EST. Everybody has a 2-day extension, so the due date is effectively Thursday Sept. 10 at 5pm EST. This extension does not count towards your allocated homework extensions.

Question 1. Calculus and Linear Algebra

The following questions test your basic skills in computing the derivatives, as well as applying the concept of convexity to determine the properties of the functions.

1. (4pts) Show that $f(x) = (2x-1)^2$ and $g(x) = \log_2(x)$ are convex and concave respectively. Show your reasoning through the use of derivatives.

$f(x) = (2x-1)^2 = 4x^2 - 4x + 1$ $f'(x) = 8x - 4$ $f''(x) = 8$ $f''(x)$ is 8 which is greater than or equal to 0 for all x, thus $f(x)$ is convex.	$g(x) = \log_2(x)$ domain of $g(x)$ is $(0,\infty)$ $g'(x) = \frac{1}{\ln(2)x}$ $g''(x) = -\frac{1}{\ln(2)x^2}$ for the domain of $g(x)$, $g''(x)$ is always less than or equal to zero, thus, $g(x)$ is concave.

2. (4pts) Show that $h(x) = x^3$ is neither convex or concave. You may plot the function to get some intuition, but you will only receive credit for a mathematical proof.

$h(x) = x^3$ $h''(x) = 6x$
for $h(x)$ to be concave, $h''(x)$ must be ≤ 0. By example, when $x = 1$, $h''(x) = 6$ which does not fulfill this inequality, thus $h(x)$ is not concave. Similarly, for $h(x)$ to be convex, $h''(x)$ must be ≥ 0 for all x. However, when $x = -1$, $h''(x) = -6$ which is not ≥ 0, thus $h(x)$ is not convex. Thus, $h(x)$ is neither concave nor convex.

3. (4pts) Let $A \in \mathbb{R}^{D \times D}$ be a positive definite matrix and $B \in \mathbb{R}^{D \times D}$ a positive semi-definite matrix. Show that $A + B$ is positive definite.

By definition of a positive definite matrix $x^T A x > 0$ $\forall x \in \mathbb{R}^D$ and of a positive semidefinite matrix $x^T B x \geq 0$

Show $x^T (A+B) x > 0$ $\forall x \in \mathbb{R}^D$

$x^T A x + x^T B x > 0$, by their definitions, $x^T A x$ is always positive, and $x^T B x$ is at a minimum 0. Thus the sum of a positive and zero is always positive and the inequality holds. Thus, $A + B$ is positive definite.

4. (6pts) Let $\vec{\alpha}, \vec{w} \in \mathbb{R}^D$ (D dimensional Euclidean space), and $\Sigma \in \mathbb{R}^{D \times D}$. Let

$$r(\vec{\alpha}) = \vec{\alpha}^\top \vec{w} + \vec{\alpha}^\top \Sigma \vec{\alpha}.$$

Calculate the first and second derivatives. Under which conditions is the function $r(\vec{\alpha})$ convex?

$r(\vec{\alpha}) = \vec{\alpha}^T \vec{w} + \vec{\alpha}^T \Sigma \vec{\alpha}$

$r'(\vec{\alpha}) = \vec{w} + 2 \Sigma \vec{\alpha}$

$r''(\vec{\alpha}) = 2 \Sigma$

function $r(\vec{\alpha})$ is convex if $\forall x, y \in \mathbb{R}^D$ and all $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$

as r is continuous, we can check using $\lambda = \frac{1}{2}$

$f(\frac{1}{2}x + \frac{1}{2}y) \leq \frac{1}{2} f(x) + \frac{1}{2} f(y)$

$(\frac{1}{2}x + \frac{1}{2}y)^T \vec{w} + (\frac{1}{2}x + \frac{1}{2}y)^T \Sigma (\frac{1}{2}x + \frac{1}{2}y) \leq \frac{1}{2}(x^T w + x^T \Sigma x + y^T w + y^T \Sigma y)$

so long as this inequality holds $r(\vec{\alpha})$ is convex

Question 2. Optimization

1. (5pts) Show that the function $f(x) = (2x-1)^2$ has only one local optimum. Is it a minimum or maximum? You can plot the function to get some intuition but you will only receive credit for a mathematical proof **based on derivatives.**

$f'(x) = 8x - 4 \quad f''(x) = 8$

an inflection point happens when $f'(x) = 8x - 4 = 0$, thus at only $x = \frac{1}{2}$. As $f''(x) = 8 \geq 0$, $f(x)$ is convex and $(\frac{1}{2}, 0)$ is the only local minimum. The inflection point of a convex function is a minimum.

2. (6pts) The following question tests your familiarity with the method of Lagrange multipliers. Consider the function

$$h(x, y) = (x-2)^2 + (y-2)^2.$$

What is the minimum if we constrain $2x^2 + 2y^2 = 1$? Please show the derivation steps. First write the Lagrangian, then write down the set of constraints you get from setting the derivative of the Lagrangian to zero. Then, write the final solution.

Lagrangian: $L(x,y,\lambda) = (x-2)^2 + (y-2)^2 - \lambda(2x^2 + 2y^2 - 1)$
$= x^2 - 4x + 4 + y^2 - 4y + 4 - 2x^2\lambda - 2y^2\lambda + \lambda$

$\frac{dL}{dx} = 2x - 4 - 4x\lambda = 0 \qquad \frac{dL}{dy} = 2y - 4 - 4y\lambda = 0 \qquad \frac{dL}{d\lambda} = 2x^2 + 2y^2 - 1 = 0$

$x - 2x\lambda = 2 \qquad y - 2y\lambda = 2$

$x = \frac{2}{1-2\lambda} \qquad y = \frac{2}{1-2\lambda}$

$2\left(\frac{2}{1-2\lambda}\right)^2 + 2\left(\frac{2}{1-2\lambda}\right)^2 = 1$

$\left(\frac{2}{1-2\lambda}\right)^2 = \frac{1}{4}$

$\frac{2}{1-2\lambda} = \pm\frac{1}{2}$

$\frac{2}{1-2\lambda} = \frac{1}{2} \qquad \lambda = \frac{-3}{2}$

$\frac{2}{1-2\lambda} = -\frac{1}{2} \qquad \lambda = \frac{5}{2}$

$\boxed{x = y = \frac{1}{2} \quad \text{min}}$

$x = y = -\frac{1}{2}$ ~~max~~

$h(\frac{1}{2}, \frac{1}{2}) = (\frac{1}{2} - 2)^2 + (\frac{1}{2} - 2)^2 = 2(-\frac{3}{2})^2 = \boxed{\frac{9}{2} = \text{min of } h(x,y) \text{ with contraint } 2x^2 + 2y^2 = 1}$

Question 3. Probability and Statistics

1. (6pts) Consider a random variable $X = (Z+1)^2$ where $Z \sim Ber(p)$ for some $0 \le p \le 1$. In other words, Z follows the Bernoulli distribution with parameter p.
   Calculate the mean E[X] and variance Var(X) of X. Please show your work.

$$P(Z=1)=p \qquad P(Z=0)=1-p$$
$$E(x) = p(1+1)^2 + (1-p)(0+1)^2$$
$$E(x) = 4p + 1 - p = \boxed{3p-1}$$
$$Var(x) = E(x^2) - (E(x))^2$$
$$= E((\cancel{z+1})^{\cancel{4}}) - (3p-1)^2$$
$$= p(1+1)^4 - 9p^2 + 6p - 1$$
$$= 16p - 9p^2 + 6p - 1$$
$$\boxed{Var(x) = 22p - 9p^2 - 1}$$

2. (6pts) Suppose we have observed N independent samples from X, $(x_1, \ldots, x_N)$. X still follows the same distribution from the previous question.
   What is the maximum likelihood estimate of p? Please show the likelihood objective, the log-likelihood objective, and derive the MLE $\hat{p}$. Show your work. A useful function is

$$1\{x_i = a\} = \begin{cases} 1, & \text{if } x_i = a \\ 0, & \text{otherwise} \end{cases}$$

$$L(\theta) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$LL(\theta) = \sum_{i=1}^{n} \log p^{x_i}(1-p)^{1-x_i} = \sum_{i=1}^{n} x_i \log(p) + (1-x_i)\log(1-p)$$
$$\frac{d\,LL(\hat{p})}{d\hat{p}} = \sum_{i=1}^{N} x_i \frac{1}{p} + \left(1 - \sum_{i=1}^{N} x_i\right)\left(\frac{-1}{1-p}\right)$$
$$\hat{p} = \frac{\sum_{i=1}^{N} x_i}{N} = Avg(l(x_i))$$

3. (3pts) Is the maximum likelihood estimator (MLE) unbiased? Provide the derivation.

MLE is unbiased if $E(..(x_1, x_2, \ldots, x_n)) = p$

$E(\hat{p}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} p = \frac{1}{n}(np) = p$

thus, $E(\hat{p}) = p$ thus MLE is an unbiased estimate of p.

4. (5pts) Let $\vec{U} = (U_1, \ldots, U_D) \sim N(\vec{\mu}, \Sigma)$ be drawn from a multivariate Gaussian distribution with mean vector $\vec{\mu} \in \mathbb{R}^D$, where $\vec{\mu} = (\mu_1, \ldots, \mu_D)$ and $\mu_i \neq \mu_j$ for all $i, j \in \{1, \ldots, D\}$ *s.t.* $i \neq j$. Assume that $(U_1, \ldots, U_D)$ are independent.
We will perform various operations on the elements of $\vec{U}$. For each of the below transformations, identify which are still normally distributed. For those that are, compute their means and covariances.

$$S = \sum_{i=1}^{D} U_i$$

$$P = \Pi_{i=1}^{D} U_i$$

$$T = \begin{cases} U_1, & \text{with probability } 1/D \\ \ldots & \\ U_D, & \text{with probability } 1/D \end{cases}$$

S is normally distributed

$E\left(\sum_{i=1}^{D} u_i\right) = \sum_{i=1}^{D} E(u_i) = \sum_{i=1}^{D} \mu_i$

$Var(S) = E\left(\left(\sum_{i=1}^{D} u_i\right)^2\right) - E\left(\sum_{i=1}^{D} u_i\right)^2$

As for P, P is not normally distributed due to skewedness,

As for T, T is not normally distributed, for example assume $D=2$, $\mu_1=-5$, $\mu_2=5$, $\Sigma_1=\Sigma_2=1$, the resulting distribution will be bimodal around -5 and 5 and thus clearly not normal.

5. (6pts) You have two friends that are identical twins. If you ask them for help in answering a crossword puzzle question, one will give you the correct answer with 50% probability. The other will give you the correct answer with 75% probability.

If you randomly go up to one of the twins and ask them a crossword puzzle question, what is the probability you will be told the correct answer? If the answer you are told is correct, what is the probability that it was the 75% twin that told you it?

Please use the following variables in your solution:

Let C be the event of hearing a correct answer, A be the event of asking the friend with 50% correct answers, and B be the event of asking the friend with 75% correct answers.

(Hint: Use the rules of probability and Bayes's rule).

$P(C|A) = 0.5 \qquad P(C|B) = 0.75 \qquad P(A) = 0.5$ identical

$P(NC|A) = 0.5 \qquad P(NC|B) = 0.25 \qquad P(B) = 0.5$

$$P(C) = P(C|A)P(A) + P(C|B)P(B)$$
$$= (0.5)(0.5) + (0.75)(0.5) = \boxed{0.625 = P(C)}$$

$$P(B|C) = \frac{P(C|B)P(B)}{P(C)} = \frac{(0.75)(0.5)}{0.625} = \boxed{0.6 = P(B|C)}$$

6. (6 pts) Consider a biased coin that lands heads side up with 80% probability. Let $N \geq 2$. Toss this coin N times and denote the number of heads by H. In this question, we attempt to understand how large H typically gets.

   (a) (2 pts) Write the mathematical formula for the exact probability that there is at most one tail out of N coin flips.

   (b) (4 pts) Show that $\Pr[|\frac{H}{N} - 0.8| \geq N^{-\frac{1}{4}}] \leq 0.12$. Show all steps and justify all substitutions you make. Explain how terms from this problem map into Hoeffding's inequality, and describe how the problem fulfills the conditions of Hoeffding's inequality.

   You may use the following fact to assist you with this question. Hoeffding's Inequality: Consider a random variable X such that $X \in [0, 1]$. Let $X_1, \ldots, X_N$ be independent copies of this random variable. Then,

$$\Pr\left[\left|\frac{\sum_{i=1}^{N} X_i}{N} - \mathbb{E}[X]\right| \geq t\right] \leq 2\exp(-2Nt^2).$$

(a) $P(N-H \geq N-1) = P(H=N) + P(H=N-1)$

$P(\text{At most 1 Tail}) = (0.8)^N + (0.8)^{N-1}(0.2)$

$P(\text{At most 1 Tail}) = (0.8)^{N-1}(0.8 + 0.2)$

$P(\text{At most 1 Tail}) = (0.8)^{N-1}$

(b) Show connection between $P[|\frac{H}{N} - 0.8| \geq N^{-\frac{1}{4}}] \leq 0.12$ and Hoeffding's,

$H = \sum_{i=1}^{N} X_i$ where $X_i = 1$ if heads, 0 if tails

$E(X) = 0.8$, the probab of flipping heads

let $t = N^{-\frac{1}{4}}$ then show $0.12 \geq 2e^{-2Nt^2}$

$0.12 \geq 2e^{-2N(N^{-\frac{1}{4}})^2}$

$0.12 \geq 2e^{-2\sqrt{N}}$

$0.06 \geq e^{-2\sqrt{N}}$

$1.98 = \left(\frac{-\ln(0.06)}{2}\right)^2 = N$

as $N \to \infty$ $e^{-2\sqrt{N}} \to 0$

thus $\forall N \geq 1.98$, the equality holds.

As was defined above, $N \geq 2$, thus Hoeffding's inequality holds and clearly $P(|\frac{H}{N} - 0.8| \geq N^{-\frac{1}{4}})$ less than or equal to 0.12