CS534 — Written Homework 0 (40pts) — Due Oct 1st 11:59pm, 2021

This first written assignment focuses on some of the basic math concepts including gradient, probability theory, expectation, and maximum likelihood estimation.

1. (Gradient) Compute the gradient $\nabla_{\mathbf{x}} f$ of the following functions.

   a. (1pt)
   $$f(z) = \log(1+z),\ z = \mathbf{x}^T\mathbf{x},\ \mathbf{x} \in R^D$$

   Use the property for derivative of a log first, then apply the chain rule and vector derivative properties.

   $\nabla_{\mathbf{x}} f = \frac{1}{1+z} * \nabla_{\mathbf{x}} z$, *where* $z = \mathbf{x^T x}$.

   Now, we have $\nabla_{\mathbf{x}} z = \nabla_{\mathbf{x}} \mathbf{x^T x} = 2\mathbf{x}$ *(see the Matrix derivative cheatsheet or Matrix cookbook). Finally, sub in* $\mathbf{x^T x}$ *for z:*
   $\nabla_{\mathbf{x}} f = \frac{2\mathbf{x}}{1+\mathbf{x^T x}}$

   b. (2pts)
   $$f(z) = \exp^{-\frac{1}{2}z}$$
   $$z = g(\mathbf{y}) = \mathbf{y}^T S^{-1} \mathbf{y}$$
   $$\mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \mu$$
   $$\text{where } \mathbf{x}, \mu \in R^D, S \in R^{D\times D}$$

   Recall first the property (from Matrix cheatsheet/cookbook) that:
   $\nabla_{\mathbf{x}} \mathbf{x^T} M \mathbf{x} = 2M\mathbf{x}$, *for some matrix* M. *Here we have a case where* $M = S^{-1}$. *The rest comes directly from derivatives of functions and the chain rule:*

   $\nabla_{\mathbf{x}} f = -\frac{1}{2} * \exp^{\frac{1}{2}z} * \nabla_{\mathbf{x}} z$
   $\nabla_{\mathbf{x}} f = -\frac{1}{2} * \exp^{\frac{1}{2}z} * 2S^{-1}\mathbf{y} * \nabla_{\mathbf{x}} \mathbf{y}$ *(from matrix property)*
   $\nabla_{\mathbf{x}} f = -\frac{1}{2} * \exp^{\frac{1}{2}z} * 2S^{-1}\mathbf{y} * \nabla_{\mathbf{x}} (\mathbf{x} - \mu)$
   $\nabla_{\mathbf{x}} f = -\frac{1}{2} * \exp^{\frac{1}{2}z} * 2S^{-1}\mathbf{y} * 1$ *(now, substitute in the* $\mathbf{x} - \mu$*)*
   $\nabla_{\mathbf{x}} f = -\frac{1}{2} * \exp^{\frac{1}{2}(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu)} * 2S^{-1}(\mathbf{x} - \mu)$

2. (Probability) Consider two coins, one is fair and the other one has a 1/10 probability for head. Now you randomly pick one of the coins, and toss it twice. Answer the following questions.

   (a) (1pt) What is the probability that you picked the fair coin?

   (b) (1pt) What is the probability of the first toss being head?
   Let x_1 *denote the outcome of the first toss and let* y *denote the coin that is selected. We can write down the following probabilities.*
   $P(y = f) = P(y = uf) = \frac{1}{2}$
   $P(x_1 = h|y = f) = \frac{1}{2}$
   $p(x_1 = h|y = uf) = 1/10$

   Now we can write out the probability of the first toss being head as:
   $P(x_1 = h) = P(x_1 = h, y = f) + P(x_1 = h, y = uf)$
   $= P(y = f)P(x_1 = h|y = f) + P(y = uf)P(x_1 = h|y = uf) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{10}$
   $= \frac{1}{2} \times \frac{6}{10} = \frac{3}{10}$

   (c) (4pts) If both tosses are heads, what is the probability that you have chosen the fair coin (Hint: Bayes Rule)?
   Let x_1, x_2 *denote the outputs of the first two tosses. It is easy to see that*
   $P(x_1 = h, x_2 = h|y = f) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(x_1 = h, x_2 = h|y = uf) = \frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$

Now we need to compute $P(y = f|x_1 = h, x_2 = h)$, to do so, we use Bayes Theorem:
$P(y = f|x_1 = h, x_2 = h) = \frac{P(x_1=h,x_2=h|y=f)P(y=f)}{P(x_1=h,x_2=h)}$
To compute the denominator, we use the same approach as used in (a):
$P(x_1 = h, x_2 = h) = P(x_1 = h, x_2 = h|y = f)P(y = f) + P(x_1 = h, x_2 = h|y = uf)P(y = uf)$
$= \frac{1}{4} \times \frac{1}{2} + \frac{1}{100} \times \frac{1}{2} = \frac{13}{100}$

Plug this into the Bayes Theorem, we have:
$P(y = f|x_1 = h, x_2 = h) = \frac{P(x_1=h,x_2=h|y=f)P(y=f)}{P(x_1=h,x_2=h)} = \frac{1/4\times 1/2}{13/100} = \frac{25}{26}$

3. (Maximum likelihood estimation for uniform distribution.) Given a set of i.i.d. samples $x_1, x_2, ..., x_n \sim uniform(0, \theta)$.

   (a) (3pts) Write down the likelihood function of θ.

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta) \tag{1}$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} I(x_i \leq \theta) = \begin{cases} \frac{1}{\theta^n} & \textit{if } \forall x_i \leq \theta \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

   (b) (4pts) Find the maximum likelihood estimator for θ.

$$\operatorname*{argmax}_\theta L = \operatorname*{argmax}_\theta \frac{1}{\theta^n} \quad \textit{subject to} \quad \forall x_i \leq \theta \tag{3}$$

$$= \max\{x_1, ..., x_n\} = x_{max} \tag{4}$$

   From (3) to (4) is because in order to maximize L, we want θ to be as small as possible, but θ has to be no smaller than any observed values (otherwise L goes to zero). So the smallest value θ is allowed to take is x_{max}.

4. (12pts) (Maximum likelihood estimation of categorical distribution.) A DNA sequence is formed using four bases Adenine(A), Cytosine(C), Guanine(G), and Thymine(T). We are interested in estimating the probability of each base appearing in a DNA sequence. Here we consider each base as a random variable x following a categorical distribution of 4 values (a, c, g and t) and assume a sequence is generated by repeatedly sampling from this distribution. This distribution has 4 parameters, which we denote as p_a, p_c, p_g, and p_t. Given a collection of DNA sequences with accumulated length of N, we counted the number of times that we observe of the four values, denoted by n_a, n_c, n_g and n_t respectively. Please show that the maximum likelihood estimation for p_x is $\frac{n_x}{N}$, where $x \in \{a, c, g, t\}$. Note that the four parameters are constrained to sum up to 1. This can be captured as a constrained optimization problem, solved using the method of Lagrange multiplier.

   Helpful starting point: the probability mass function for the discrete random variable can be written compactly as

$$p(x) = \prod_{s=a,c,g,t} p_s^{I(x=s)}$$

   Here $I(x = s)$ is an indicator function, and takes value 1 if x is equal to s, and 0 otherwise.

   We can think of x as the outcome of rolling a 4-sided die. To simplify the notation, we will use index $1, 2, 3, 4$ to replace the four letters. The parameters are $\mathbf{p} = [p_1, p_2, p_3, p_4]$, subject to the constraint that $\sum_{i=1}^{4} p_i = 1$.

   Given a collection of DNA sequences, we could concatenate them into one long sequence of length N: $[x_1, x_2, ..., x_N]$.

Because we consider each position as independent, the log likelihood function can be written as:

$$\begin{aligned} l(\mathbf{p}) &= \sum_{m=1}^{n} \log \prod_{i=1}^{4} p_i^{I(x_m=i)} \\ &= \sum_{i=1}^{4} \log p_i \sum_{m=1}^{n} I(x_m = i) \\ &= \sum_{i=1}^{4} n_i \log p_i \end{aligned}$$

To take the constraints into consideration, we form the lagrangian:

$$l(\mathbf{p}, \lambda) = \sum_{i=1}^{4} n_i \log p_i - \lambda(\sum_{i=1}^{4} p_i - 1)$$

To find the optimizing parameters, we take the partial derivative of l wrt each parameter p_i, and setting it to zero:

$$\frac{\partial}{\partial p_i} l = \frac{n_i}{p_i} - \lambda = 0 \text{ for } i = 1, \cdots, 4$$

Using the constraints $\sum_{i=1}^{4} p_i = 1$, we arrive at the conclusion that $\lambda = N$. This leads to the final solution:

$$p_i = \frac{n_i}{N}, \text{ for } i = 1, \cdots, 4$$

5. (Expected loss). Sometimes the cost of classification is not symmetric, one type of mistake is much more costly than the other. For example, the cost of misclassifying a normal email as spam can be substantially higher than letting a spam slip through. This can be captured by using a mis-classification loss matrix like the following:

predicted label $\hat{y}$	true label y = 0	true label y = 1
0	0	10
1	5	0

where misclassifying a positive example ($y = 1$, $\hat{y} = 0$) has a cost of 10, and misclassifying a negative example ($y = 0$, $\hat{y} = 1$ has a smaller cost of 5.

Suppose we have a probabilistic model that estimates $P(y = 1|\mathbf{x})$ for given $\mathbf{x}$. Here we will go through some questions to figure out how to prediction for $\mathbf{x}$ so what the expected loss is minimized.

(a) (2pts) Say $P(y = 1|\mathbf{x}) = 0.4$, what is the expected loss of predicting $\hat{y} = 1$?
If $y = 0$, the loss will be 5, this has 60% chance because $P(y = 1|\mathbf{x}) = 0.4$. If $y = 1$, the loss will be 0. This has 40% chance. So the expected loss for $\hat{y} = 1$ is $0.6 * 5 + 0.4 * 0 = 3$.

(b) (3pts) What is the best prediction that minimizes the expected loss?
Let's consider the alternative $\hat{y} = 0$. The expected loss for that prediction will be: $0.6*0+0.4*10 = 4$. Therefore to minimize the expected loss, we will predict $\hat{y} = 1$.

(c) (4pts) Show that to minimize the expected loss for our decision for this loss matrix, we should set a probability threshold θ and predict $\hat{y} = 1$ if $P(y = 1|x) > \theta$ and $\hat{y} = 0$ otherwise.
We want to predict $\hat{y} = 1$ if the expected loss of predicting 1 is less than the expected loss of predicting zero. The loss for predicting a particular value $\hat{y}$ is:

$$L(\hat{y}, 0) * P(y = 0|x) + L(\hat{y}, 1) * P(y = 1|x)$$

where $L(\hat{y}, y)$ is the loss of predicting $\hat{y}$ when the true label is y. So the loss of predicting 1 is $0 * P(y = 1|x) + 5 * P(y = 0|x)$ whereas the loss of predicting 0 is $10 * P(y = 1|x) + 0 * P(y = 0|x)$. Thus we will predict 1 iff:

$$P(y = 0|\mathbf{x})5 \quad < \quad P(y = 1|\mathbf{x})10 \tag{5}$$

Define $p_1 = P(y = 1|\mathbf{x})$, then this becomes

$$(1 - p_1) \times 5 \quad < \quad p_1 \times 10 \tag{6}$$

$$p_1 \quad > \quad \frac{1}{3} \tag{7}$$

So we should set the threshold to 1/3.

(d) (3pts) Show a loss matrix where the threshold is 0.1.

There are many loss matrices that will result in a threshold of 0.1, so long as the cost of misclassifying a positive example is 9 times the cost of misclassifying a negative example (assuming zero cost for correct prediction). For example

predicted label $\hat{y}$	*true label y*	
	0	*1*
0	*0*	*9*
1	*1*	*0*