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

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

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?

   (c) (4pts) If both tosses are heads, what is the probability that you have chosen the fair coin (Hint: Bayes Rule)?

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.

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

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.

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 $\hat{y} = 1$?

(b) (3pts) What is the best prediction that minimizes the expected loss?

(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.

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