Homework 2 for Math 173B - Winter 2025

1. Consider the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, given by the probability distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

   (a) Show that $\mathbb{E}(x) = \mu$.

   (b) Show that $Var(x) = \sigma^2$.

The next two questions are designed to help us 1) understand one of the "cost" functions associated with logistic regression, and 2) experiment with logistic regression and SGD.

2. Let z be a random variable drawn from a distribution where the CDF is given by

$$G_Z(z) = \frac{1}{1 + e^{-z}}. \tag{1}$$

   Consider the random variable

$$y := sign(a + z)$$

   where $a \in \mathbb{R}$ is fixed and z is drawn according to the distribution above. (Here $sign(x) = 1$ when $x \geq 0$ and $sign(x) = -1$ when $x < 0$.)

   (a) Verify that $G_Z(z)$ satisfies the defining properties of a CDF.

   (b) Write $\mathbb{P}(y = 1)$ and $\mathbb{P}(y = -1)$ in terms of the CDF above.

   (c) Deduce that the probability mass function associated with y is $p(y) = \frac{1}{1+e^{-ay}}$, where $y = \pm 1$.

   (d) Suppose the random variables $y_1, y_2, ..., y_N$ satisfy $y_i = sign(a_i + z_i)$, where the a_i's are fixed and the z_i's are independently drawn from the distribution associated with (1). Deduce that

$$p(y_1, y_2, ..., y_N) = \prod_{i=1}^{N} \frac{1}{1 + e^{-a_i y_i}}. \tag{2}$$

Recall that one way to think of certain classification problems is the following:

- **Consider that you have some data $x_i \in \mathbb{R}^d$, $i =, 1..., N$. Each x_i is associated with class label y_i where y_i is either 1 or -1.**
- **You, as a data scientist observe labeled data points (x_i, y_i), $i = 1, ..., N$.**
- **Loosely speaking, logistic regression assumes that the labels y_i are assigned randomly, according to the model**

$$y_i = sign(w^T x_i + z_i),$$

  with z_i being random, as in (1), and $w \in \mathbb{R}^d$ being a parameter vector to be found.

- **Given the "training" data $(x_i, y_i), i = 1, ..., N$, you'd like to be able to classify new data points x, i.e., you'd like to assign a class label y to them.**
- **Your strategy is, given x, to assign $y = sign(w^T x)$.**
- **For this to potentially work, you need to first estimate the value of $w \in \mathbb{R}^d$. You will now derive an optimization function to find the best value of w given our data (x_i, y_i), as well as an SGD algorithm for solving it.**

3. Recall problem 2(d) but now with $a_i = w^T x_i$. In order to execute the strategy above, we will try to find the vector w that maximizes the probability of the observed class labels. That is, we seek w^* that maximizes
$$H(w) = \prod_{i=1}^{N} \frac{1}{1 + e^{-w^T x_i y_i}}.$$
   (a) Explain, in your own words, why maximizing $H(w)$ makes sense.
   (b) Explain why this is equivalent to finding w^* that maximizes
$$\widetilde{F}(w) = \sum_{i=1}^{N} \log\left(\frac{1}{1 + e^{-w^T x_i y_i}}\right).$$
   Hint: Take logs on both sides of the expression for $H(w)$.
   (c) Deduce that this is equivalent to finding w^* that minimizes
$$F(w) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-w^T x_i y_i}\right).$$
   (d) Derive an SGD algorithm for solving this problem. Make sure to specify what your random variables are, as well as your update steps.
   (e) For the SGD algorithm you derived, verify that $\nabla F(w) = \mathbb{E}_{i_t} \nabla f_{i_t}(w)$ by computing both.