Part 1: Theoretical Exercise (16 points)

In class we developed the logistic regression model by requiring that the class labeling probability (for class 1) follows is a sigmoid function of the inner product of the feature vector (x) and the weight vector (w):

$$\Pr[Y = 1|X = x, w] \quad = \quad \sigma(w^\top x) \quad = \quad \frac{1}{1 + e^{-w^\top x}}$$

Using this function for the class labeling probability, we defined the data likelihood, the binary cross entroy (BCE) loss, and a gradient descent algorithm for minimizing the BCE loss.

In this question, you will derive a model and optimization scheme under the following assumptions on the class labeling probabilities:

$$\Pr[Y = 1|X = x, w] = \Phi(w^\top x), \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

This is called the *probit probability model* (unlike the logistic probability model in logistic regression). Recall that $\Phi(x)$ is the CDF of the standard normal variable (with mean 0 and variance 1). The probit model is similar to logistic regression but uses the cumulative normal distribution instead of the sigmoid function. Both models typically yield similar results in practice.

1. Derive the log-likelihood $\ell(w; D)$ under this model. You may assume that the marginal probability of the data, $P_X(\{x^{(i)}\}_{i=1}^n)$, does not depend on the weight vector w (as we assumed in the logistic probability model).
2. Express the problem of maximizing the log-likelihood in this model as a problem of minimizing the appropriate binary cross-entropy (BCE) loss. The BCE loss you specify here **should not** be normalized by the number of samples (n).
3. Find the gradient of the BCE loss and describe how to minimize it using gradient descent.

(1) We notice $P(Y=1/X=X_i, \omega) = \Phi(\omega^t X_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t/x} dt$

so in that case $P(Y=0/X=X_i,\omega) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_i} e^{-t/X_i} dt$

$$\ell(\omega;D) = \prod_{i=1}^{n} [\Phi(\omega^T X_i)]^{y(i)} \cdot [1-\Phi(\omega^T X_i)]^{1-y(i)}$$

$$\log \ell(\omega;D) = \sum_{i=1}^{n} y(i) \log \Phi(\omega^T X_i) + (1-y(i)) \log(1-\Phi(\omega^T \cdot X_i))$$

$$= \sum_{i=1}^{n} y(i) \log\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t/X_i} dt\right) + (1-y(i)) \log\left(1-\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t/X_i} dt\right)\right)$$

(2) maximizing log likelihood $\Longleftrightarrow$ minimizing negative log likelihood

that means $BCE(\omega;D) = -\ell(\omega;D) =$

$$= \sum_{i=1}^{n} y(i) \log \Phi(\omega^T X_i) + (1-y(i)) \log(1-\Phi(\omega^T X_i)$$

$$= \sum_{i=1}^{n} y(i) \log\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t/X_i} dt\right) + (1-y(i)) \log\left(1-\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-t/X_i} dt\right)\right)$$

(3) we denote $\omega^T x^{(i)} = z_i$

$$\nabla_\omega BCE = \left(-\sum_{i=1}^{n} y(i) \log \Phi(z_i) + (1-y(i)) \log (1-\Phi(z_i))\right)'$$

$$= -\sum_{i=1}^{n} \left(y(i) \log \Phi(z_i) + (1-y(i)) \log (1-\Phi(z_i))\right)'$$

$$= -\sum_{i=1}^{n} \left(y(i) \frac{\Phi(z_i)'}{\Phi(z_i)} + (1-y(i)) \cdot \frac{\Phi(z_i)'}{1-\Phi(z_i)} \right) x(i)$$

(4) no, we can see there already a col of 1's (bias)

(5)

step	gradient	loss	weights
1	[-1.5958 -7.9788 -4.7873 -5.5852 -3.9894]	5.5452	[0.016 0.0798 0.0479 0.0559 0.0399]
2	[-0.3158 -6.11 -3.3782 -3.6027 -0.378]	4.436	[0.0191 0.1409 0.0817 0.0919 0.0437]
3	[0.1454 -5.1612 -2.6061 -2.7931 0.8036]	3.8773	[0.0177 0.1925 0.1077 0.1198 0.0356]
4	[0.3326 -4.5485 -2.0851 -2.3789 1.2065]	3.4854	[0.0143 0.238 0.1286 0.1436 0.0236]

(6) predicted lables - [1, 0, 1, 1, 1, 1, 1, 1]

errors

Accuracy - 0.75