Problem Set 3
ECE 685 Fall 2020

Name: Guillem Amat Castello

Problem 1: Supervised Learning of a Probabilistic Model

1. Our cost function will be the following:

$$\mathcal{L}(w|\mathcal{D}) = -ln(f(z|x,w)) = z + e^{\mu(x,w)-z} - \mu(x,w)$$

2. Compute the gradient of $\mathcal{L}(w|\mathcal{D})$ with respect to the parameter vector w:

$$\nabla_w \mathcal{L}(w|\mathcal{D}) = \begin{pmatrix} \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1} \\ \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2} \\ \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_3} \end{pmatrix} = \begin{pmatrix} 2x_1(e^{\mu(x,w)} - 1)(1 + \sigma(a_2)(1 - \sigma(a_2))) \\ x_2(e^{\mu(x,w)} - 1)(\sigma(a_2)(1 - \sigma(a_2))) \\ 0 \end{pmatrix}$$

3. Write pseudo-code for minimizing $\mathcal{L}(w|\mathcal{D})$ using Stochastic Gradient Descent:

Initialize $\eta > 0$, $batch = n$, $w = 0^n$
1. for t in $[1, 2...T]$ $\leftarrow$ Number of iterations
2. for b in $[1, 2...B]$ $\leftarrow$ Number of batches
   2.1 Draw a sample of size n
   2.2 Calculate the loss: $\mathcal{L}(w|\mathcal{D})$
   2.3 Compute the gradient: $\nabla W = \nabla_w \mathcal{L}(w|\mathcal{D})$
   2.4 Update weights: w = w - $\eta \nabla W$

4. Compute the Hessian of $\mathcal{L}(w|\mathcal{D})$ with respect to the parameter vector w:

$$H_f(p) = \begin{bmatrix} \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1^2} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1 w_2} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1 w_3} \\ \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2 w_1} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2^2} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2 w_3} \\ \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_3 w_1} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_3 w_2} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_3^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1^2} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_1 w_2} & 0 \\ \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2 w_1} & \frac{\partial \mathcal{L}(w|\mathcal{D})}{\partial w_2^2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

5. Write a pseudo-code for minimizing $\mathcal{L}(w|\mathcal{D})$ using Newton's method:

Initialize $\eta > 0$, $w = 0^n$
1. for t in $[1, 2...T]$ $\leftarrow$ Number of iterations
  1.1 Calculate the loss: $\mathcal{L}(w|\mathcal{D})$
  1.2 Compute the Gradient: $\nabla W = \nabla_w \mathcal{L}(w|\mathcal{D})$
  1.3 Compute the Hessian: $H_f(p) = \nabla_w^2 \mathcal{L}(w|\mathcal{D})$
  1.4 Apply the Hessian to the Gradient: $\nabla \mathcal{H} = [H_f(p)]^{-1} \nabla W$
  1.5 Update weights: w = w - $\eta \nabla \mathcal{H}$

Problem 2: Bayes Decision Rule

The indicator loss for a function of f(x) of a binary target t is given by the expression below:

$$L(f(x), t) = \begin{cases} 0, & \text{if } f(x) = t \\ 1, & \text{if } f(x) \neq t \end{cases}$$

The expectation of the total loss will be the expected amount of miss-classified observations:

$$L(f(x), t) = \begin{cases} P(t = 1|x), & \text{if } f(x) = -1 \\ P(t = -1|x), & \text{if } f(x) = 1 \end{cases}$$

Let $x = P(t = 1|x)$ and $y = P(t = -1|x)$. Notice that $P(t = -1|x) = 1 - P(t = 1|x)$. The Bayes Decision rule for the Binary case will be given by the following expression:

$$L(f(x), t) = \begin{cases} 1, & \text{if } x > y \\ 0, & \text{if } x < y \end{cases}$$

And the Bayes Decision Boundary will be the point where x and y are equal, that is:

$$P(t = 1|x) > 1 - P(t = -1|x) \Leftrightarrow 2P(t = 1|x) > 1 \Leftrightarrow P(t = 1|x) > \frac{1}{2}$$

Problem 3.3: Binary Classification with Generalized Linear Models

The log-odds for the binary case in LDA are given by the expression below:

$$\begin{aligned}\frac{P(G=k|X=x)}{P(G=K|X=x)} &= log\frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K) + x^T\Sigma^{-1}(\mu_k-\mu_K) \\ &= \alpha_{k_0} + \alpha_k^T x\end{aligned} \tag{1}$$

The log-odds for the binary case in Logistic Regression are given by the expression below:

$$\frac{P(G=k|X=x)}{P(G=K|X=x)} = \beta_{k_0} + \beta_k^T x \tag{2}$$

Although the two models have the same form as it can be seen in (1) and (2), the difference lies in the way the linear coefficients are estimated. The Logistic Regression Model only specifies the conditional distribution $P(G=k|X=x)$, it does not attempt to identify $P(X)$. Instead it leaves the marginal density of X as an arbitrary density function. Logistic Regression then proceeds to maximize the conditional likelihood of G given X.

LDA, on the other hand, makes more assumptions on the distribution of the data and it maximizes the full log-likelihood of the joint density. It estimates the parameters π_k, μ_k and a common covariance matrix Σ. LDA then assumes a mixture of Gaussian densities for $P(X)$:

$$P(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X; \mu_k, \Sigma)$$