CSE – 426 Homework 2

Griffin Kent

1) Verify that $\sigma(-z) = 1 - \sigma(z)$.

(Proof): By the definition of the sigmoid function, we know that

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Thus, we have

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

Therefore, completing the proof.■

2) Suppose the probability of $y = 1$ given z is $\sigma(z) = \mathbb{P}(y = 1|z)$. Using the fact that $y \in \{0,1\}$, verify that $f(y) = \sigma(z)^{y}\big(1-\sigma(z)\big)^{1-y}$ is indeed the conditional probability of the random variable y given z.

(Proof): Since $y \in \{0,1\}$, it can only take on one of two values and thus is a Bernoulli distributed random variable. The standard form of the probability density function of a Bernoulli random variable is given by

$$f(Y = y) = \mathbb{P}(y)^{y}(1-\mathbb{P}(y))^{1-y}$$

Where $\mathbb{P}(y)$ is defined as the probability of y taking on the value of 1. Since in this case $\mathbb{P}(y = 1|z) = \sigma(z)$, we obtain the following conditional probability density function for y given z by substituting in the defined probability:

$$f(y|z) = \sigma(z)^{y}\big(1-\sigma(z)\big)^{1-y}$$

Thus, completing the proof.■

3) Find $\frac{d}{dz}\log\big(1-\sigma(z)\big)$ and $\frac{d}{dz}\log\big(\sigma(z)\big)$.

(Solution): Solving for $\frac{d}{dz}\log\big(1-\sigma(z)\big)$ first, we have

$$\frac{d}{dz}\log\big(1-\sigma(z)\big) = \frac{d}{dz}\log\big(\sigma(-z)\big) = \frac{d}{dz}\log\left(\frac{1}{1+e^{z}}\right)$$
$$= (1+e^{z})\frac{d}{dz}\left(\frac{1}{1+e^{z}}\right)$$

$$= (1+e^{z})\left(\frac{(1+e^{z})\frac{d}{dz}(1)-(1)\frac{d}{dz}(1+e^{z})}{(1+e^{z})^2}\right)$$

$$= \frac{0-e^{z}\frac{d}{dz}(z)}{(1+e^{z})} = -\frac{e^{z}}{1+e^{z}} = -\frac{e^{-z}e^{z}}{e^{-z}(1+e^{z})}$$

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

Now, solving for $\frac{d}{dz}\log(\sigma(z))$, we have the following

$$\frac{d}{dz}\log(\sigma(z)) = \frac{d}{dz}\log\left(\frac{1}{1+e^{-z}}\right) = (1+e^{-z})\frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right)$$

$$= (1+e^{-z})\left(\frac{(1+e^{-z})\frac{d}{dz}(1)-(1)\frac{d}{dz}(1+e^{-z})}{(1+e^{-z})^2}\right)$$

$$= \frac{0-e^{-z}\frac{d}{dz}(-z)}{(1+e^{-z})} = \frac{e^{-z}}{1+e^{-z}}$$

Then, from the properties we proved in question (1) above, we have

$$\frac{d}{dz}\log(\sigma(z)) = \frac{e^{-z}}{1+e^{-z}} = 1-\sigma(z).$$

■

4) Find the partial derivatives of the first entry of the output of the softmax function

$$softmax([z_1, z_2, \ldots, z_k]) = \left[\frac{e^{z_1}}{\sum_j e^{z_j}}, \frac{e^{z_2}}{\sum_j e^{z_j}}, \ldots, \frac{e^{z_k}}{\sum_j e^{z_j}}\right] = [\phi_1, \phi_2, \ldots, \phi_k]$$

With respect to z_1 and z_2.

(Solution): To start, we will take the partial derivative of the first entry of the output of the softmax function with respect to z_1.

$$\frac{\partial \phi_1}{\partial z_1} = \frac{\partial}{\partial z_1}\left(\frac{e^{z_1}}{\sum_j e^{z_j}}\right) = \frac{\left(\sum_j e^{z_j}\right)\frac{\partial}{\partial z_1}(e^{z_1}) - (e^{z_1})\frac{\partial}{\partial z_1}\left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{\left(\sum_j e^{z_j}\right)(e^{z_1})}{\left(\sum_j e^{z_j}\right)^2} - \frac{(e^{z_1})\frac{\partial}{\partial z_1}(e^{z_1}+e^{z_2}+\cdots+e^{z_k})}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{(e^{z_1})}{\sum_j e^{z_j}} - \frac{(e^{z_1})(e^{z_1})}{\left(\sum_j e^{z_j}\right)^2} = \phi_1 - \phi_1^2.$$

Now, taking the partial derivative of the first entry of the output of the softmax function with respect to z_2, we have

$$\frac{\partial \phi_1}{\partial z_2} = \frac{\partial}{\partial z_2}\left(\frac{e^{z_1}}{\sum_j e^{z_j}}\right) = \frac{\left(\sum_j e^{z_j}\right)\frac{\partial}{\partial z_2}(e^{z_1}) - (e^{z_1})\frac{\partial}{\partial z_2}\left(\sum_j e^{z_j}\right)}{\left(\sum_j e^{z_j}\right)^2}$$

$$= \frac{0 - (e^{z_1})(e^{z_2})}{\left(\sum_j e^{z_j}\right)^2} = 0 - \frac{(e^{z_1})(e^{z_2})}{\left(\sum_j e^{z_j}\right)^2} = -\phi_1\phi_2.$$

■

5) (Graduate Only) Prove that the Hessian matrix of the negative log-likelihood function for logistic regression is positive semi-definite.

(Proof): The negative log-likelihood function (cross-entropy loss function) for logistic regression is defined as the following

$$-\ell(\boldsymbol{\theta}) = -\log L(\boldsymbol{\theta}) = -\sum_{i=1}^{m}\left\{y^{(i)}\log\sigma\left(z^{(i)}\right) + \left(1 - y^{(i)}\right)\log\sigma\left(-z^{(i)}\right)\right\}$$

Where m is the number of training examples $\left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$. Likewise, the Hessian matrix of the negative log-likelihood function is defined as

$$\boldsymbol{H} = \sum_{i=1}^{m}\boldsymbol{x}^{(i)}\left(\boldsymbol{x}^{(i)}\right)^T\sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\left(1 - \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\right)$$

$$\boldsymbol{H} = \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T$$

Where $\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\left(1 - \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\right)\right)$ is a diagonal matrix with the i-th diagonal element being $\sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\left(1 - \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\right)$ and $\boldsymbol{X}^T = \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(m)}\right]$, where each $\boldsymbol{x}^{(i)}$ is a column vector. (*Logistic Regression Lecture Notes*).

By definition, a matrix will be positive semi-definite if all its eigenvalues $\lambda_i \geq 0$. Since $\boldsymbol{H}$ is a symmetric matrix, (Using the properties defined in Section 3.13 of the *Linear Algebra Review and Reference*) we know that $\mathbf{X}$ is an orthogonal matrix and $\mathbf{\Sigma}$ is the diagonal matrix whose entries are the eigenvalues of $\boldsymbol{H}$. Thus, we have

$$\lambda_i = \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\left(1 - \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\right)$$

Since the sigmoid function maps any real number to the range of (0,1), we know that

$$\lambda_i = \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\left(1 - \sigma\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)\right) \geq 0, \forall i = 1,2,\dots,m.$$

Therefore, since we have shown for the hessian matrix $\boldsymbol{H}$ that the eigenvalues $\lambda_i \geq 0$, we know that the matrix $\boldsymbol{H}$ must be positive semi-definite.

■