8th Homework

Thomas Saltos

Exercise 1

The negative log-likelihood cost function of the logistic regression problem is defined as

$$L(\theta) = -\sum_{n=1}^{N}\big(y_n ln(s_n) + (1-y_n)ln(1-s_n)\big)$$

Where $s_n = \sigma(\theta^T x_n)$. The gradient $\nabla_\theta L(\theta)$ of $L(\theta)$ is estimated as follows:

$$\nabla_\theta L(\theta) = -\sum_{n=1}^{N}\frac{y_n \partial s_n}{s_n \partial\theta} + \frac{1-y_n}{1-s_n}\left(-\frac{\partial s_n}{\partial\theta}\right)$$

Applying the chain rule, we have $\frac{\partial s_n}{\partial\theta} = \frac{\partial s_n}{\partial t}\frac{\partial t}{\partial\theta}$, where $t = \theta^T x$, and hence

$$\frac{\partial s_n}{\partial\theta} = s_n(1-s_n)x_n$$

Since $\frac{\partial s_n}{\partial t} = s_n(1-s_n)$. Therefore, the previous equation is written as

$$\nabla_\theta L(\theta) = -\sum_{n=1}^{N}\frac{y_n}{s_n}s_n(1-s_n)x_n + \frac{y_n-1}{1-s_n}(s_n(1-s_n)x_n) = -\sum_{n=1}^{N}(y_n(1-s_n) + (y_n-1)s_n)x_n$$

$$= \sum_{n=1}^{N}(s_n - y_n)x_n = X^T(s-y)$$

where $X^T = [x_1, x_2, \dots, x_N], s = [s_1, s_2, \dots, s_N]^T, y = [y_1, y_2, \dots, y_N]^T$

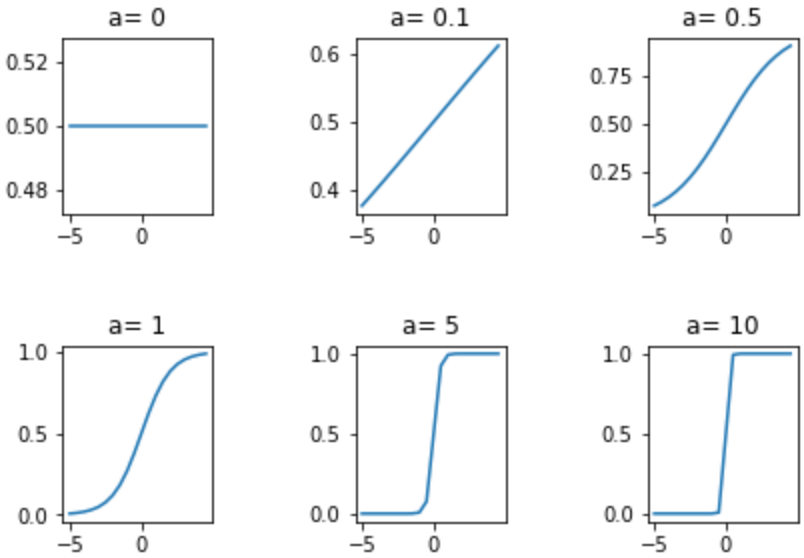
The gradient descent iteration consists of the following updating rule:

$$\theta_i = \theta_{i-1} - \mu\, X^T\left(s^{(i-1)} - y\right)$$

where $s^i = \left[\sigma\left(\theta_i^T x_1\right), \sigma\left(\theta_i^T x_2\right), \dots, \sigma\left(\theta_i^T x_N\right)\right]^T$ and μ is the learning rate. After convergence of the iterative algorithm we get the estimate for θ i.e.,$\hat{\theta}$. Then, a given x_{test} is classified to ω_1 if $\sigma\left(\widehat{\theta^T} x_{test}\right) > 0$ and to class ω_2 otherwise.

Exercise 2

a)



It can be observed that as a increases, $f(z)$ approaches the unit step function.

b) A gradient scheme for estimating the involved parameters based on the minimization of $J(\theta)$ will have the following form:

$$\theta_i = \theta_{i-1} - \mu[\nabla_\theta J(\theta)]_{\theta=\theta_{i-1}}$$

Where $J(\theta) = \sum_{n=1}^{N}\left(y_n - f(\theta^T x_n)\right)^2$. The gradient is:

$$\nabla_\theta J(\theta) = -2\sum_{n=1}^{N}\left(y_n - f(\theta^T x_n)\right)(\alpha f(\theta^T x_n)\left(1 - f(\theta^T x_n)\right)x_n$$

We have that:

$$\theta_i = \theta_{i-1} + 2\mu\sum_{n=1}^{N}\left(y_n - f(\theta^T x_n)\right)(\alpha f(\theta^T x_n)\left(1 - f(\theta^T x_n)\right)x_n$$

c) As it can be also shown from the figures above, the model will tend to 1 as $e^{-\alpha z} \Rightarrow 0$, which implies that $\alpha\theta^T x \to +\infty$. On the other hand, 0 is approached as $\alpha\theta^T x \to -\infty$. Therefore, clear 0 or 1 is impossible for finite values of α and x in practice.

d) For a given x_n, the model classifies it based in the value of $f(\theta^T x_n)$. This value actually can be interpreted as the probability of x_n, to belong to the class corresponding to $y_n = 1$. Therefore, if $f(\theta^T x_n) > 0.5, x_n$ is classified to 1 while if $f(\theta^T x_n) < 0.5$ to class 0.

e) A way to force the model responses to 1 for class 1 and 0 for class 0 is via the increase of the parameter α so that the model to approach the unit step function. This can be also verified from the figures given in part (a) above.