Homework 8

Exercise 1

- Likelihood function is:

$\prod_{n=1}^{N} P(y_n|x_n;\theta) = \prod_{n=1}^{N} s_n^{y_n}(1-s_n)^{1-y_n}$, where $s_n = \sigma(\theta^T x_n)$

- Negative log-likelihood function is:

$L(\theta) = -\sum_{n=1}^{N}(y_n \ln s_n + (1-y_n)\ln(1-s_n))$, where $s_n = \sigma(\theta^T x_n)$

- The gradient of the negative log-likelihood is:

$\nabla L(\theta) = -\sum_{n=1}^{N} \frac{y_n}{s_n}\frac{\partial s_n}{\partial \theta} + \frac{1-y_n}{1-s_n}\left(-\frac{\partial s_n}{\partial \theta}\right)$, but $\frac{\partial s_n}{\partial \theta} = \frac{\partial s_n}{\partial t}\frac{\partial t}{\partial \theta} = \frac{\partial s_n}{\partial \theta^T x_n}\frac{\partial \theta^T x_n}{\partial \theta} = s_n(1-s_n)x_n$

- Thus, gradient of the negative log-likelihood becomes:

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

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

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

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

Where $X^T = [x_1, \dots, x_N]$, $s = [s_1, \dots, s_N]^T$, $y = [y_1, \dots, y_N]^T$

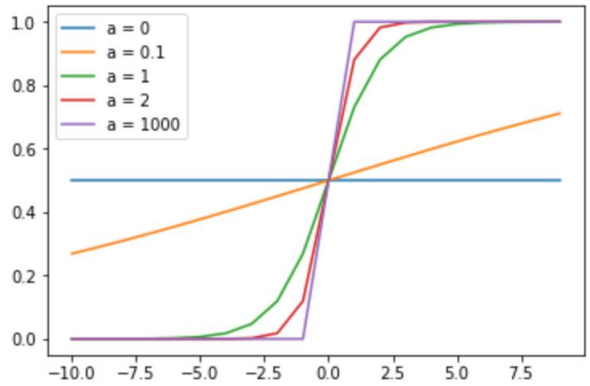
- The Hessian of L(θ) is positive definite. Thus L(θ) is convex and it has a single minimum. Due to the presence of the nonlinear function, equating the gradient to zero and solving does not lead to an analytic form solution. Thus, we resort to approximate techniques, such as gradient descent. Updating rule for logistic regression is:

$\theta = \theta_{i-1} - \mu X^T\left(s^{i-1} - y\right)$, ${s_n}^{i-1} = \sigma({\theta_{i-1}}^T x_n)$, μ is the learning rate.

Using the estimate of θ, we calculate the amount $\sigma(\theta^T x_0)$ for a given x_0. If $\sigma(\theta^T x_0) > 0$ we assign x_0 to class 1. Otherwise, we assign to class 2.

Exercise 2

(a) Plot function $f(z) = \frac{1}{1+\exp(-az)}$ for various values of a:



(b) As shown in Exercise 1, the gradient descent step is $\theta = \theta_{i-1} - \mu\nabla L(\theta)$, with $L(\theta)$ being a cost function. By substituting this cost function with the MSE one we get:

$\theta = \theta_{i-1} - \mu\nabla L(\theta)$,

$$L(\theta) = \sum_{n=1}^{N} (y_n - \mathrm{f}(\theta^T x_n))^2$$

$$\nabla L(\theta) = -2\sum_{\mathrm{n=1}}^{\mathrm{N}} (y_n - \mathrm{f}(\theta^T x_n))\mathrm{f}(\theta^T x_n)(1 - \mathrm{f}(\theta^T x_n))x_n$$

Thus, $\theta = \theta_{i-1} + 2\mu\sum_{\mathrm{n=1}}^{\mathrm{N}}(y_n - \mathrm{f}(\theta^T x_n))\mathrm{f}(\theta^T x_n)(1 - \mathrm{f}(\theta^T x_n))x_n$

(c) As it can be seen from the diagrams above, the value of f tends to 0 and 1 but never actually reaches it. Therefore, we cannot respond with a clear 0 or 1.

(d) For a given x, we could say that the value of the function gives us the probability of this x belonging to a class 1. If f(.)>0.5 we assign to class 1, otherwise to class 0.

(e) By increasing the value of a we transform the sigmoid function to the step function, and we force it to give values 0 or 1 as a response.

Exercise 3

Use an index that either starts from the smallest or the largest number of the data set. Let's choose the smallest and set the threshold there. Count how many points of class 1 where classified incorrectly and how many points of class -1 where classified incorrectly. Add the two numbers and keep the result (cost function). Add 1 to the index (move it to the right) and calculate the value of cost function. Finally, when reaching the largest number stop and choose as threshold the value of the index that gives the smallest value to the cost function. If more than one values satisfies the criterion choose one randomly.