CS4780/5780 Homework 5

Due: TBD

Problem: Logistic Regression

In this problem, we are going to assume the same notation setup in class. For logistic regression, we model the class probability by

$$P(y|\vec{x}_i) = \sigma(y(\vec{w}^T\vec{x}_i))$$

where we define

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

(Note: We dropped the bias term b since we can always absorb the bias into $\vec{w}$)

1. Show that the sigmoid function $\sigma(\cdot)$ has the following property

   $$\sigma(-s) = 1 - \sigma(s)$$

   By proving this property, we have shown that we have properly defined a probabilistic model, namely, $P(y_i = 1|\vec{x}_i) + P(y_i = -1|\vec{x}_i) = 1$

2. In class, we mentioned about using Gradient Descent to find the MLE estimate for $\vec{w}$. One of the problems with Gradient Descent is that the algorithm sometimes gets stuck at a local optima rather than the global optima. In this question, we are going to show that for logistic regression, Gradient Descent always return the global optima.

   (a) To make things easier, first show that

   $$\sigma'(s) = \sigma(s)(1 - \sigma(s))$$

   (b) Show that the gradient of the log likelihood function, namely, $\nabla_w \log P(\vec{y}|X, \vec{w})$ where $\vec{y} = [y_1, y_2, ..., y_n]^T$ and $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ is

   $$\sum_{i=1}^{n}(1 - \sigma(y_i(\vec{w}^T\vec{x}_i)))y_i\vec{x}_i$$

   (c) Show that the Hessian H of the log likelihood function is

   $$-\sum_{i=1}^{n}\sigma(y_i(\vec{w}^T\vec{x}_i))(1 - \sigma(y_i(\vec{w}^T\vec{x}_i)))y_i^2\vec{x}_i\vec{x}_i^T$$

To show this, first show that

$$H_{ab} = \frac{\partial^2}{\partial w_a \partial w_b} \log P(\vec{y}|X, \vec{w}) = -\sum_{i=1}^{n} \sigma(y_i(\vec{w}^T\vec{x}_i))(1 - \sigma(y_i(\vec{w}^T\vec{x}_i)))y_i^2 x_{ia} x_{ib}$$

where w_k is the k-th entry of the weight vector $\vec{w}$ and x_{ik} is the k-th entry of $\vec{x}_i$. Then verify that (a,b)-th entry of the matrix $\vec{x}_i\vec{x}_i^T$ is indeed $x_{ia}x_{ib}$.

(d) Show that the Hessian is negative semi-definite, namely, $\vec{z}^T H \vec{z} \leq 0$ for any vector $\vec{z} \in \mathbb{R}^d$. By showing this, we have shown that the log likelihood is concave and has no local maxima except the global one.