CSE 151A - Discussion 05

Quick Review

Logistic Regression
Predict a probability $H_{\vec{w}}(\vec{x}) = \sigma(\vec{w} \cdot Aug(\vec{x}))$, with logistic function $\sigma(t) = \frac{1}{1+e^{-t}}$
Goal : find the value of $\vec{w}$ to maximize the log-likelihood $f(\vec{w}) = log\mathcal{L}(\vec{w})$ using gradient ascent

Maximum Likelihood
In general, the likelihood is : $\mathcal{L}(\vec{w}) = \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i \vec{w} \cdot Aug(\vec{x}^{(i)})}}$

The log likelihood is : $f(\vec{w}) = log\mathcal{L}(\vec{w}) = -\sum_{i=1}^{n} log\big[1 + e^{-y_i \vec{w} \cdot Aug(\vec{x}^{(i)})}\big]$

*NOTE : Here we have assumed $y_i = 1$ for a positive class label, and $y_i = -1$ for a negative class label. You may encounter equations that look different to the ones used here, and this is likely due the use of $y_i = 0$ rather than $y_i = -1$ for a negative class label.

Gradient Ascent
Setting : $f(\vec{w})$ is differentiable but we cannot explicitly solve for $\vec{w}$ like before
Strategy : pick a starting guess $\vec{w}^{(0)}$ and iterate $\vec{w}^{(i)} = \vec{w}^{(i-1)} + \alpha \cdot \nabla f(\vec{w}^{(i-1)})$ until convergence, where $\nabla f(\vec{w}^{(i-1)}) = \sum_{k=1}^{n} y_k \vec{x}^{(k)} H_{\vec{w}^{(i-1)}}(-y_k \vec{x}^{(k)})$

Making Classifications
Predict class 1 if $H_{\vec{w}}(\vec{x}) > \tau$ $\quad\tau$ can be thought of as a threshold probability (y-value of logistic function)
Predict class 1 if $\vec{w} \cdot Aug(\vec{x}) > t$ $\quad t$ can be thought of as an x-value threshold on the logistic function

Problem 1.

Consider the equation $f(w) = -(x^2 - 5x + 4)$.

a) Show that this function is strictly concave. What does this tell us about the number of maxima?

b) Use gradient ascent to solve for the value of x that maximizes $f(x)$.
For this problem, start with $x^{(0)} = 0$, and compute all intermediate $x^{(i)}$ values up to $x^{(4)}$.
Do this three times with the following values of $\alpha : 0.3, 0.8, 1.2$. For each, note any interesting findings and determine if the algorithm will eventually converge.

Solution:
a. We will take advantage of the fact that f is twice differentiable.
By definition, a function $f(x)$ is strictly concave on $[a, b]$ if $f''(x) < 0$ for all $x \in (a, b)$.
Solving the derivative yields $f'(x) = -(2x - 5)$ and $f''(x) = -2$.
We have shown that $f''(x) < 0$ for all x and that f is therefore strictly concave.
As a result, we can also conclude that there exists only one local maximum, which must also be the global maximum.

b. Recall the update rule for gradient ascent : $\vec{w}^{(i)} = \vec{w}^{(i-1)} + \alpha \cdot \nabla f(\vec{w}^{(i-1)})$.
We will simplify this equation to fit the context of our problem, where $\vec{w}$ is simple a single value x.
We now have the following update rule : $x^{(i)} = x^{(i-1)} + \alpha \cdot f'(x^{(i-1)})$.
We solved for $f'(x)$ in part a, so we can rewrite this equation as $x^{(i)} = x^{(i-1)} + \alpha \cdot -(2x^{(i-1)} - 5)$.

For $x^{(0)} = 0$ and $\alpha = 0.3$
$x^{(1)} = 0 + 0.3 \cdot -(2(0) - 5) = \boxed{1.5}$
$x^{(2)} = 1.5 + 0.3 \cdot -(2(1.5) - 5) = \boxed{2.1}$
$x^{(3)} = 2.1 + 0.3 \cdot -(2(2.1) - 5) = \boxed{2.34}$
$x^{(4)} = 2.34 + 0.3 \cdot -(2(2.34) - 5) = \boxed{2.436}$
With $\alpha = 0.3$, the value of $x^{(i)}$ at each iteration slowly approaches from the left the value of x that maximizes $f(x)$. **Yes** we will eventually get **convergence**.

For $x^{(0)} = 0$ and $\alpha = 0.8$
$x^{(1)} = 0 + 0.8 \cdot -(2(0) - 5) = \boxed{4}$
$x^{(2)} = 4 + 0.8 \cdot -(2(4) - 5) = \boxed{1.6}$
$x^{(3)} = 1.6 + 0.8 \cdot -(2(1.6) - 5) = \boxed{3.04}$
$x^{(4)} = 3.04 + 0.8 \cdot -(2(3.04) - 5) = \boxed{2.176}$
With $\alpha = 0.8$, the value of $x^{(i)}$ at each iteration bounces back and forth around the value of x that maximizes $f(x)$. **Yes** we will eventually get **convergence**.

For $x^{(0)} = 0$ and $\alpha = 1.2$
$x^{(1)} = 0 + 1.2 \cdot -(2(0) - 5) = \boxed{6}$
$x^{(2)} = 6 + 1.2 \cdot -(2(6) - 5) = \boxed{\text{-2.4}}$
$x^{(3)} = -2.4 + 1.2 \cdot -(2(-2.4) - 5) = \boxed{9.36}$
$x^{(4)} = 9.36 + 1.2 \cdot -(2(9.36) - 5) = \boxed{\text{-7.104}}$
With $\alpha = 1.2$, the value of $x^{(i)}$ at each iteration also bounces back and forth around the value of x that maximizes $f(x)$. However, **no**, the algorithm will **NOT converge**.

Problem 2.

After running gradient descent, suppose that we have solved for $\vec{w} = (0.5, 2, -1)^T$ that minimizes some convex function f.
We have the following validation set, consisting of four data points and their corresponding labels:

i	$x_1^{(i)}$	$x_2^{(i)}$	y_i
1	2	4	-1
2	3	2	1
3	0	-1	-1
4	-1	2	-1

* Note : Don't forget to augment each $\vec{x}^{(i)}$

a) We will use the following rule : Predict class 1 if $H_{\vec{w}}(\vec{x}) > \tau$, else predict class -1.
What is the classification accuracy over the above validation set when $\tau = 0.5$?

b) Is there a value of τ that would result in 100% validation accuracy? If so, compute such a value.

Solution:
a. Recall $H_{\vec{w}}(\vec{x}^{(i)}) = \sigma(\vec{w} \cdot Aug(\vec{x}^{(i)})) = \sigma(w_0 \cdot 1 + w_1 \cdot x_1^{(i)} + w_2 \cdot x_2^{(i)})$.
Computing this value for each data point yields the following:
$H_{\vec{w}}(\vec{x}^{(1)}) = \sigma(w_0 \cdot 1 + w_1 \cdot x_1^{(1)} + w_2 \cdot x_2^{(1)})$
$= \sigma((0.5)(1) + (2)(2) + (-1)(4))$
$= \sigma(0.5) = 0.622$
$0.622 > 0.5 \rightarrow$ Predict 1 (INCORRECT because $y_1 = -1$)

$H_{\vec{w}}(\vec{x}^{(2)}) = \sigma(w_0 \cdot 1 + w_1 \cdot x_1^{(2)} + w_2 \cdot x_2^{(2)})$
$= \sigma((0.5)(1) + (2)(3) + (-1)(2))$
$= \sigma(4.5) = 0.989$
$0.989 > 0.5 \rightarrow$ Predict 1 (CORRECT because $y_2 = 1$)

$H_{\vec{w}}(\vec{x}^{(3)}) = \sigma(w_0 \cdot 1 + w_1 \cdot x_1^{(3)} + w_2 \cdot x_2^{(3)})$
$= \sigma((0.5)(1) + (2)(0) + (-1)(-1))$
$= \sigma(1.5) = 0.818$
$0.818 > 0.5 \rightarrow$ Predict 1 (INCORRECT because $y_3 = -1$)

$H_{\vec{w}}(\vec{x}^{(4)}) = \sigma(w_0 \cdot 1 + w_1 \cdot x_1^{(4)} + w_2 \cdot x_2^{(4)})$
$= \sigma((0.5)(1) + (2)(-1) + (-1)(2))$
$= \sigma(-3.5) = 0.029$
$0.029 \not> 0.5 \rightarrow$ Predict -1 (CORRECT because $y_4 = -1$)

Therefore, the classification accuracy with $\tau = 0.5$ is $\boxed{50\%}$.

b. To determine if we can achieve 100% validation accuracy, we need to look at the largest value of $H_{\vec{w}}(\vec{x}^{(i)})$ that corresponds to a negative label of -1 and the smallest value of $H_{\vec{w}}(\vec{x}^{(i)})$ that corresponds to a positive label of 1. This will tell us if our validation data is perfectly separable into classes. From the data we see that these values are 0.818 (with label -1), and 0.989 (with label 1).
Therefore, we can conclude that any value of τ in the range $\boxed{0.818 < \tau < 0.989}$ will result in 100% validation accuracy.