1 Vectors & Planes Review (7 pts)

a) Problem A (2pts)

Solution: The magnitude of each vector can be found using formula below:
$\|\bar{x}^{(1)}\| = \sqrt{(\sqrt{3})^2 + (-1)^2} = 2$ and $\|\bar{x}^{(2)}\| = \sqrt{(1)^2 + (0)^2} = 1$
The angle between them in degrees is shown below:

$$\begin{aligned}\cos(\theta) &= \frac{\bar{x}^{(1)} \cdot \bar{x}^{(2)}}{\|\bar{x}^{(1)}\|\|\bar{x}^{(2)}\|} \\ &= \frac{\sqrt{3}}{2 \cdot 1} \\ &= \frac{\sqrt{3}}{2}\end{aligned}$$

Since $\cos(\theta)) = \frac{\sqrt{3}}{2}$, applying inverse cos to that value gives the θ value of 30 degrees.
$\bar{x}^{(1)} = 2, \bar{x}^{(2)} = 1, \theta = 30°$

b) Problem B (1pt)

Solution: All d-dimensional vectors that are orthogonal to S are parallel to the vector $\bar{\theta}$. The vectors, however, have to be the length C. Thus the vector $\bar{v}$ is equal to $C\frac{\bar{\theta}}{\|\bar{\theta}\|}$ Make sure to add sign to as there are two orthogonal vectors on the plus side and negative side.
Final Answer: $\bar{v} = \pm C\frac{\bar{\theta}}{\|\bar{\theta}\|}$

c) Problem C (4pt)

Solution:

(i) Suppose there is an arbitrary point $\bar{x}_0$ on the plane and a vector $\bar{v}$ pointing from $\bar{x}_0$ to $\bar{x}$. To find the signed distance of $\bar{x}$ from hyperplane, first define the projection of $\bar{v}$ onto $\bar{\theta}$

$proj_{\bar{\theta}}(\bar{v}) = \frac{\bar{\theta}\cdot\bar{v}}{||\bar{\theta}||} \cdot \frac{\bar{\theta}}{||\bar{\theta}||}$ Note that the projection is normalized by multiplying with the unit vector.

Following shows the distance between the vector $\bar{x}$ and the hyperplane. Note that $\bar{\theta}\cdot\bar{x}_0 = 0$ as they are on the same plane

$$d = \frac{\bar{\theta}\cdot\bar{v}}{||\bar{\theta}||} = \frac{\bar{\theta}\cdot(\bar{x}-\bar{x}_0)}{||\bar{\theta}||} = \frac{\bar{\theta}\cdot\bar{x}}{||\bar{\theta}||}$$

(ii) Adding offset b changes $\bar{\theta}\cdot\bar{x}_0 = 0$ to $\bar{\theta}\cdot\bar{x}_0 = -b$. Add b to the numerator.

$$d = \frac{\bar{\theta}\cdot\bar{v}}{||\bar{\theta}||} = \frac{\bar{\theta}\cdot(\bar{x}-\bar{x}_0)}{||\bar{\theta}||} = \frac{\bar{\theta}\cdot\bar{x}+b}{||\bar{\theta}||}$$

(iii)

i)

$$d = \frac{(-2,-5)\cdot(4,1)+6}{\sqrt{(-2)^2+(-5)^2}} = \frac{-7}{\sqrt{29}}$$

ii)

$$d = \frac{(-2,-5)\cdot(0,0)+6}{\sqrt{(-2)^2+(-5)^2}} = \frac{6}{\sqrt{29}}$$

2 Linear Binary Classifiers (5 pts)

a) Problem A (1pt)

Solution: Each feature has mapping 0,1. Since there are three words in the bag of model, the feature space is $\{0,1\}^3$

b) Problem B (1pt)

Solution: There are only two possible label space as the only options are positive or negative.

$y \in \{-1, 1\}$

c) Problem C (3pts)

Solution: YES, if we plot the data set, $\bar{x}^{(1)} = [1,1,0]^T, \bar{x}^{(2)} = [1,1,1]^T, \bar{x}^{(3)} = [1,0,1]^T, \bar{x}^{(4)} = [1,0,0]^T$ in $\mathbb{R}^3$ and it is linearly separable by a hyperplane passing through origin.

One hyperplane that classifies the data is defined by $\bar{\theta} = [1,1,-3]^T$ and no offset, so $b = 0$

3 Decision Boundaries (4 pts)

a) Problem A (1pt)

Solution: Inside or outside of a circle centered at the origin with radius 2.

$$f(x,y) = \begin{cases} +1, & \text{if } x^2 + y^2 > 4 \quad \text{(outside the circle, positive class)} \\ -1, & \text{if } x^2 + y^2 \leq 4 \quad \text{(inside or on the circle, negative class)} \end{cases}$$

b) Problem B (1pt)

Solution: Above or below a line through the origin with normal [2, -2].

$$f(x,y) = \begin{cases} +1, & \text{if } 2x - 2y \geq 0 \quad \text{(below or on the line, positive class)} \\ -1, & \text{if } 2x - 2y < 0 \quad \text{(above the line negative class)} \end{cases}$$

c) Problem C (1pt)

Solution: A square centered at [-1,1] with sides parallel to the axes and side length 2

$$f(x,y) = \begin{cases} +1, & \text{if } x \notin [-2,0] \cup y \notin [0,2] \quad \text{(outside the square, positive class)} \\ -1, & \text{if } x \in [-2,0] \cap y \in [0,2] \quad \text{(inside or on the square, negative class)} \end{cases}$$

d) Problem D (1pt)

Solution: Origin with side length $2\sqrt{2}$ and counter-clockwise rotation $\frac{\pi}{4}$ or 45°

$$f(x,y) = \begin{cases} +1, & \text{if } |x+y| > 2 \text{ or } |x-y| > 2 \quad \text{(outside the square, positive class)} \\ -1, & \text{if } |x+y| \leq 2 \text{ and } |x-y| \leq 2 \quad \text{(inside or on the square, negative class)} \end{cases}$$

4 Perceptron Algorithm with Offset (6 pts)

a) Problem A (1pt)

Solution:

The perceptron algorithm initializes the parameters $\bar{\theta}$ and b to zero. The algorithm uses the data point value as the magnitude of the parameter and updates on θ happen whenever the points are misclassified. As a side note, α_i denotes the number of time $\bar{x}^{(i)}$ are misclassified during training.

The final value of $\bar{\theta}$ is a total sum of the product of the number of missclassification, the sign of each miss, and the value of each data point that was missed.

The value of b is a total sum of the number of misclassification and the sign of each miss.

$\bar{\theta}^{(k+1)} = \bar{\theta}^{(k)} + y^{(i)}\bar{x}^{(i)}$, thus $\bar{\theta} = 0 + \sum_{i=1}^{n} \alpha_i y^{(i)}\bar{x}^{(i)}$

$b^{(k+1)} = b^{(k)} + y^{(i)}$, thus $\sum_{i=1}^{n} \alpha_i y^{(i)}$

The final decision boundary is defined by $\bar{\theta} = 0 + \sum_{i=1}^{n} \alpha_i y^{(i)}\bar{x}^{(i)}$ and $b = 0 + \sum_{i=1}^{n} \alpha_i y^{(i)}$

b) Problem B (3pt)

Solution:

i	number of Misclassifications
0	1
1	0
2	11
3	2
4	0
5	4

$\bar{\theta} = [4, 0]^T$ and $b = -6$

$\bar{\theta}$ and b calculated from the expression derived in (a) matches the values produced by the code. Details can be find in check_formula function in the code.

c) Problem C (1pt)

Solution: Changing the order of points will NOT affect the convergence of the perceptron algorithm. Given that the data is linearly separable, the perceptron algorithm is guaranteed to converge to a decision boundary within a finite number of updates.

d) Problem D (1pt)

Solution: YES, Changing the order of the data points can change the final decision boundary that the algorithm converges to. There are multiple possible decision boundaries in perceptron algorithm as the order of the algorithm processing misclassified points gives a variable updates on the weight vector $\bar{\theta}$.

5 LinearRegression- Optimization Methods (18 pts)

5.1 Comparing Optimization Algorithms (12 pts)

a) Problem A (1pt)

Solution: Check the code for all three implementations.

Algorithm	$\boldsymbol{\eta}$	$\boldsymbol{\theta_0}$	$\boldsymbol{\theta_1}$	**# iterations**	**Runtime (s)**
GD	10^{-4}	0.2991	-0.1273	434791	45.74s
GD	10^{-3}	0.3041	-0.1363	60744	5.74s
GD	10^{-2}	0.3057	-0.1391	7800	0.69s
GD	10^{-1}	0.3062	-0.1400	951	0.09s
SGD	10^{-4}	0.3048	-0.1375	659060	4.02s
SGD	10^{-3}	0.3056	-0.1397	82020	0.41s
SGD	10^{-2}	0.3023	-0.1421	7300	0.04s
SGD	10^{-1}	0.2663	-0.1778	2080	0.01s
Closed form	-	0.3063	-0.1408	-	0.001s

b) Problem B (1pt)

Solution: Runtime: For a given learning rate, running time for SGD is significantly faster than GD as theta is updated at each individual point for SGD whereas theta is updates after the whole data iteration for GD.

Number of iteration: Number of iteration in all data points is greater for SGD than GD because $\bar{\theta}$ gets updated on each point for SGD while $\bar{\theta}$ gets updated only once after whole iteration through the data for GD.

Coefficients: Coefficients converge to approximately 0.30 for all GD and SGD. However, SGD has more variability (noise) in coefficient when the algorithm has larger learning rate.

c) Problem C (1pt)

Solution: The runtime for the closed form solution is significantly faster than SGD. Learning rate of 10^{-3} used for SGD $[0.3056, -0.1397]^T$ has the closest coefficients to the coefficients of the closed form solution $[0.3063, -0.1408]^T$

d) Problem D (1pt)

Solution: For the adaptive solution, I used the following function: $\eta_k = \frac{1}{1+k}$ The coefficients for the adaptive solution is $[0.3064, -0.1403]^T$ with a runtime of 0.1469 seconds and 27980 iterations. These coefficients are extremely close to the coefficients from the closed form solution $[0.3063, -0.1404]^T$ and it took 27980 iterations through the data to converge with my proposed learning rate function.

5.2 Weighted Regression (6 pts)

a) Problem A (2pt)

Solution:

$$\begin{aligned} J(\bar{\theta}) &= \frac{1}{2}\sum_{i=1}^{n} w^{(i)}(\bar{\theta}\cdot\bar{x}^{(i)} - \bar{y}^{(i)})^2 \\ &= \frac{1}{2}\sum_{i=1}^{n} w^{(i)}(X\bar{\theta} - \bar{y}^{(i)})^2 \\ &= (X\bar{\theta} - \bar{y}^{(i)})^T W (X\bar{\theta} - \bar{y}^{(i)}) \end{aligned}$$

Note that diagonal matrix W is a diagonal matrix with elements $w^{(1)}, w^{(2)}, \ldots, w^{(n)}$ that "weighs" individual training elements differently. Also, identity $A^2 = A^T A$ is used for the solution.

b) Problem B (2pt)

Solution:

$$\begin{aligned} \nabla_{\bar{\theta}} J &= 2X^T W (X\bar{\theta} - \bar{y}^{(i)}) \\ 2X^T W (X\bar{\theta} - \bar{y}^{(i)}) &= 0 \text{ (set gradient to 0)} \\ X^T W X\bar{\theta} - X^T W\bar{y} &= 0 \\ X^T W X\bar{\theta} &= X^T W\bar{y} \\ \bar{\theta} &= (X^T W X)^{-1} X^T W\bar{y} \end{aligned}$$

Thus the close form solution of $\bar{\theta}$ is $(X^T W X)^{-1} X^T W\bar{y}$

c) Problem C (2pt)

Solution: In time series forecasting, such as predicting stock prices, more recent data points are generally more valuable than older data. Therefore, making weighted regression would be useful here as assigning higher weights/priority to recent observations is crucial.

6 Logistic regression (10 pts)

a) Problem A (2pt)

Solution:

$$\begin{aligned}
\nabla_{\bar{\theta}} \frac{1}{N} \sum_{i=1}^{N} \ln(1 + e^{-y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}) &= \frac{1}{N} \sum_{i=1}^{N} \frac{e^{-y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}}{1 + e^{-y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} (-y^{(i)} \bar{x}^{(i)}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{-y^{(i)} \bar{x}^{(i)}}{1 + e^{y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} \\
&= \frac{1}{N} \sum_{i=1}^{N} -y^{(i)} \bar{x}^{(i)} \frac{1}{1 + e^{y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} \\
&= \frac{1}{N} \sum_{i=1}^{N} -y^{(i)} \bar{x}^{(i)} \sigma(-(y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})) \\
&= \frac{1}{N} \sum_{i=1}^{N} -y^{(i)} \bar{x}^{(i)} (1 - \sigma((y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})))
\end{aligned}$$

b) Problem B (2pt)

Solution:

$$\begin{aligned}
\bar{\theta}^{(k+1)} &= \bar{\theta}^{(k)} - \eta \nabla_{\bar{\theta}} J(\bar{\theta}) \\
&= \bar{\theta}^{(k)} - \eta \frac{-y^{(i)} \bar{x}^{(i)}}{1 + e^{y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} \\
&= \bar{\theta}^{(k)} + \eta \frac{y^{(i)} \bar{x}^{(i)}}{1 + e^{y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} \\
&= \bar{\theta}^{(k)} + \eta y^{(i)} \bar{x}^{(i)} \frac{1}{1 + e^{y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)})}} \\
&= \bar{\theta}^{(k)} + \eta y^{(i)} \bar{x}^{(i)} \sigma(-(y^{(i)}(\bar{\theta} \cdot \bar{x}^{(i)}))
\end{aligned}$$

c) Problem C (1pt)

Solution: $\bar{\theta} =$ [0.2905402 -0.01951301 0.01714351]

d) Problem D (1pt)

Solution: NO, There is no closed-form solution for logistic regression because the cost function involves a non-linear combination of the parameters due to the sigmoid function. This makes it impossible to set the gradient to zero and solve for $\bar{\theta}$