Home Work 1:

Perceptron Mistakes:

Perceptron algorithm through the origin:

for $i = 1, \dots, n$ do

if $y^{(i)}(\theta \cdot x^{(i)}) \leq 0$ then

$\theta = \theta + y^{(i)} x^{(i)}$

return θ

dot product

* $y^{(i)}(\theta \cdot x^{(i)}) =$ number

$\theta \cdot x^{(i)} = [\theta_1, \theta_2] \cdot [x_1^{(i)}, x_2^{(i)}]$

$= \theta_1 \times x_1^1 + \theta_2 \times x_2^1$

* $y^{(i)} x^{(i)} = y^{(1)} \cdot [x_1^{(1)}, x_2^{(1)}] =$ vector

$= [y^{(1)} x_1^1, y^{(1)} x_2^1]$

$x^{(1)} = [-1, -1] \longrightarrow y^{(1)} = 1$

$x^{(2)} = [1, 0] \longrightarrow y^{(2)} = -1$

$x^{(3)} = [-1, 1.5] \longrightarrow y^{(3)} = 1$

$\theta^{(0)} = 0$

1. number of mistakes of Perceptron alg. if it starts with $x^{(1)}$

$y^{(1)}(\theta \cdot x^{(1)}) = 1\,([0,0] \cdot x^{(1)}) = 0$ 1st mistake

update $\theta \rightarrow \theta = \theta^{(0)} + y^{(1)} x^{(1)} = [0,0] + 1\,[-1,-1] = [-1,-1]$

$y^{(2)}(\theta \cdot x^{(2)}) = -1\,([-1,-1] \cdot [1,0]) = -1\,(-1 \times 1 + -1 \times 0) = 1$ no mistake

$y^{(3)}(\theta \cdot x^{(3)}) = 1\,([-1,-1] \cdot [-1, 1.5]) = 1(-1 \times -1 + -1 \times 1.5) = -0.5$ 2nd M.

update $\theta \rightarrow \theta = \theta + y^{(3)} x^{(3)} = [-1,-1] + 1\,[-1, 1.5] = [-2, 0.5]$

Progression of the separating hyperplane: $[[-1,-1],[-2,0.5]]$

2. In this part, what are the factors that affect the number of mistakes made by the algorithm?

Iteration order

Linear Support vector Machines:

In this problem, we will investigate minimizing the training objective for a support vector Machine (with margin loss).

The training objective for the Support vector machine (with margin loss) can be seen as optimizing a balance between the average Hinge loss over the examples and a Regularization term that tries to keep the parameters small (increase the margin). This balance is set by the regularization term $\lambda > 0$. Here we only consider the case without the offset parameter θ_0.

$$\text{Training objective} = \frac{1}{m} \sum_{i=1}^{m} \left[\text{Loss}_h \left(y^{(i)} (\theta \cdot x^{(i)}) + \frac{\lambda}{2} \|\theta\|^2 \right] \right.$$

$$\text{Hinge loss} \Longrightarrow \text{Loss}_h (y \cdot (\theta \cdot x)) = \max \{ 0, 1 - y(\theta \cdot x) \}$$

$$\hat{\theta} = \text{Argmin}_\theta \left[\text{Loss}_h (y(\theta \cdot x) + \frac{\lambda}{2} \|\theta\|^2 \right]$$

1. Suppose $\text{Loss}_h (y(\hat{\theta} \cdot x)) > 0$. Express $\hat{\theta}$ in terms of x, y and λ.

$$\hat{\theta} = \text{Argmin}_\theta \left[\text{Loss}_h (y(\theta \cdot x) + \frac{\lambda}{2} \|\theta\|^2 \right]$$

The above loss can be minimized by solving for the following equation:

$$0 = \nabla_\theta \left[\text{Loss}_h (y (\theta \cdot x))\right] + \nabla_\theta \left[\frac{\lambda}{2} \|\theta\|^2 \right]$$

Given that $\text{Loss}_h (y(\hat{\theta} \cdot x)) > 0 \Longrightarrow \text{Loss}_h (y(\hat{\theta} \cdot x)) = 1 - y(\theta \cdot x)$

$\downarrow$

$$\nabla_\theta \left[\text{Loss}_h (y (\theta \cdot x))\right] = -yx$$

and

$$\nabla_\theta \left[\frac{\lambda}{2} \|\theta\|^2 \right] = \lambda \hat{\theta}$$

$$\Longrightarrow 0 = \lambda \hat{\theta} - yx$$

$$\boxed{\hat{\theta} = \frac{1}{\lambda} yx}$$

Passive - Aggressive algorithm:

The Passive - Aggressive algorithm (without offset) responds to a labeled training example (x, y) by finding θ that minimizes:

$$\frac{\lambda}{2} \| \theta - \theta^{(k)} \|^2 + \mathrm{Loss}_h (y (\theta \cdot x))$$

where $\theta^{(k)}$ is the current setting of the parameters prior to encountering (x, y)

Hinge loss $\Rightarrow \mathrm{loss}_h (y \, \theta \cdot x) = \max \{ 0, \; 1 - y \theta \cdot x \}$

We could replace the loss function with something else:

$$\theta^{(k+1)} = \theta^{(k)} + \eta \, y x$$

but the real-valued step-size parameter η is no longer equal to one. it now depends on both $\theta^{(k)}$ and the training example (x, y).

If λ is large, the step size of the algorithm (η) would be small

1. Suppose $\mathrm{Loss}_h (y \, \theta^{(k+1)} \cdot x) > 0$. Express the value of η in terms of λ in this case:

$$f(\theta) = \frac{\lambda}{2} \| \theta - \theta^{(k)} \|^2 + \mathrm{Loss}_h (y \, \theta \cdot x)$$

$$= \frac{\lambda}{2} \| \theta - \theta^{(k)} \|^2 + 1 - y \, \theta \cdot x$$

We compute the minimum by setting gradient of $f(\theta)$:

$$\nabla_\theta f = \lambda (\theta - \theta^{(k)}) - y x = 0$$

$$\theta = \theta^{(k)} + \frac{1}{\lambda} y x$$

Where $\frac{1}{\lambda}$ is just the step size, thus $\eta = \frac{1}{\lambda}$ when $\mathrm{Loss}_h (y \, \theta^{(k+1)} \cdot x) > 0$