Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 1

Given Gaussian prior over w as $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$,MAP objective for the logistic regression model $p(y_n|\mathbf{x}_n, \mathbf{w}) = \frac{1}{1+exp(-y_n\mathbf{w}^T\mathbf{x}_n)}$ can be written as:

$$\hat{\mathbf{w}}_{MAP} = \arg\min_{\mathbf{w}} \mathcal{NLL} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$= \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

where $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \ln(1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$
Setting gradient of $\mathcal{L}(\mathbf{w}) = 0$,we get

$$\nabla_{\mathbf{w}}\mathcal{L}(w) = 0$$

$$\sum_{i=1}^{N} \frac{-y_n\mathbf{x}_n exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)} + \lambda\mathbf{w} = 0$$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_n y_n \mathbf{x}_n$$

where $\alpha_n = \frac{exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{\lambda(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))}$.
Note that we can use this expression of $\mathbf{w}$ for updation of $\mathbf{w}$.Here α_n will be large for incorrect prediction at $\mathbf{x}_n$ as for incorrect prediction $y_n\mathbf{w}^T\mathbf{x}_n$ will be large negative and hence will contribute more towards updation of $\mathbf{w}$.Similarily α_n will be small for correct prediction at $\mathbf{x}_n$ as for correct prediction $y_n\mathbf{w}^T\mathbf{x}_n$ will be large positive and hence will contribute less towards updation of $\mathbf{w}$. For simplification consider the SGD update of $\mathbf{w}$ using the above expression. Select a random iϵ{1,2,....,N}.Let say the model mispredicts at $\mathbf{x}_i$ and $y_i = 1$.At this point $\mathbf{w^{t-1}}^T\mathbf{x}_i$ will be negative and α_i^{t-1} is positive.Consider the update for $\mathbf{w^{t-1}}$ to $\mathbf{w^t}$ using above expression.Then

$$\mathbf{w^t}^T\mathbf{x}_i = \alpha_i^{t-1}\mathbf{x}_i^T\mathbf{x}_i \qquad \text{which is positive as } \mathbf{x}_i^T\mathbf{x}_i \text{is always positive}$$

. Similar arguement holds when $y_n = -1$.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 2

The generative model can be written as

$$p(y=1|\mathbf{x},\theta) = \frac{p(y=1|\theta)p(\mathbf{x}|y=1,\theta)}{p(y=1|\theta)p(\mathbf{x}|y=1,\theta) + p(y=0|\theta)p(\mathbf{x}|y=0,\theta)}$$

$$p(y=1|\mathbf{x},\theta) = \frac{p(y=1|\theta)\prod_{d=1}^{D} p(x_d|y=1,\theta)}{p(y=1|\theta)\prod_{d=1}^{D} p(x_d|y=1,\theta) + p(y=0|\theta)\prod_{d=1}^{D} p(x_d|y=0,\theta)}$$

Here $\theta = (\pi, \mu_{d,0}, \mu_{d,1})$ where $d = 1, 2, ..., D$

$$p(y=1|\mathbf{x},\theta) = \frac{\pi\prod_{d=1}^{D} \mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}}{\pi\prod_{d=1}^{D} \mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d} + (1-\pi)\prod_{d=1}^{D} \mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}}$$

Dividing numerator and denominator by $(1-\pi)\prod_{d=1}^{D} \mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}$ we get,

$$p(y=1|\mathbf{x},\theta) = \frac{\frac{\pi}{1-\pi}\prod_{d=1}^{D} (\frac{\mu_{d,1}}{\mu_{d,0}})^{x_d}(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})^{1-x_d}}{\frac{\pi}{1-\pi}\prod_{d=1}^{D} (\frac{\mu_{d,1}}{\mu_{d,0}})^{x_d}(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})^{1-x_d} + 1}$$

Rewriting R.H.S

$$\frac{exp(\ln(\frac{\pi}{1-\pi}) + \sum_{d=1}^{D} x_d \ln(\frac{\mu_{d,1}}{\mu_{d,0}}) + (1-x_d)\ln(\frac{1-\mu_{d,1}}{1-\mu_{d,0}}))}{exp(\ln(\frac{\pi}{1-\pi}) + \sum_{d=1}^{D} x_d \ln(\frac{\mu_{d,1}}{\mu_{d,0}}) + (1-x_d)\ln(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})) + 1}$$

$$p(y=1|\mathbf{x},\theta) = \frac{exp(b + \sum_{d=1}^{D} x_d \ln(\frac{\mu_{d,1}(1-\mu_{d,0})}{\mu_{d,0}(1-\mu_{d,1})}))}{1 + exp(b + \sum_{d=1}^{D} x_d \ln(\frac{\mu_{d,1}(1-\mu_{d,0})}{\mu_{d,0}(1-\mu_{d,1})}))}$$

This model is same as discriminative model for logistic regression with:
$p(y=1|\mathbf{x},\mathbf{w}) = \frac{exp(\mathbf{w}^T\mathbf{x}+b)}{1+exp(\mathbf{w}^T\mathbf{x}+b)}$ and each of $w_d = \ln\frac{\mu_{d,1}(1-\mu_{d,0})}{\mu_{d,0}(1-\mu_{d,1})}$ for $d = 1, 2, ...D$ and $b = \ln(\frac{\pi}{1-\pi}) + \sum_{d=1}^{D} \ln(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})$.
Also the decision boundary that this model will learn is linear.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 3

First take the constrained least square model for some c.The Langrangian setup for it will be:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \arg\max_{\alpha} \alpha(\|\mathbf{w}\|^2 - c^2)$$

Here $\alpha \geq 0$. Solving as dual formulation for this Lagrangian $\hat{\mathbf{w}}(\hat{\alpha})$=$\arg\max_{\alpha}(\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I})^{-1}\mathbf{X}^T y$.
Also using KKT condition:

$$\hat{\alpha}(\|\hat{\mathbf{w}}\|^2 - c^2) = 0$$

Implies either $\hat{\alpha} = 0$ or $\|\hat{\mathbf{w}}\|^2 - c^2 = 0$
Now consider the the Ridge Regression Model:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Solution for $\hat{\mathbf{w}}(\lambda)$=$(\mathbf{X}^T\mathbf{X} + \frac{\lambda}{2}\mathbf{I})^{-1}\mathbf{X}^T y$
If $\hat{\alpha} = 0$ then it simply gives us an unregularized model which is not of our interest.So the other case that remains is when $\|\hat{\mathbf{w}}\|^2 - c^2 = 0$.In this case simply put $\frac{\lambda}{2} = \hat{\alpha}$ and $\|\hat{\mathbf{w}}\| = c$. Since these substitution satisfies all the above conditions,hence the two problems are exact same.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 4

$$p(\mathbf{y}|\mathbf{X},\mathbf{W}) = \prod_{n=1}^{N} \mu_{ny_n}$$

$$= \prod_{n=1}^{N} \frac{exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T \mathbf{x}_n)}$$

$$\text{loglikelihood } = \sum_{n=1}^{N} (\mathbf{w}_{y_n}^T \mathbf{x}_n - \ln \sum_{l=1}^{K} exp(\mathbf{w}_l^T \mathbf{x}_n))$$

$$\nabla_{\mathbf{w}_k} \mathcal{LL}(\mathbf{W}) = \sum_{n=1}^{N} \left(\mathbb{1}[y_n = k]\mathbf{x}_n - \frac{exp(\mathbf{w}_k^T \mathbf{x}_n)\mathbf{x}_n}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T \mathbf{x}_n)} \right)$$

Clearly,setting it to zero doesn't give closed form solution for $\mathbf{w}_k$.

$$\nabla_{\mathbf{w}_k} \mathcal{LL}(\mathbf{W}) = \sum_{n=1}^{N} (\mathbb{1}[y_n = k]\mathbf{x}_n - \mu_{nk}\mathbf{x}_n)$$

$$\nabla_{\mathbf{w}_k} \mathcal{NLL}(\mathbf{W}) = \sum_{n=1}^{N} (\mu_{nk}\mathbf{x}_n - \mathbb{1}[y_n = k]\mathbf{x}_n)$$

Updation rule using Gradient Descent for each $\mathbf{w}_k$ where $k = \{1, 2, ..., K\}$ and assuming $\eta = 1$:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - \sum_{n=1}^{N} \left(\mu_{nk}^{t-1}\mathbf{x}_n - y_{nk}\mathbf{x}_n\right) \quad \text{where } y_{nk} = \mathbb{1}[y_n = k]$$

Also the SGD update for each $\mathbf{w}_k$ where $k = \{1, 2, ..., K\}$ with a randomly chosen example $(\mathbf{x}_n, y_n)$ can be written as:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - \left(\mu_{nk}^{t-1}\mathbf{x}_n - y_{nk}\mathbf{x}_n\right)$$

Algorithm for SGD Update for each $\mathbf{w}_k$:
Step1. Initialize the weights as $\mathbf{w}_k^{(0)}$ for all $k = \{1, 2, ..., K\}$
Step2. Pick a random $n\epsilon\{1, 2, ...N\}$ and update each $\mathbf{w}_k$ for all $k = \{1, 2, ..., K\}$ as:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - \left(\mu_{nk}^{t-1}\mathbf{x}_n - y_{nk}\mathbf{x}_n\right)$$

Step3. Repeat until convergence
For the case where probablities μ_{nk} are replaced by the hard class assignments in SGD algorithm,The updation rule becomes:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - (\mathbf{x}_n - y_{nk}\mathbf{x}_n) \quad \text{where } k = \arg\max_l \{\mu_{nl}^{t-1}\}_{l=1}^{K}$$

$$\mathbf{w}_{k'}^t = \mathbf{w}_{k'}^{t-1} + (y_{nk'}\mathbf{x}_n) \quad \text{where } k' \neq k$$

Algorithm for SGD Update with hard assignment:
Step1. Initialize the weights as $\mathbf{w}_k^{(0)}$ for all $k = \{1, 2, ..., K\}$
Step2. Pick a random $n \epsilon \{1, 2, ...N\}$ and update $\mathbf{W}$ as:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - (\mathbf{x}_n - y_{nk}\mathbf{x}_n) \quad \text{where } k = \arg\max_l \{\mu_{nl}^{t-1}\}_{l=1}^K$$

$$\mathbf{w}_{k'}^t = \mathbf{w}_{k'}^{t-1} + (y_{nk'}\mathbf{x}_n) \quad \text{where } k' \neq k$$

Step3. Repeat until convergence
Note that while updating weights with hard class assignment the contribution of the example with correct prediction is zero whereas in the soft class assignment case contribution of correctly predicted example in the weight update is non-zero.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 5

We can assume the two convex hulls to be coming from positive class and negative class.Now they are linearly separable if there exists a vector $\mathbf{w}$ and a scalar b such that $\mathbf{w}^T\mathbf{x}_n + b > 0$ for $n = 1, 2, ..., N$ and $\mathbf{w}^T\mathbf{y}_m + b < 0$ for $m = 1, 2, ..., M$ respectively.

$$\mathbf{C}_+(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

$$= \mathbf{w}^T\sum_{n=1}^{N}\alpha_n\mathbf{x}_n + b$$

$$= \sum_{n=1}^{N}\alpha_n(\mathbf{w}^T\mathbf{x}_n + b) \text{ as } \sum_{n=1}^{N}\alpha_n = 1$$

Hence $\mathbf{C}_+(\mathbf{x}) > 0$ as $\alpha_n \geq 0$ and $\mathbf{w}^T\mathbf{x}_n + b > 0$

Similarily,

$$\mathbf{C}_-(\mathbf{y}) = \sum_{m=1}^{M}\beta_m(\mathbf{w}^T\mathbf{y}_m + b)$$

and Hence $\mathbf{C}_-(\mathbf{y}) < 0$ as $\beta_m \geq 0$ and $\mathbf{w}^T\mathbf{y}_m + b < 0$

Claim: If they are linearly separable then they don't intersect.

Proof: Assume that they intersect then there must be a point $\mathbf{z}$ common to the two convex hulls.At this point both $\mathbf{C}_+(\mathbf{z}) > 0$ and $\mathbf{C}_-(\mathbf{z}) < 0$ should be true which is impossible.Hence contradiction.So they can't intersect.

Claim: If they don't intersect then they are linearly separable.

Proof: Since the convex hulls don't intersect so we can always find vector $\mathbf{w}$ and a scalar b such that $\mathbf{w}^T\mathbf{x}_n + b > 0$ for $n = 1, 2, ..., N$ and $\mathbf{w}^T\mathbf{y}_m + b < 0$ for $m = 1, 2, ..., M$ respectively and there won't exist any such point violating the conditions for $\mathbf{C}_+(\mathbf{x})$ and $\mathbf{C}_-(\mathbf{y})$ in their respective convex hulls as they don't intersect.Hence they are linearly separable.