CS 5824/ECE 5424: Advanced Machine Learning

Homework Assignment 1: Section A

Ankit Parekh
MS (Non-Thesis) Computer Engineering
ECE Department
ankitparekh@vt.edu

Problem 1: Question 1)

If $x_1, x_2, \dots, x_N$ are N observations sampled from a Bernoulli distribution with parameter p, the probability mass function for each x_i would be:

$$f(x_i : p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ and 1 & $0 < p < 1$.

Likelihood Function $L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \dots p^{x_n}(1-p)^{1-x_n}$

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

To find maximum liklihood, one can maximize $\ln L(p)$ instead of $L(p)$, since $y = \ln(x)$ is a naturally increasing function.

So, we can find p at which ~~ln~~ $L(p)$ is maximum.
or $\log L(p)$

$$\log L(p) = \left(\sum x_i\right)\log(p) + \left(n - \sum x_i\right)\log(1-p)$$

Taking derivative and setting it to zero: $\frac{\partial \log(L(p))}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} = 0$ 1

Solving 1: multiply $p(1-p)$ on both sides:

$$\left(\sum x_i\right)(1-p) - \left(n - \sum x_i\right)p = 0$$

To indicate p as an estimate $\Rightarrow \boxed{\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}}$

$\boxed{\therefore \text{MLE of } p = \frac{\sum_{i=1}^{n} x_i}{n}}$

To show it is the maxima:

$$\frac{\partial^2 \log(L(p))}{\partial p} = \underbrace{-\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{\sum_{i=1}^{n}(1-x_i)}{(1-p)^2}}_{\text{This term in negative indicating } \hat{p} \text{ is a maxima.}}$$

Problem 1: Question 2)

(1) $x_1, x_2, \dots, x_N$ are N observations sampled from a poisson distribution with probability mass function ∋

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Likelihood function is $L(\lambda; x_1 \dots x_N) = \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$

Since $y = \ln(x)$ is a naturally increasing function, we can try to find maxima of Log-liklihood to get the maxima of liklihood.

Log-Liklihood function is $l(\lambda; x_1 \dots x_N) = -N\lambda - \sum_{i=1}^{N} \ln(x_i!) + \ln(\lambda) \sum_{i=1}^{N} x_i$

We need to maximize Log-Liklihood and find the λ for the derivative is 0.

$$\frac{d}{d\lambda} l(\lambda; x_1 \dots x_N) = -N + \frac{1}{\lambda} \sum_{i=1}^{N} x_i = 0$$

$\therefore \boxed{\lambda = \frac{1}{N} \sum_{i=1}^{N} x_i}$ MLE is $\hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} x_i$ which is the mean of the observations.

2) By definition of expectation: $E(Y) = \sum_{y \in Y} y P(Y=y)$

For Poisson Distribution: $E(Y) = \sum_{y \geq 0} y \cdot \frac{\lambda^y e^{-\lambda}}{y!}$ ... (I)

Solving (I): $E(Y) = \lambda e^{-\lambda} \sum_{y \geq 1} \frac{1}{(y-1)!} \lambda^{y-1}$ as $y=0$ term variables

$= \lambda e^{-\lambda} \sum_{z \geq 0} \frac{\lambda^z}{z!}$ putting $z = y-1$

$= \lambda e^{-\lambda} e^{\lambda}$ Taylor Series Expansion for e^{λ}

$\boxed{E(Y) = \lambda}$ $\therefore$ Expectation of Poisson Distribution

Problem 2: Question 1) Used Python for calculations

```python
import numpy as np

vehicle_speed_dataset = {'number_of_wheels':[4, 4, 2, 8, 4, 3], 'cost':[15000, 25000, 5000, 40000, 22000, 17000]}
def standardize_feature(arr):
  feature = np.array(arr)
  mean = np.mean(feature)
  print("Mean: ", mean)
  std = np.std(feature)
  print("Standard Deviation: ", std)
  arr = [(x - mean)/std for x in arr]
  print("Standardized Values:", arr)
print("Number of Wheels : ", vehicle_speed_dataset['number_of_wheels'])
standardize_feature(vehicle_speed_dataset['number_of_wheels'])
print("Cost : ", vehicle_speed_dataset['cost'])
standardize_feature(vehicle_speed_dataset['cost'])
```

Answer (Output):

```
Number of Wheels: [4, 4, 2, 8, 4, 3]
Mean:  4.167
Standard Deviation:  1.863
Standardized Values: [-0.0894, -0.0894, -1.1628, 2.057, -0.0894, -0.626]

Cost: [15000, 25000, 5000, 40000, 22000, 17000]
Mean: 20666.667
Standard Deviation:  10687.479
Standardized Values: [-0.530, 0.405, -1.466, 1.809, 0.125, -0.343]
```

Problem 2: Question 2)

Q.2) Linear Model $h_w(x) = w_0 + w_1 x_1 + w_2 x_2$

Mean squared error is the sum of squared differences between the predicted and the true values.

Cost function $J(w)$ for least squares is given by:

Features:
x_1: no. of wheels
x_2: cost

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} (\text{predicted}_i - \text{true}_i)^2$$

where n = number of samples

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \left(h_w(x^{(i)}) - y^{(i)}\right)^2$$

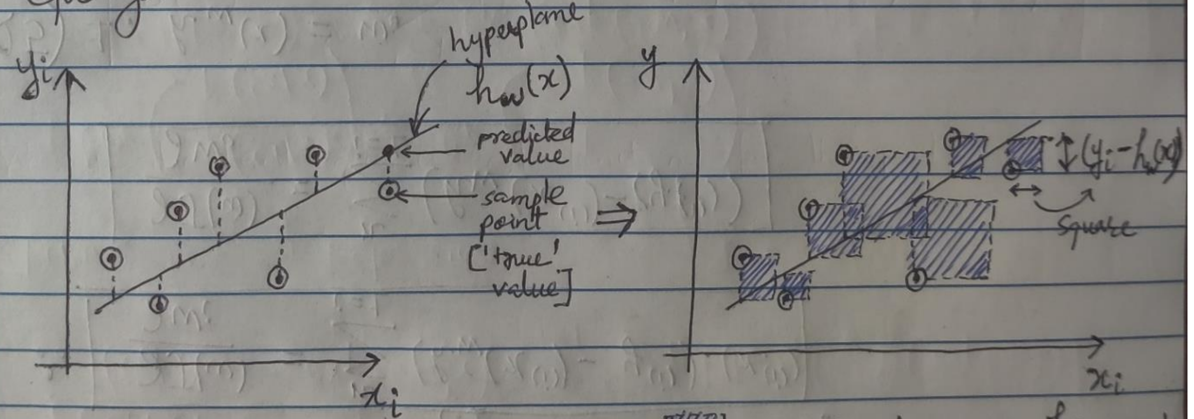
$$\boxed{J(w) = \frac{1}{n} \sum_{i=1}^{n} \left(\left(w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}\right) - \left(y^{(i)}\right)\right)^2}$$

Problem 2: Question 3)

Q.3) * We have used Least Squares Cost function for this problem to find the parameters (w) of a hyperparameter plane which best fits the linear regression dataset.

* Computing the total squared error between the associated hyperparameter plane and the true values gives an appropriate measure of how well does the line fits the dataset.

* It also takes into account that the difference between predicted values and true values might have different signs for different points, and hence squaring the individual terms so that both positive and negative values are treated equally.



* We get the area marked in blue using Least squares framework and we want to minimize this area. We try to minimize this area using parameter (w) tuning.

$$\underset{w}{\text{minimize}}\ J(w) = \underset{w}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}\left((w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}) - y^{(i)}\right)^2$$

* The residuals being normally distributed, L2 regularization performs better for linear regression.

* We want our approximation to penalize outliers more and capture the general trend. The Least Squares is also differentiable at all places and hence can be used as a cost function in gradient descent.

Problem 2: Question 4)

Q4) Partial differentiating the cost function $J(w)$: $x^{(j)}$: j^{th} sample

$$\frac{\partial}{\partial w_i} J(w) = \frac{\partial}{\partial w_i} \frac{1}{n} \sum_{j=1}^{n} \left(h_w(x^{(j)}) - y^{(j)}\right)^2$$

$$= \frac{1}{n} \cdot 2 \sum_{j=1}^{n} \left(h_w(x^{(j)}) - y^{(j)}\right) \cdot \frac{\partial}{\partial w_i} \left(h_w(x^{(j)})\right)$$

$$= \frac{1}{n} \cdot 2 \sum_{j=1}^{n} \left(h_w(x^{(j)}) - y^{(j)}\right) \cdot \frac{\partial}{\partial w_i} \left(w_0 + w_1 x_1 + w_2 x_2\right)$$

$$\therefore \boxed{\frac{\partial J(w)}{\partial w_i} = \sum_{j=1}^{n} \left(h_w(x^{(j)}) - y^{(j)}\right) \cdot \left(x_i^{(j)}\right)}$$

$$\boxed{\frac{\partial J(w)}{\partial w_0} = \sum_{j=1}^{n} \left(\left(h_w(x^{(j)})\right) - y^{(j)}\right)}$$

Problem 2: Question 5)

Q5) 1. $h_w(x) = w_0$

The model is only using a single parameter (bias weight). It does not have any flexibility (variance) and is heavily biased. It will underfit on the dataset.

2. $h_w(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$

The model has a large number of features resulting in low bias and high variance. On introducing regularization, it will most likely overfit on the dataset.

Problem 2: Question 6)

Problem 2: Q6)

1. We can increase the number of features in the model to reduce bias. If we decrease the regularization it will increase variance, and the model might fit better to the distribution. Model complexity will increase.

2. We can decrease the model complexity by decreasing the number of features and increasing regularization. This will introduce more bias and reduce variance respectively which should fix the over-fitting issue.

Problem 3: Question 1)

Q1) $P(C_0)$: Probability of patient having infectious (class C_0) disease.

$P(C_0/x)$: Probability of patient having infectious (class C_0) disease given he/she exhibits observed symptoms in data vector x.

$P(x/C_0)$: Probability of patient exhibiting observed symptoms (in data vector x) given that he/she has the infectious disease (class C_0).

According to Baye's Theorem:

$$\underset{\text{[Posterior]}}{P(C_0/x)} = \frac{\overbrace{P(x/C_0)}^{\text{Likelihood}} \cdot \overbrace{P(C_0)}^{\text{Prior}}}{P(x)}$$

Problem 3: Question 2)

Q2) $P(C_0/x) = \dfrac{P(x/C_0)\,P(C_0)}{P(x)} \quad \ldots (1)$; $P(C_1/x) = \dfrac{P(x/C_1)\,P(C_1)}{P(x)} \quad \ldots (2)$

$$P(x) = P(x/C_1)\,P(C_1) + P(x/C_0)\,P(C_0) \quad \ldots (3)$$

Adding (1) & (2) and substituting (3):

We get:

$$P(C_1/x) + P(C_0/x) = 1$$

$$P(C_0/x) = 1 - P(C_1/x) \quad \ldots (\mathrm{I})$$

We are given: $P(C_1/x) = \dfrac{1}{1 + \exp(-w^T x)}$

Substituting this in (I), we get:

$$\boxed{P(C_0/x) = 1 - \frac{1}{1+\exp(-w^Tx)} = \frac{\exp(-w^Tx)}{1+\exp(-w^Tx)}}$$

$$\frac{P(C_1/x)}{P(C_0/x)} = \left(\frac{1}{1+\exp(-w^Tx)}\right) \Big/ \left(\frac{\exp(-w^Tx)}{1+\exp(-w^Tx)}\right)$$

$$\frac{P(C_1/x)}{P(C_0/x)} = \frac{1}{\exp(-w^Tx)}$$

Taking ~~log~~ on both sides:

$$\ln\left(\frac{P(C_1/x)}{P(C_2/x)}\right) = \ln\left(\frac{1}{\exp(-w^Tx)}\right)$$

$$= -\ln\left(\exp(-w^Tx)\right)$$

$$= -(-w^Tx)$$

$$\boxed{\ln\left(\frac{P(C_1/x)}{P(C_2/x)}\right) = w^Tx}$$

Problem 3: Question 3)

Q3) 1) Sigmoid function is mapping range of y_i values into a $[0,1]$ interval.

Class Mapping: $\sigma(y_i) \geq 0.5 \Rightarrow$ Class 1
$\sigma(y_i) < 0.5 \Rightarrow$ Class 2 $\Big\}$ & $c_i \in \{0,1\}$

$y_i = W^T \cancel{x}_i + w_0$ } Logistic Regression for Binary Classification

$$\text{P.M.F. for } \sigma(y_i) \Rightarrow f(x)^{c_i}(1-f(x))^{c_i}$$

$$\text{Likelihood Function} = \prod_{i=1}^{N} f(x)^{c_i}(1-f(x))^{1-c_i}$$

$$\boxed{\text{Likelihood Function} = L(w, w_0) = \prod_{i=1}^{N} \sigma(y_i)^{c_i}(1-\sigma(y_i))^{1-c_i}}$$

2) Taking log of likelihood function:

$$\log(L(w,w_0)) = \log \prod_{i=1}^{N} \sigma(y_i)^{c_i}(1-\sigma(y_i))^{1-c_i}$$

$$= \sum_{i=1}^{N} c_i \log(\sigma(y_i)) + \sum_{i=1}^{N}(1-c_i)\log(1-\sigma(y_i))$$

$$= \sum_{i=1}^{N} c_i \log\left[\frac{\sigma(y_i)}{1-\sigma(y_i)}\right] + \sum_{i=1}^{N} \log(1-\sigma(y_i)) \quad \ldots (\mathrm{I})$$

$$\frac{\sigma(y_i)}{1-\sigma(y_i)} = \frac{1/1+\exp(-y_i)}{1 - 1/1+\exp(-y_i)} = \frac{1}{\exp(-y_i)} \Rightarrow \log\left[\frac{\sigma(y_i)}{1-\sigma(y_i)}\right] = -\log(\exp(-y_i)) = y_i \quad \ldots (\mathrm{II})$$

Using (II) in (I): $\log(L(w,w_0)) = \sum_{i=1}^{N} c_i y_i + \sum_{i=1}^{N} \log(1-\sigma(y_i))$

$$\log(L(w, w_0)) = \sum_{i=1}^{N} c_i y_i - \sum_{i=1}^{N} \log\left(\frac{1}{1-\sigma(y_i)}\right)$$

$$= \sum_{i=1}^{N} c_i y_i - \sum_{i=1}^{N} \log\left(\frac{1}{1 - \frac{1}{1+\exp(-y_i)}}\right)$$

$$= \sum_{i=1}^{N} c_i y_i - \sum_{i=1}^{N} \log\left(\frac{1+\exp(-y_i)}{\exp(-y_i)}\right)$$

$$= \sum_{i=1}^{N} c_i y_i - \sum_{i=1}^{N} \log\left[\left(1+\exp(-y_i)\right)\left(\exp(y_i)\right)\right]$$

$$= \sum_{i=1}^{N} c_i y_i - \sum_{i=1}^{N} \log\left[1+\exp(y_i)\right]$$

Putting $y_i = w^T x_i + w_0$

$$\boxed{\log(L(w, w_0)) = \sum_{i=1}^{N} c_i (w^T x_i + w_0) - \log\left[1 + \exp(w^T x_i + w_0)\right]}$$

3) i) $$\frac{\partial}{\partial w} \log(L(w, w_0)) = \sum_{i=1}^{N} c_i x_i - \frac{1}{1+\exp(w^T x_i + x_0)} \exp(w^T x_i + w_0).(x_i)$$

$$\Rightarrow \frac{\partial \log}{\partial w}(L(w, w_0)) = \sum_{i=1}^{N} c_i x_i - \frac{x_i \exp(w^T x_i + x_0)}{1+\exp(w^T x_i + x_0)}$$

$$\Rightarrow \boxed{\frac{\partial \log}{\partial w}(L(w, w_0)) = \sum_{i=1}^{N} x_i \left[c_i - \frac{\exp(w^T x_i + w_0)}{1+\exp(w^T x_i + w_0)}\right]} \quad - (1)$$

ii) $$\frac{\partial \log(L(w, w_0))}{\partial w_0} = \sum_{i=1}^{N} c_i - \frac{1}{1+\exp(w^T x_i + x_0)} . \exp(w^T x_i + x_0)$$

$$\Rightarrow \boxed{\frac{\partial \log(L(w, w_0))}{\partial w_0} = \sum_{i=1}^{N} c_i - \frac{\exp(w^T x_i + x_0)}{1+\exp(w^T x_i + x_0)}}$$