Anubha Bhargava – Homework 1, Machine Learning

1. Part 1, a and b:

Anubha Bhargava

HW # 1

Part 1

iid Bernoulli random variables

$p(x_i = 1 | \pi) = \pi$, π is unknown

$0 \le \pi \le 1$

$P(x=0|\pi) = 1-\pi$

1) a) N observations $(x_1 \ldots x_N)$ $x_i \in \{0,1\}$

$$p(x_1 \ldots x_n | \theta) = \prod_{i=1}^{n} p(x_i | \theta)$$

$$E[x] = \mu \quad var[x] = \mu(1-\mu)$$

$$\prod_{n=1}^{N} p(x_1 \ldots x_N | \mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\prod_{n=1}^{N} p(x_1 \ldots x_N | \pi) = \boxed{\prod_{n=1}^{N} (\pi)^{x_n}(1-\pi)^{1-x_n}}$$

b) Take the log of the probability distribution $p(x_1 \ldots x_N | \pi)$

$$\ln(p(x_1 \ldots x_N|\pi)) = \sum_{n=1}^{N} \ln p(x_n|\pi) = \ln(\pi^{x_n}) + \ln\left(1-\pi^{(1-x_n)}\right)$$

$$= \frac{d}{dx}\left(\sum_{n=1}^{N} x_n \ln \pi + (1-x_n)\ln(1-\pi)\right)$$

$$= \sum_{n=1}^{N} \frac{x_n}{\pi} + \sum_{n=1}^{N} \frac{(1-x_n)}{(\pi-1)} = \left(\frac{1}{\pi}\right)\sum_{n=1}^{N} x_n + \frac{1}{(\pi-1)}\sum_{n=1}^{N}(1-x_n)$$

$$= \frac{1}{\pi}\sum_{n=1}^{N} x_n + \frac{1}{\pi-1}\left(N - \sum_{n=1}^{N} x_n\right)$$

$$= \sum_{n=1}^{N} \frac{x_n}{\pi} + \left(\frac{N}{\pi-1}\right) - \sum_{n=1}^{N} \frac{x_n}{(\pi-1)}$$

$$\sum_{n=1}^{N} \frac{x_n(\pi-1) - x_n\pi}{\cancel{\pi(\pi-1)}} + \frac{N\pi}{\cancel{(\pi-1)}\pi} = 0$$

$$\sum_{n=1}^{N} \cancel{x_n\pi} - \cancel{x_n\pi} - x_n + N\pi$$

$$\sum_{n=1}^{N} -x_n + N\pi = 0$$

$$\sum_{n=1}^{N} x_n = -N\pi$$

$$\boxed{\hat{\pi}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n}$$

c) The maximum likelihood estimator is equal to π itself. This is also equal to the expected value/mean, which is the average value of the function under the probability distribution. This makes intuitive sense because the maximum likelihood should occur the most often.

1. Part 2, a and b

Part 2

N observations $(x_1 \ldots x_N)$ $x_i \in \{0, 1, 2 \ldots\}$

Poisson λ

a) $\boxed{\prod_{i=1}^{N} \frac{(\lambda)^{x_n} e^{-\lambda}}{(x_n)!}} = \prod_{i=1}^{N} p(x_1 \ldots x_N | \lambda)$

b) $\hat{\lambda}_{ML}: \nabla_{\lambda}\left(\sum_{i=1}^{N} \ln(p(x_n|\lambda))\right) = (x_n \ln(\lambda) - \lambda \cancel{\ln(e)}) - \cancel{\ln(x_n!)})$

0

$$\left(\sum_{i=1}^{N} \frac{x_n}{\lambda} - \sum_{i=1}^{N} 1\right) = 0$$

$$\left(\sum_{i=1}^{N} \frac{x_n}{\lambda} - N\right) = 0$$

$$\frac{\sum_{i=1}^{N} x_n}{\lambda} = N \quad \boxed{\frac{1}{N}\sum_{i=1}^{N} x_n = \hat{\lambda}_{ML}}$$

c) The maximum likelihood estimator is equal to the mean. The maximum likelihood should occur the most often, so this makes intuitive sense.

2.

② N non-negative integer-valued observations $(x_1, \ldots x_N)$

iid Poisson random variables with λ

gamma prior distribution on λ, $\lambda \sim Gam(\lambda | a, b)$

$$Gam(\lambda | a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

a) use Bayes rule to derive the posterior distribution of λ

posterior = prior • likelihood

$$p(\lambda | \{x_1 \ldots x_N\}) \propto \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \cdot \prod_{i=1}^{N} \frac{(\lambda)^{x_n} e^{-\lambda}}{(x_n)!}$$

$$\propto \lambda^{a-1} e^{-b\lambda} \prod_{i=1}^{N} (\lambda)^{x_n} e^{-\lambda}$$

note $\prod \lambda^{x_0} e^{-\lambda} = \lambda^{x_0} e^{-\lambda} \cdot \lambda^{x_1} e^{-\lambda} \ldots \lambda^{x_n} e^{-\lambda} = \lambda^{x_0 + x_1 + x_n} \cdot e^{-\lambda N}$

$$= \lambda^{a-1+\sum_{i=1}^{N} x_n} e^{-b\lambda} e^{-\lambda N} = \boxed{\lambda^{a-1+\sum_{i=1}^{N} x_n} e^{-\lambda(b+N)}}$$

This is a gamma distribution $\rightarrow Gam(\lambda | \underbrace{a + \sum_{i=1}^{N} x_i}_{\alpha}, \underbrace{b + N}_{\beta})$

b) ~~mean~~ variance of λ under this posterior

$$\boxed{E[\lambda] = \frac{\alpha}{\beta}} \qquad \boxed{Var[\lambda] = \frac{\alpha}{\beta^2}}$$

$$E[\lambda] = \frac{\sum_{i=1}^{N} x_i + a}{b+N} \qquad var[\lambda] = \frac{a + \sum_{i=1}^{N} x_i}{(b+N)^2}$$

The mean would equal the solution to part 2 of problem 1 if $a = 0$ and $b = 0$.

3. <u>Part 1</u>

a)

Label	Weights
Intercept Term	23.471
Number of Cylinders	-0.564

Displacement	0.994
Horsepower	0.070
Weight	-6.023
Acceleration	0.206
Model Year	2.776

A positive sign of the weight means there is a positive correlation with the output. A negative sign means there is a negative correlation. The miles per gallon (y) increases with displacement, horsepower, intercept term, acceleration and model year. Additionally, the miles per gallon decreases with the number of cylinders and weight. The miles per gallon of the car altered the most with weight and model year. The horsepower of the car slightly increased the miles per gallon, but not by much.

b) Mean = 2.681968, Standard Deviation = 0.478090

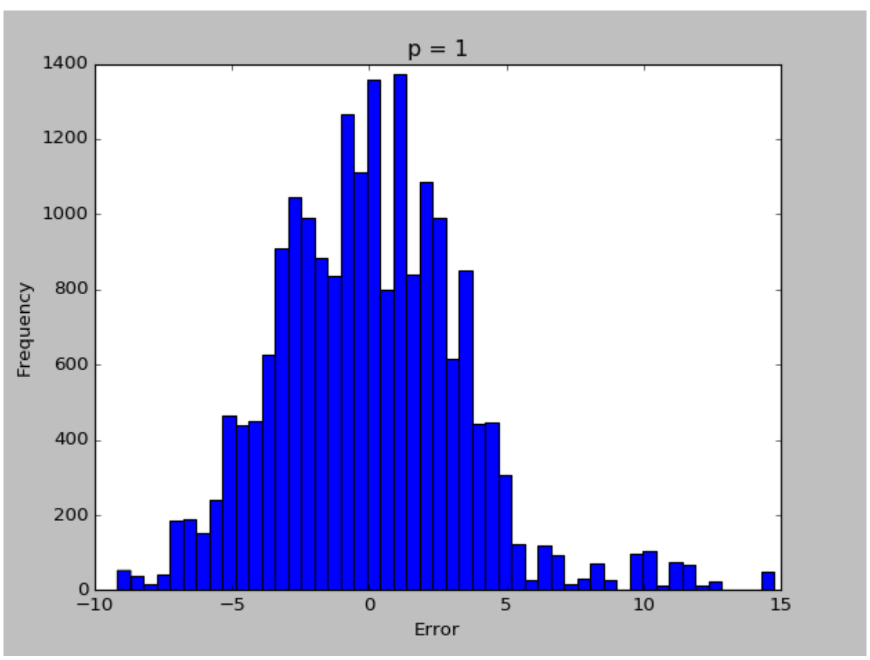
Part 2

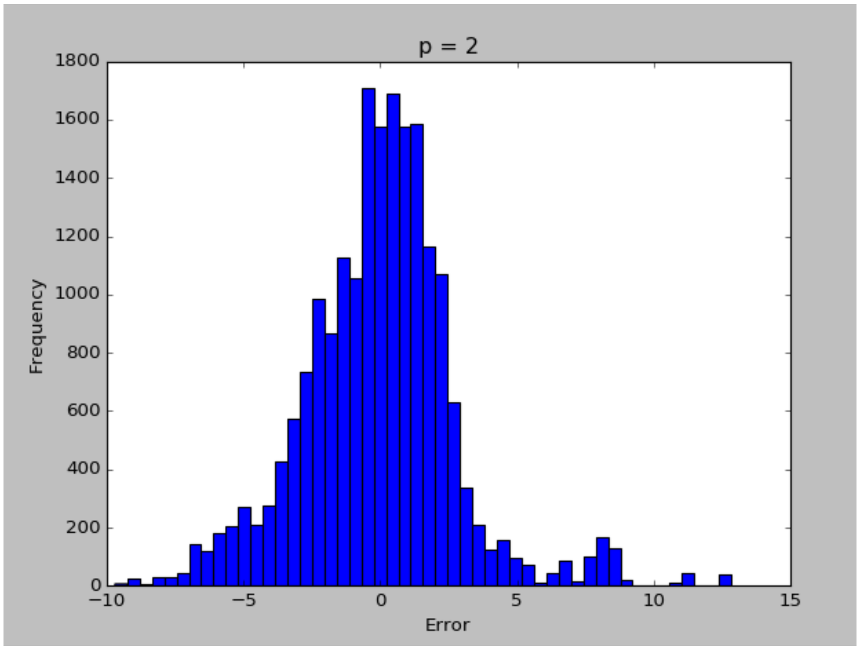
a)

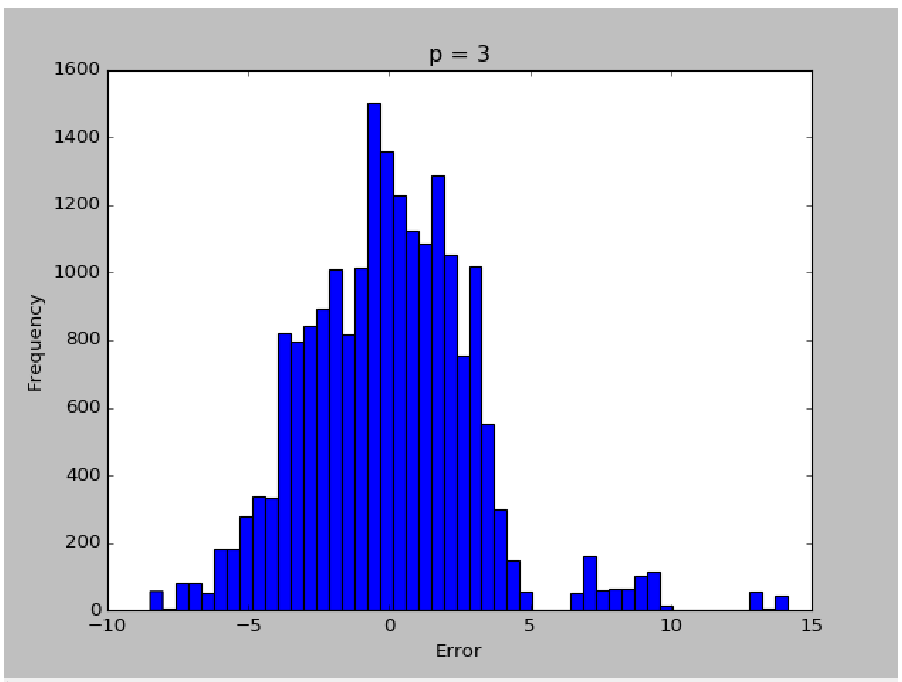
P	Mean	Standard Deviation
1	3.41602	0.673024
2	**2.75676**	**0.624128**
3	2.94815	0.637418
4	2.97078	0.631708

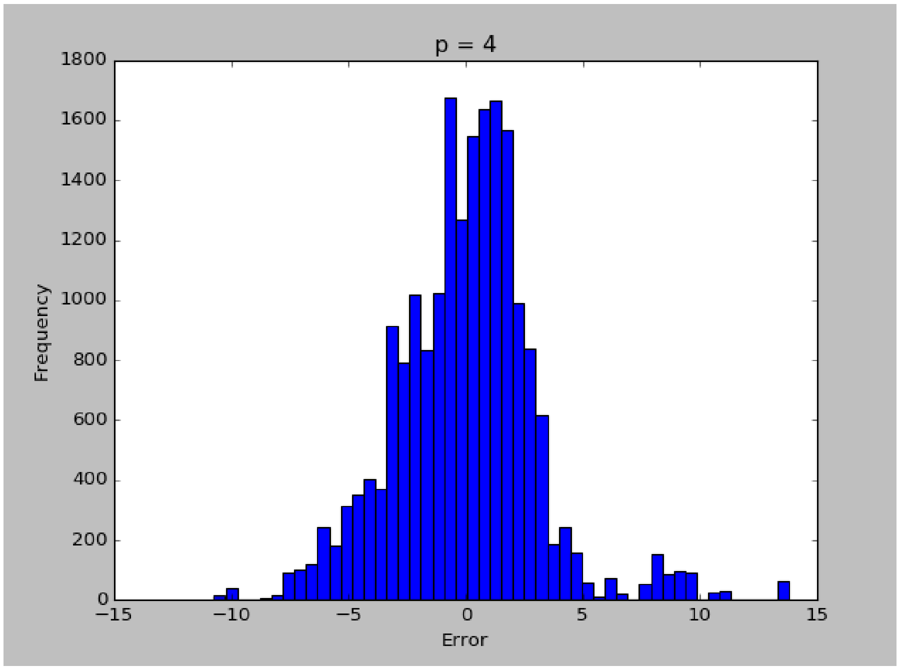
When the value of p equals 2, it has the lowest mean and standard deviation. Thus, p=2 is the best because we want the mean and standard deviation of the root mean square error to be low.

b) Plot a histogram of these errors for each p.




p = 2
Frequency
Error
1800
1600
1400
1200
1000
800
600
400
200
0
−10
−5
0
5
10
15


p = 3
Frequency
Error
1600
1400
1200
1000
800
600
400
200
0
−10
−5
0
5
10
15



c) I calculated the maximum likelihood values for the mean and variance using the following formulas: $\frac{1}{n}\sum_{i=1}^{n}(x_i)$ and $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{ML})^2$. The formula for maximum likelihood used is: $w_{ML} = (X^TX)^{-1}X^Ty$
In order to calculate the log likelihood, the following formula was used: $\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
The value for p=2 is the highest for the log likelihood, which agrees with the conclusion of Part 2(a). For higher values for log likelihood, the variance was smaller. Also, we assume that this is a normal distribution.

P	Mean	Variance	Log Likelihood
1	-0.035	11.825	-43081.722
2	**-0.025**	**7.982**	**-39150.868**
3	0.016	9.337	-40718.939
4	-0.001	9.224	-40597.294