PROBLEM 1

1. **Maximum likelihood estimate for lambda**

Referred from CASE -2 of http://www2.imm.dtu.dk/courses/02711/lecture3.pdf

$X \sim \frac{\lambda^x e^{-\lambda}}{x!}$ for some real valued $\lambda > 0$

$\Rightarrow \lambda_{MLE} = \underset{\lambda}{\arg\max}\ P(D|\lambda)$ → Maximizing likelihood function

$P(D|\lambda) = \prod_{i=1}^{m} \frac{\lambda^{x^{(m)}} e^{-\lambda}}{x^{(m)}!}$ → MAXIMIZED s.t $\lambda > 0$

$\rightarrow \max_{\lambda} \prod_{i=1}^{m} \frac{\lambda^{x^{(m)}} e^{-\lambda}}{x^{(m)}!} \Rightarrow L(\lambda, \mu) = \frac{\lambda^{\sum_i^m x^{(m)}} \cdot e^{-M\lambda}}{\prod x^{(m)}!} - \mu(-\lambda)$

s.t $-\lambda < 0$

Applying KKT conditions gives us.

STATIONARITY : $\frac{\partial}{\partial \lambda}\left(\frac{\lambda^{\sum x^{(m)}} \cdot e^{-M\lambda}}{\prod (x^{(m)}!)} + \mu\lambda\right) = 0$ i.e $\frac{\partial L}{\partial \lambda} = 0$.

COMPLEMENTARY SLACKNESS : $\mu(\lambda) = 0$

PRIMAL FEASIBILITY : $-\lambda < 0$.

DUAL FEASIBILITY : $\mu \geq 0$.

Two cases arise inorder to satisfy this constraint.

① $\mu = 0$ ② $\mu \neq 0$.

CASE-1 : $\mu = 0$.

[TWO COMPLEMENTARY CONDITIONS]

$$L(\lambda) \rightarrow \max_{\lambda} \frac{\lambda^{\sum x^{(m)}} \cdot e^{-M\lambda}}{\prod x^{(m)}!}$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \frac{\partial}{\partial \lambda}\left(\frac{\lambda^{\sum x^{(m)}} \cdot e^{-M\lambda}}{\prod x^{(m)}!} \right) = 0$$

$$\Rightarrow \frac{\sum x^{(m)} \cdot \lambda^{\left(\sum x^{(m)} - 1\right)} \cdot e^{-M\lambda} + \lambda^{\sum x^{(m)}} \cdot (-M) e^{-M\lambda}}{\prod x^{(m)}!} = 0$$

$$\Rightarrow \sum x^{(m)} - M\lambda = 0 \Rightarrow \lambda = \frac{\sum x^{(m)}}{M} \quad L > 0$$

[for this λ]

CASE-2 : $\mu > 0$, $\lambda = 0$.

$\lambda = 0 \rightarrow$ doesn't satisfy the primal feasibility condition $(\lambda > 0)$ of the considered KKT conditions.

$\therefore$ Optimal Solution occurs when $\lambda = \frac{\sum x^{(m)}}{M}$

2) Prior probability $P(\lambda) = \frac{1}{5} \max\left\{1 - \frac{\lambda}{10}, 0\right\}$ is introduced

[MAP ESTIMATE]

$$P(D|\lambda) = \frac{\lambda^{\sum_i^M x^{(m)}} \cdot e^{-M\lambda}}{\prod(x^{(m)}!)} \Rightarrow P(\lambda|D) = P(D|\lambda)\,P(\lambda)$$

s.t $\lambda > 0$, $\lambda < 10$

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}}\ P(\lambda|D) \Rightarrow \max_{\lambda} \frac{\lambda^{\sum_i^M x^{(m)}} \cdot e^{-M\lambda}}{\prod(x^{(m)}!)} \cdot \left(\frac{1}{5}\left(1 - \frac{\lambda}{10}\right)\right)$$

s.t $-\lambda < 0$, $\lambda - 10 < 0$

Let $\sum_m^M x^{(m)} = S$, $\prod_m (x^{(m)}!) = P$

$$L(\lambda, \mu_1, \mu_2) = \frac{\lambda^S \cdot e^{-M\lambda}}{P} \cdot \left(\frac{10-\lambda}{50}\right) - \mu_1(-\lambda) - \mu_2(\lambda - 10)$$

Applying KKT conditions.

STATIONARITY : $\frac{\partial L}{\partial \lambda} = 0$

PRIMAL FEASIBILITY : $\lambda > 0$. $\lambda < 10$

DUAL FEASIBILITY : $\mu_1, \mu_2 \geq 0$

COMPLEMENTARY SLACKNESS : $\mu_1(\lambda) = 0$, $\mu_2(10 - \lambda) = 0$

Two conditions leads to 4 cases

1. $\mu_1 = 0$, $\mu_2 = 0$
2. $\mu_1 \neq 0$, $\mu_2 = 0$
3. $\mu_1 = 0$, $\mu_2 \neq 0$
4. $\mu_1 \neq 0$, $\mu_2 \neq 0$

CASE-1 : $\mu_1 = 0 . \ \mu_2 = 0 .$

By stationarity condition

$$\frac{\partial L}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(\frac{1}{50P}\left(\lambda^s \cdot e^{-M\lambda}(10-\lambda)\right)\right) = 0 .$$

$$\lambda^s\left(\frac{\partial}{\partial \lambda}\left(e^{-M\lambda}(10-\lambda)\right)\right) + e^{-M\lambda}(10-\lambda)\, s \cdot \lambda^{s-1} = 0 .$$

$$= \lambda^s\left(e^{-M\lambda}(-1) + (10-\lambda)(-M)e^{-M\lambda}\right) + e^{-M\lambda}(10-\lambda)\, s \cdot \lambda^{s-1} = 0 .$$

$$= \lambda\left(M\lambda - 10M - 1\right) + s(10-\lambda) = 0 .$$

$$\Rightarrow M\lambda^2 - 10M\lambda - \lambda + 10s - \lambda s = 0 \Rightarrow M\lambda^2 - \lambda(10M+1+s) + 10s = 0 .$$

$$\lambda = \frac{(10M+s+1) \pm \sqrt{(10M+s+1)^2 - 40Ms}}{2M} .$$

① ②

Consider the term $(10M+s+1)^2 - 40Ms = (10M+s)^2 + 1 + 2(10M+s) - 40Ms$

$= (10M-s)^2 + 1 + 2(10M+s)$

$= (10M-s)^2 + 1 + 2(10M-s) + 4s$

$= (10M-s+1)^2 + 4s$ — ①

For ⊕ term

$$\lambda > \frac{(10M+s+1) + (10M-s+1)}{2M}$$

$$\lambda > \frac{20M+2}{2M} \Rightarrow \boxed{\lambda > 10}$$ → Primal feasibility not satisfied.

So, only ⊖ term exists satisfying all KKT conditions

CASE-2; $\mu_1 = 0$, $\mu_2 \neq 0 \Rightarrow \lambda = 10$.

↳ Primal feasibility not satisfied. $(\lambda < 10)$

Value of likelihood function shall also be $\Rightarrow \frac{10^5 \cdot e^{-10m}}{P} \left(\frac{1}{5}\left(1 - \frac{10}{10}\right)\right) \rightarrow \boxed{0}$

CASE-3 : $\mu_2 = 0$, $\mu_1 \neq 0 \Rightarrow \lambda = 0$.

↳ PRIMAL FEASIBILITY not satisfied $(\lambda > 0)$

Value of function is also $\boxed{0}$

CASE-4 : $\mu_1 \neq 0$, $\mu_2 \neq 0 \Rightarrow \lambda = 0$, $\lambda = 10$.

PRIMAL FEASIBILITY is not satisfied for both $(\lambda > 0, \lambda < 10)$

For both the values Likelihood function shall also be $\boxed{0}$

Thus only candidate for local maximum is

$$\lambda = \frac{(10M + S + 1) - \sqrt{(10M + S + 1)^2 - 40MS}}{2M}.$$

Satisfying all the KKT conditions

3. We generally choose conjugate priors for analysis as that leads to
    a. Posterior distributions having the same functional form as prior distributions
        i. So that posterior distribution can act as prior for the subsequent data observed
    b. Greatly simplifies Bayesian Analysis

For the choice of prior distribution $P(\lambda) = \frac{1}{5} \max\left\{1 - \frac{\lambda}{10}, 0\right\}$ — (I)

Considering our data likelihood function

$$P(D|\lambda) = \frac{\lambda^{\sum_m x^{(m)}} e^{-M\lambda}}{\prod_m (x^{(m)}!)} \quad \text{— (II)}$$

(I) is not a conjugate distribution for (II)

Thus choosing (I) doesn't result in POSTERIOR having same functional as PRIOR

As data likelihood has. product of λ & exponential in λ

A good choice of prior here should also be of the same form

Considering the [GAMMA DISTRIBUTION]

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

Thus resulting POSTERIOR shall be of the same of PRIOR

PROBLEM 2

1. **Log likelihood of the data observations**

For positive real valued r.v X, distributed according to LOG-NORMAL

$$X \sim \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \quad \text{for real valued parameters}$$

$$P(D \mid \mu, \sigma^2) = \prod_{m=1}^{M} P(x^{(m)} \mid \mu, \sigma^2) \leftarrow [\text{Likelihood function}]$$

Log likelihood function $\rightarrow \ln P(D \mid \mu, \sigma^2)$

$$\sum_{m=1}^{M} \ln P(x^{(m)} \mid \mu, \sigma^2) = \sum_{m=1}^{M} \left(-\frac{(\ln x^{(m)}-\mu)^2}{2\sigma^2} - \ln\left(x^{(m)}\sigma\sqrt{2\pi}\right)\right)$$

This Log likelihood (μ, σ^2) is not concave but

$L(\mu, \lambda)$ where $\left[\lambda = \frac{1}{\sigma}\right]$ is concave $[\sigma \neq 0]$.

$$\Rightarrow L(\mu, \lambda) = \sum_{m=1}^{M} -\left(\frac{\lambda^2}{2}(\ln x^{(m)}-\mu)^2\right) - \ln\left(\frac{x^{(m)}\sqrt{2\pi}}{\lambda}\right)$$

CONVEX

CONCAVE

CONCAVE

[SUM OF CONCAVE FUNCTIONS ← IS CONCAVE]

So, $L(\mu, \lambda)$ is concave

$$L(\mu,\lambda) = \sum_{m=1}^{M} -\left(\frac{\lambda^2}{2}(\ln x^{(m)} - \mu)^2\right) - \ln\left(\frac{x^{(m)}\sqrt{2\pi}}{\lambda}\right)$$

$$\Rightarrow \frac{\partial L}{\partial \mu} = \sum_{m=1}^{M} -\left(\lambda^2(\ln x^{(m)} - \mu)(-1)\right) = 0.$$

$$\Rightarrow \sum_{m=1}^{M} \ln x^{(m)} - M\mu = 0 \Rightarrow \mu = \frac{\sum_{m=1}^{M} \ln x^{(m)}}{M}.$$

$$\frac{\partial L}{\partial \lambda} = \sum_{m=1}^{M} -\left(\lambda(\ln x^{(m)} - \mu)^2\right) + \frac{1}{\left(\frac{\lambda^m\sqrt{2\pi}}{x}\right)}\left(+\frac{x^{(m)}\sqrt{2\pi}}{\lambda^2}\right)$$

$$= \sum_{m=1}^{M}\left(-\lambda(\ln x^{(m)} - \mu)^2 + \frac{1}{\lambda}\right) = 0$$

$$\Rightarrow \frac{M}{\lambda} = \lambda \sum_{m=1}^{M} (\ln x^{(m)} - \mu)^2$$

$$\Rightarrow \frac{1}{\lambda^2} = \frac{\sum_{m=1}^{M} (\ln x^{(m)} - \mu)^2}{M} \Rightarrow \boxed{\sigma^2}$$

3) Are the maximum likelihood estimators for μ, σ biased ?

Obtained sample mean and uncorrected sample variance of μ, σ from maximum likelihood estimation are as follows.

$$\mu_{MLE} = \frac{\sum_{m=1}^{M} \ln x^{(m)}}{M} \quad , \quad \sigma^2_{MLE} = \frac{\sum_{m=1}^{M} (\ln x^{(m)} - \mu_{MLE})^2}{M}$$

Let μ, σ^2 be true mean and variances of distribution.

$$\Rightarrow E[\mu_{MLE}] = E\left[\frac{1}{M}\sum_{m=1}^{M} \ln x^{(m)}\right] \quad \text{(By linearity of expectation)}$$

$$= \frac{1}{M}\sum_{m=1}^{M} E[\ln x^{(m)}] = E[\ln X]$$

↑ Since each $x^{(m)}$ is IID rv

By LOTUS $\rightarrow$ $E[\ln x] = \int_0^\infty \ln x \, P(x)\, dx \leq \int_0^\infty x \, P(x)\, dx$

$$E[\mu_{MLE}] \leq \mu \leftarrow E[X]$$

[BIASED ESTIMATOR OF MEAN]

A simple intuition by looking at the estimates they are like estimates for guassian with rv's $\ln x_1, \ln x_2, \ln x_3 \ldots \ln x_n$

In general σ^2 for guassian is biased. Here variance can be BIASED

Now checking for σ^2_{MLE}

$$\sigma^2 = E\left[\frac{1}{n}\sum_1 (x-\mu)^2\right]$$

[TRUE VARIANCE] [TRUE MEAN]

$$E\left[\sigma^2_{MLE}\right] = E\left[\frac{1}{M}\sum_{i=1}^{M}\left(\ln x^{(m)} - \mu_{MLE}\right)^2\right]$$

$$= E\left[\frac{1}{M}\sum_{i=1}^{M}\left(\ln x^{(m)} - x^{(m)} + x^{(m)} - \mu + \mu - \mu_{MLE}\right)^2\right] \quad \text{By linearity}$$

$$= E\left[\frac{1}{M}\sum_1\left[\left(\ln x^{(m)} - x^{(m)}\right)^2 + \left(x^{(m)} - \mu\right)^2 + \left(\mu - \mu_{MLE}\right)^2\right.\right.$$

$$\left.\left. + 2\left(x^{(m)} - \mu\right)\left(\ln x^{(m)} - x^{(m)}\right) + 2\left(x^{(m)} - \mu\right)\left(\mu - \mu_{MLE}\right) + 2\left(\ln x^{(m)} - x^{(m)}\right)\left(\mu - \mu_{MLE}\right)\right]\right]$$

$$= \sigma^2 + E\left[\frac{1}{M}\sum\left(\ln x^{(m)} - x^{(m)}\right)^2\right] + E\left[\left(\mu - \mu_{MLE}\right)^2\right]$$

$$+ 2E\left[\frac{1}{M}\sum \underbrace{\left(x^{(m)} - \mu\right)\left(\ln x^{(m)} - x^{(m)}\right)}_{①}\right] + 2E\left[\frac{1}{M}\sum \underbrace{\left(x^{(m)} - \mu\right)\left(\mu - \mu_{MLE}\right)}_{②}\right]$$

$$+ 2E\left[\frac{1}{M}\sum \underbrace{\left(\ln x^{(m)} - x^{(m)}\right)\left(\mu - \mu_{MLE}\right)}_{③}\right]$$

Observing terms ①, ②, ③

$$\left(x^{(m)} - \mu\right)\left(\ln x^{(m)} - x^{(m)}\right) = x^{(m)}\left(\ln x^{(m)} - x^{(m)}\right) - \mu\left(\ln x^{(m)} - x^{(m)}\right)$$

$$\left(\ln x^{(m)} - x^{(m)}\right)\left(\mu - \mu_{MLE}\right) = \mu\left(\ln x^{(m)} - x^{(m)}\right) - \mu_{MLE}\left(\ln x^{(m)} - x^{(m)}\right)$$

$$\left(x^{(m)} - \mu\right)\left(\mu - \mu_{MLE}\right) = x^{(m)}\left(\mu - \mu_{MLE}\right) - \mu\left(\mu - \mu_{MLE}\right).$$

$$\Rightarrow 2E\left[\frac{1}{M}\sum x^{(m)}\ln x^{(m)} - (x^{(m)})^2 - \mu_{MLE}\ln x^{(m)} + \mu x^{(m)} - \mu^2 + \mu\mu_{MLE}\right] + (I)$$

$$\Rightarrow 2E\left[\frac{1}{M}\sum x^{(m)}\ln x^{(m)} - (x^{(m)})^2 - \mu_{MLE}\ln x^{(m)} + \mu\mu_{MLE}\right] + (I)$$

$$\Rightarrow 2E\left[\frac{1}{M}\sum x^{(m)}\ln x^{(m)}\right] - 2E\left[\frac{1}{M}\sum (x^{(m)})^2\right] - 2E\left[\frac{1}{M}\sum \mu_{MLE}\ln x^{(m)}\right]$$

$$+ E\left[(\mu-\mu_{MLE})^2\right] + 2E\left[\frac{1}{M}\sum \mu\mu_{MLE}\right] + \sigma^2 + E\left[\frac{1}{M}\sum (\ln x^{(m)} - x^{(m)})^2\right]$$

$$\longrightarrow \quad E[\sigma^2_{MLE}] < \sigma^2 \quad [\text{TRUE-VARIANCE}]$$

Thus σ^2 is BIASED

4)

Log normal distribution insists rv (X)

to be positive, real valued

$E[X]$ shall also be POSITIVE (MEAN)

Now, choosing a GUASSIAN PRIOR over $[\mu]$ results in accounting for [-ve] values of $[\mu]$ which the data distribution is unlikely to produce

Thus Guassian Prior for μ may not be a good option

PROBLEM 3

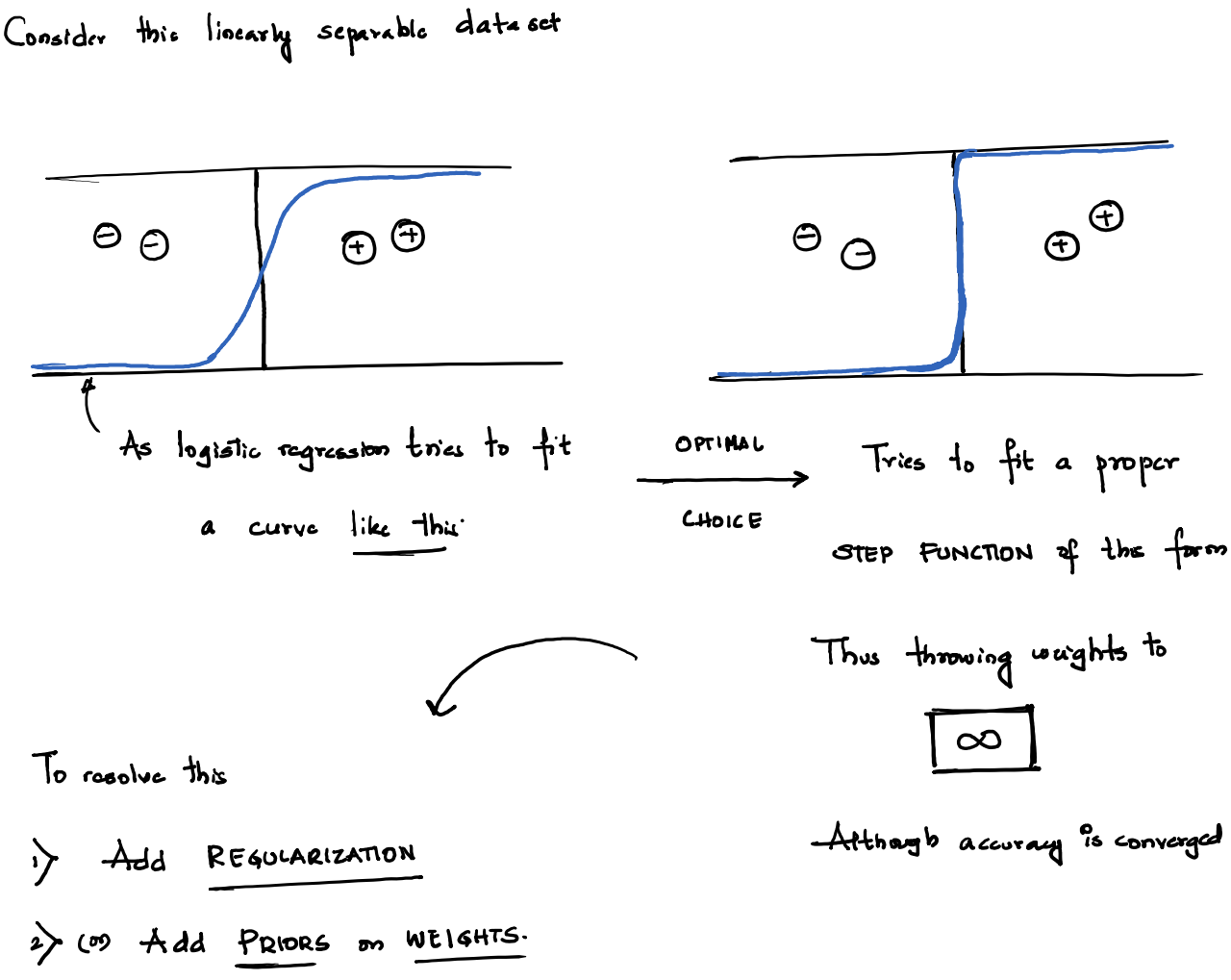
1. Fitting logistic regression model to training data no regularization involved

ACCURACIES :

TRAINING ACCURACY : 100%
VALIDATION ACCURACY : 82.69230769230769%
TEST ACCURACY : 75.0%

If there is a feature that would perfectly separate the two classes, the logistic regression model can no longer be trained. This is because the weight for that feature would not converge, because the optimal weight would be infinite

This problem of complete separation can be solved by introducing penalization of the weights or defining a prior probability distribution of weights.



2. Fitting logistic regression with L2 penalty on the weights following are the accuracies for different values of C

C	TRAINING	VALIDATION	TEST
1.00E-06	0.509615	0.615385	0.5
1.00E-05	0.509615	0.615385	0.5
0.0001	0.509615	0.615385	0.5
0.001	0.509615	0.615385	0.5
0.01	0.634615	0.711538	0.596154
0.1	0.740385	0.75	0.634615
0.5	0.826923	0.826923	0.75
1	0.826923	0.826923	0.769231
10	0.884615	0.807692	0.807692
100	0.942308	0.769231	0.826923
1000	0.990385	0.807692	0.788462
10000	1	0.807692	0.75
100000	**1**	**0.846154**	**0.75**

Best validation accuracy is obtained for C = 100000
Best weights and bias calculated for this C are

Accuracy on the test set : 75%
Bias = array([30.29822813])
W = array([[-109.3506758 , -72.03251953, 76.77956541, -60.14462796,
-28.63187635, 44.91359683, -12.00867791, 131.98871562,
-123.65074068, 10.24641001, 5.80717194, -11.5652987 ,
-28.15630323, 25.15885572, -22.72300216, 18.97688102,
61.08025052, -42.25998256, -7.59122594, -30.79315092,
42.17821559, -38.71056678, 17.50353478, -36.67590325,
26.031912 , -37.90732617, 46.35116147, -23.66595003,
14.15241658, -54.45700234, 77.7426917 , -30.22867364,
-15.10334468, 60.51284367, -62.07836708, 74.12762222,
-26.27115732, 1.99257962, -1.49141903, 27.48790619,
-56.44536433, 43.33197411, -54.61678608, 4.3863678 ,
-21.31217387, 36.24032036, -12.75947559, -88.54028366,
-69.0936089 , 73.24709733, -4.03034888, -31.76883203,
-28.13468045, -20.46467365, 7.90493342, -1.50289267,
13.38501368, -23.24556838, -55.26823679, -33.43392126]])

3. Fitting logistic regression with L1 penalty on weights, following are the accuracies

C	TRAINING	VALIDATION	TEST
1.00E-06	0.509615	0.615385	0.5
1.00E-05	0.509615	0.615385	0.5
0.0001	0.509615	0.615385	0.5
0.001	0.490385	0.384615	0.5
0.01	0.509615	0.615385	0.5
0.1	0.509615	0.615385	0.5
0.5	0.75	0.692308	0.711538
1	0.817308	0.788462	0.807692
10	0.923077	0.769231	0.788462
100	1	0.788462	0.788462
1000	1	0.769231	0.75
10000	1	0.807692	0.75
100000	**1**	**0.826923**	**0.75**

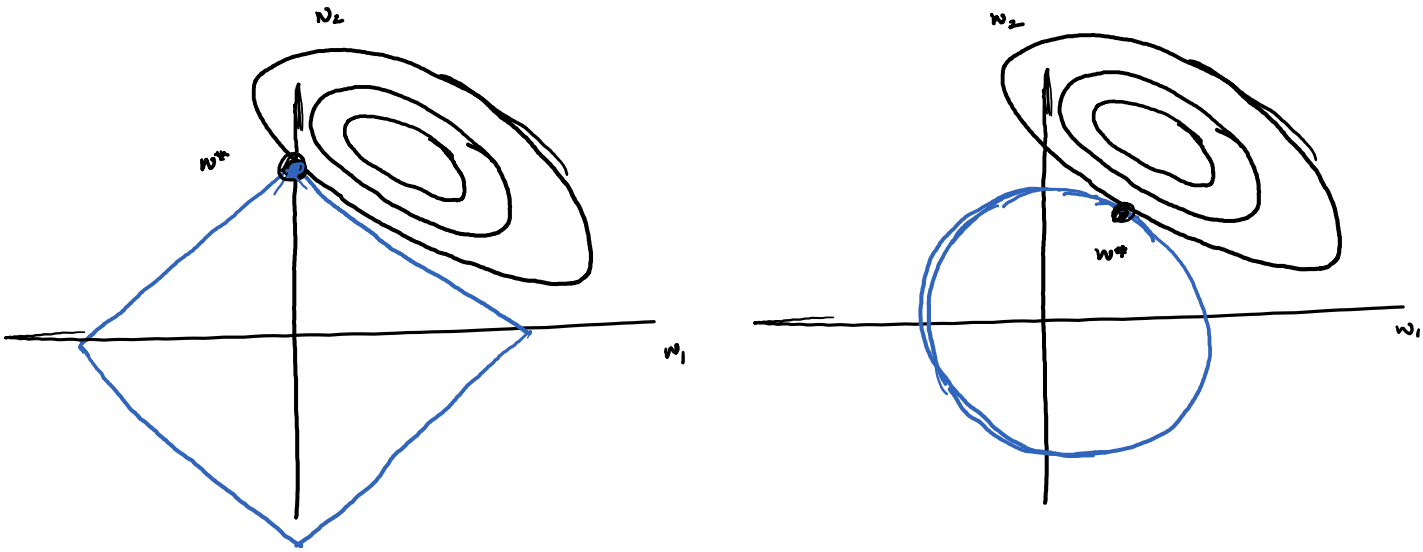
Best validation accuracy is obtained for C = 100000
Corresponding accuracy on the test set = 75%

Bias = array([43.30640427])

Weights = array([[-160.1945443 , -105.80725621, 114.78170216, -86.88794287,
-42.02491905, 66.32214868, -20.88374954, 192.12512919,
-179.13769098, 16.61122584, 4.07186831, -14.68757473,
-39.91093397, 38.52451869, -32.37626728, 24.87398742,
87.88732102, -59.82707773, -11.21993887, -45.73345418,
63.28176764, -56.05864627, 25.42596941, -52.31893448,
37.18580769, -57.09291681, 68.92534319, -34.28130793,
19.0614109 , -78.46307405, 112.76784387, -42.26988118,
-22.77675412, 89.07934674, -91.25545354, 106.80698489,
-36.60797823, 0.95900351, -1.11895256, 37.68713966,
-82.83541057, 60.53515299, -77.88782372, 6.99943983,
-29.43999232, 53.28326561, -16.30656688, -123.65912787,
-96.67104192, 107.267997 , -3.76916167, -45.78622287,
-40.56999317, -29.12157315, 8.95330377, -0.32430307,
18.73394906, -31.1382513 , -80.83139619, -48.64800047]])

4) L_1 tends to produce sparser weights.

Intuition



Formulaic Intuition

L_1: $L(w) - \lambda \|w\|_1$ $\quad\quad$ $\frac{\partial L_1}{\partial \vec{w}} : \frac{\partial L}{\partial \vec{w}} - \lambda$

L_2: $L(w) - \frac{\lambda}{2}\|w\|_2^2$ $\quad\quad$ $\frac{\partial L_2}{\partial \vec{w}} : \frac{\partial L}{\partial \vec{w}} - \lambda \vec{w}$

As weights get smaller

L_1 updates weights by same $\boxed{\lambda}$ but L_2 updates by $\boxed{\lambda w}$

Net change in $w \to \textcircled{\Delta w}$ $\uparrow$ in L_1 than L_2 as weights get smaller

So, L_1 tends to produce sparser weights than L_2

PROBLEM 4: GAUSSIAN NAÏVE BAYES

1. **Log likelihood of the Gaussian Naïve Bayes model, compute MLE for each of the parameters**

Naive Bayes makes the following conditional independence assumption

$$P\left(\mathbf{x} \mid Y=c_i\right) = \prod_{j=1}^{n} P\left(x_j \mid Y=c_i\right)$$

Guassian Naive Bayes model description :

1) Class priors : $P(Y)$

2) Conditional distributions that are guassian.

$$P\left(x_j = x_j \mid Y=c_i\right) = \mathcal{N}\left(\mu_{i,j}, \sigma_{i,j}^2\right)$$

Given a dataset with m continuous features.

$$\left\{ \left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right) \ldots \left(\mathbf{x}^{(m)}, y^{(m)}\right) \right\}.$$

Log likelihood of the data.

$$\log\left(\prod_{m=1}^{M} P\left(X = x^{(m)} \mid Y = y^{(m)}\right)\right)$$

$$\sum_{m=1}^{M} \log P\left(X = x^{(m)} \mid Y = y^{(m)}\right)$$

$$\sum_{m=1}^{M} \sum_{j=1}^{n} \log p\left(x_j = x_j^{(m)} \mid Y = y^{(m)}\right)$$

$$\underbrace{\quad\quad}_{\text{II}}$$

[ESTIMATING GUASSIAN PARAMETERS]

II $\quad \sum_{m=1}^{M} \sum_{j=1}^{n} \log p\left(x_j = x_j^{(m)} \mid Y = y^{(m)}\right)$

[INDICATOR RANDOM VARIABLE]

$$= \sum_{m=1}^{M} \sum_{j=1}^{n} \sum_{i=1}^{k} \underbrace{\log p\left(x_j = x_j^{(m)} \mid Y = c_i\right)} \cdot 1\left(y^{(m)} = c_i\right)$$

↳ Guassian (By our assumption) $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$

L

$$\underset{\mu_{ij}, \sigma_{ij}}{\operatorname{argmax}} \sum_{m=1}^{M} \sum_{j=1}^{n} \sum_{i=1}^{k} \left(-\frac{1}{2}\left(\frac{x_j^{(m)} - \mu_{ij}}{\sigma_{ij}}\right)^2 - \log\left(\sigma_{ij}\sqrt{2\pi}\right)\right) \cdot 1\left(y^{(m)} = c_i\right)$$

$$\frac{\partial L}{\partial \mu_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial \mu_{ij}} \left[\left(-\frac{1}{2}\left(\frac{x_j^{(m)} - \mu_{ij}}{\sigma_{ij}}\right)^2 - \log\left(\sigma_{ij}\sqrt{2\pi}\right)\right) \cdot 1\left(y^{(m)} = c_i\right)\right]$$

$$\sum_{m=1}^{M} \left(\frac{\mu_{ij} - x_j^{(m)}}{\sigma_{ij}}\right) \cdot 1\left(y^{(m)} = c_i\right) = 0 \Rightarrow \mu_{ij} = \frac{\sum_{m=1}^{M} x_j^{(m)} \cdot 1\left(y^{(m)} = c_i\right)}{\sum_{m=1}^{M} 1\left(y^{(m)} = c_i\right)}$$

$$\sum_{m=1}^{M} \sum_{j=1}^{n} \sum_{i=1}^{k} \left(-\frac{1}{2} \left(\frac{x_j^{(m)} - \mu_{i,j}}{\sigma_{i,j}} \right)^2 - \log\left(\sigma_{i,j} \sqrt{2\pi} \right) \right) \cdot 1\left(y^{(m)} = c_i \right)$$

$$\frac{\partial L}{\partial \sigma_{i,j}} = \sum_{m=1}^{M} \frac{\partial}{\partial \sigma_{i,j}} \left[\left(-\frac{1}{2} \left(\frac{x_j^{(m)} - \mu_{i,j}}{\sigma_{i,j}} \right)^2 - \log\left(\sigma_{i,j} \sqrt{2\pi} \right) \right) \cdot 1\left(y^{(m)} = c_i \right) \right]$$

$\rightarrow = 0$

$$= \sum_{m=1}^{M} \left(\frac{-\frac{1}{2}(-2)}{\sigma_{i,j}^3} \left(x_j^{(m)} - \mu_{i,j} \right)^2 - \frac{1}{\sigma_{i,j}} \right) \cdot 1\left(y^{(m)} = c_i \right)$$

$$= \sum_{m=1}^{M} \left(\frac{\left(x_j^{(m)} - \mu_{i,j} \right)^2}{\sigma_{i,j}^3} - \frac{1}{\sigma_{i,j}} \right) \cdot 1\left(y^{(m)} = c_i \right) = 0.$$

$$\Rightarrow \quad \sigma_{i,j}^2 = \frac{\sum_{m=1}^{M} \left(x_j^{(m)} - \mu_{i,j} \right)^2 \cdot 1\left(y^{(m)} = c_i \right)}{\sum_{m=1}^{M} 1\left(y^{(m)} = c_i \right)}$$

$\hat{\sigma}_{i,j}^2$ is a BIASED ESTIMATOR.

UNBIASED VERSION

$$\hookrightarrow \quad \sigma_{i,j}^2 = \frac{\sum_{m=1}^{M} \left(x_j^{(m)} - \mu_{i,j} \right)^2 \cdot 1\left(y^{(m)} = c_i \right)}{\sum_{m=1}^{M} 1\left(y^{(m)} = c_i \right) - 1}$$

2. Test accuracy of the GNB model : **69.23076923076923**

3. PRIOR ON NAÏVE BAYES MODEL :
   a. We generally choose a conjugate prior so that posterior also has same functional form
      i. If we choose a Gaussian Prior, it will be the conjugate distribution of the likelihood function as the corresponding posterior will be product of two exponentials of quadratic functions and hence shall be Gaussian

$$p(X|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

[LIKELIHOOD FUNCTION]

(Probability of observed data given μ)

(EXPONENTIAL OF QUADRATIC FORM.)

If we choose $p(\mu)$ by GUASSIAN → [CONJUGATE DISTRIBUTION]

POSTERIOR has same functional form as PRIOR

$$p(X|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

Likelihood function for λ

So corresponding conjugate prior → product of a power of λ & exponential of a linear function of λ.

Corresponds to a GAMMA DISTRIBUTION

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

4. I don't think Naïve Bayes is reasonable here for the following reasons
   Naïve Bayes assumes that features are conditionally independent given classes
   a. Logistic regression being a discriminative classifier without any assumptions produced higher test accuracies than the Naïve Bayes model for this dataset
      i. Logistic Regression (No penalty) - 75%
      ii. Logistic Regression (L1 penalty) - 75%
      iii. Logistic Regression (L2 penalty) - 75%

   All are higher than Naïve Bayes