PROBLEM 1

1. **Maximum likelihood estimate for lambda**

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

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

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

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

$$\hookrightarrow \max_{\lambda} \prod_{i=1}^{m} \frac{\lambda^{x^{(m)}} e^{-\lambda}}{x^{(m)}!}$$

s.t $\lambda > 0$

$$\Rightarrow L(\lambda, \mu) = \frac{\lambda^{\sum x^{(m)}} \cdot e^{-m\lambda}}{\prod x^{(m)}!} + \mu(-\lambda)$$

KKT conditions give us.

STATIONARITY : $\dfrac{\partial L}{\partial \lambda} = 0$, PRIMAL FEASIBILITY : $-\lambda < 0$

SLACKNESS : $\mu(-\lambda) = 0$. , DUAL FEASIBILITY : $\mu \geq 0$.

CASE-1 : $\mu = 0$.

[Two complementary conditions]

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

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

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

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

[for this λ]

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

Maximum of $L \rightarrow \boxed{0}$

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

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

[MAP ESTIMATE]

$$P(D|\lambda) = \frac{\lambda^{\sum_4 x^{(m)}} \cdot e^{-m\lambda}}{\prod (x^{(m)}!)} \Rightarrow P(\lambda|D) = P(D|\lambda)\, P(\lambda)$$

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

$$\Rightarrow \lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}}\; P(\lambda|D) \Rightarrow \max_{\lambda} \frac{\lambda^{\sum_4 x^{(m)}} \cdot e^{-m\lambda}}{\prod (x^{(m)}!)} \cdot \left(\frac{1}{5}\left(1 - \frac{\lambda}{10}\right)\right)$$

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

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

$$L(\lambda, \mu_1, \mu_2) = \frac{\lambda^{S} \cdot e^{-m\lambda}}{P} \cdot \left(\frac{1}{5}\left(1 - \frac{\lambda}{10}\right)\right) + \mu_1(\lambda) + \mu_2(10 - \lambda)$$

Applying KKT conditions.

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

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

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

DUAL FEASIBILITY : $\boxed{\mu_1, \mu_2 \geq 0.}$

2 Complementary conditions $\Rightarrow$ [4-cases]

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

$\Rightarrow$ Maximize $\rightarrow$ $\frac{\lambda^s \cdot e^{-m\lambda}}{P} \cdot \left(\frac{10-\lambda}{50}\right) \rightarrow$ L

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

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

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

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

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

$$\Rightarrow \lambda = \frac{(10m+1+s) \pm \sqrt{(10m+1+s)^2 - 40ms}}{2m}$$

if [+] λ turns out to be > 10

So,

$$\lambda = \frac{(10m+1+s) - \sqrt{(10m+1+s)^2 - 40ms}}{2m}$$

CASE-2 : $\mu_1 = 0$, $\lambda = 10$.

$\Rightarrow$ Value of function

$\hookrightarrow \dfrac{10^S \cdot e^{-10m}}{P} \left(\dfrac{1}{5} \left(1 - \dfrac{10}{10} \right) \right) \rightarrow \boxed{0}$

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

Value of function $\rightarrow \boxed{0}$

CASE-4 : $\lambda = 0$, $\lambda = 10$ $[\mu_1 > 0, \mu_2 > 0]$

For both the values Likelihood function shall be $\boxed{\text{zero}}$

So, Likelihood is non-zero and maximized when

$$\lambda = \frac{(10m + 1 + S) - \sqrt{(10m+1+S)^2 - 40mS}}{2m}$$

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

Here the prior considered is not conjugate distribution for likelihood and thus posterior does not have same functional form as PRIOR

Considering the likelihood function

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

[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

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

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

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

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

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

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

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

[SUM OF CONCAVE FUNCTIONS ←
IS CONCAVE]

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

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

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

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

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

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

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

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

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

Obtained sample mean and uncorrected sample variance of μ, σ from maximum likelihood estimators 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.

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

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

Since each $x^{(m)}$ is IID rv

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

$$\mu_{MLE} \leq \mu$$

[BIASED ESTIMATOR OF MEAN]

Now checking for $\underline{\sigma^2_{MLE}}$

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

[TRUE VARIANCE] [TRUE MEAN]

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

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

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

$E\left[\sigma^2_{MLE}\right] < \sigma^2$ (TRUE VARIANCE)

So it is [BIASED]

4) Observing the [LOG NORMAL DISTRIBUTION]

it is POSITIVE, REAL VALUED RANDOM VARIABLE X

True mean is always $\boxed{+ve}$

If Guassian is choosen as prior for mean

in a way account for negative values of $\boxed{\text{mean}}$ ✓

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%

   When we donot use any kind of regularizer for finding logistic regression classifier

   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

   The 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]])
```

PROBLEM 4: GUASSIAN NAÏVE BAYES

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

Naive Bayes makes the following conditional independence assumption

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

Guassian Naive Bayes model description:

1) Class priors: $P(Y)$

2) Conditional distributions that are guassian.

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

Given a dataset with m continuous features.

$$\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}) \dots (\mathbf{x}^{(m)}, y^{(m)})\}.$$

Log likelihood of the data.

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

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

$$\underbrace{\sum_{m=1}^{M} \sum_{j=1}^{n} \log P\left(X_j = x_j^{(m)} \mid Y = y^{(m)}\right)}_{\text{II}}$$

[ESTIMATING GUASSIAN PARAMETERS]

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

[INDICATOR RANDOM VARIABLE]

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

↳ Guassian (By our assumption). $\mathcal{N}\left(\mu_{ij}, \sigma_{i,j}^2\right)$

L

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

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

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

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

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

$= 0$

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

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

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

$\sigma_{i,j}^2$ is a $\underbrace{\text{BIASED ESTIMATOR.}}$

UNBIASED VERSION

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

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

$$\underbrace{P(X|\mu)}_{\text{[LIKELIHOOD FUNCTION]}} = \prod_{m=1}^{m} P(x_m|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \underbrace{\left\{ -\frac{1}{2\sigma^2} \sum_{m=1}^{N} (x_m-\mu)^2 \right\}}_{\text{(EXPONENTIAL OF QUADRATIC FORM.)}}$$

(Probability of observed data given μ)

if we choose $p(\mu)$ by GUASSIAN $\rightarrow$ [CONJUGATE DISTRIBUTION]

POSTERIOR has same functional form as PRIOR

$$P(X|\lambda) = \prod_{m=1}^{M} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{M/2} \exp \underbrace{\left\{ -\frac{\lambda}{2} \sum_{m=1}^{M} (x_m-\mu)^2 \right\}}$$

Likelihood function for λ

So corresponding conjugate prior $\rightarrow$ 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

PROBLEM SET 5

ROUGH NOT INCLUDED AS PART OF SUBMISSION

PROBLEM - 1

Non negative, integer valued random variable X

$$X \sim \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{[POISSON DISTRIBUTION]} \quad \text{for some real valued } (\lambda > 0)$$

i) Given data samples $x^{(1)}, x^{(2)} \ldots x^{(m)} \rightarrow \boxed{D}$

What is the maximum likelihood estimate for λ

$$\text{Maximum likelihood estimate } \lambda = \arg\max_{\lambda} P(D|\lambda)$$

$$= \arg\max_{\lambda} \ln P(D|\lambda) \quad \text{— (1)}$$

$$P(D|\lambda) = \prod_{i=1}^{m} \frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} \quad \text{— (2)}$$

Substituting (2) in (1) gives

$$\arg\max_{\lambda} \ln \left(\prod_{i=1}^{m} \frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} \right)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{m} \ln \left(\frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} \right)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{m} \ln\left(\frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!}\right)$$

$$= \arg\max_{\lambda} \sum_{i=1}^{m} \left(\ln\left(\lambda^{x^{(i)}} e^{-\lambda}\right) - \ln\left(x^{(i)}!\right)\right)$$

$$= \arg\max_{\lambda} \left[\sum_{i=1}^{m} \left[x^{(i)} \ln\lambda + \ln e^{-\lambda}\right]\right] - \ln\left(\prod_{i=1}^{m} x^{(i)}!\right)$$

$$= \arg\max_{\lambda} \left(\ln\lambda\left(\sum_{i=1}^{m} x^{(i)}\right) - m\lambda\right) - \ln\left(\prod_{i=1}^{m} x^{(i)}!\right)$$

$$\frac{\partial}{\partial\lambda}\left(\ln P(D|\lambda)\right) = \frac{\partial}{\partial\lambda}\left(\ln\lambda\left(\sum_{i=1}^{m} x^{(i)}\right) - m\lambda - \ln\left(\prod_{i=1}^{m} x^{(i)}!\right)\right)$$

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

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{m} x^{(i)}}{m}$$

2) Adding a prior to $P(D|\lambda)$ $\rightarrow \lambda \sim \frac{1}{5} \max\left\{1 - \frac{\lambda}{10}, 0\right\}$

Probability distribution check : $\frac{1}{5}\int_{0}^{10}\left(1 - \frac{\lambda}{10}\right) d\lambda$

$1 - \frac{\lambda}{10} > 0 \rightarrow \boxed{\lambda < 10}$

$\boxed{\lambda > 0}$

$$= 10 - \frac{\lambda^2}{20}\Big|_{0}^{10} = \frac{1}{5} \times 5 = \boxed{1}$$

[VALID PROBABILITY DISTRIBUTION]

$$\Rightarrow \quad P(D|\lambda) = \prod_{i=1}^{m} \frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} , \quad P(\lambda) = \frac{1}{5} \max\left\{1 - \frac{\lambda}{10}, 0\right\}$$

Maximum aposteriori estimate → MAP estimate $P(\lambda|D)$

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda|D) = \arg\max_{\lambda} \left(\left[\prod_{i=1}^{m} \frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} \right] \left[\frac{1}{5} \max\left\{1 - \frac{\lambda}{10}, 0\right\} \right] \right)$$

$\Downarrow$

Restrict λ in range $(0, 10)$

$\hookrightarrow P(\lambda) = \frac{1}{5}\left(1 - \frac{\lambda}{10}\right)$

$$\lambda_{MAP} = \arg\max_{\lambda} \left(\ln P(\lambda|D) \right)$$

$$= \arg\max_{\lambda} \left(\sum_{i=1}^{m} \ln \left(\frac{\lambda^{x^{(i)}} e^{-\lambda}}{x^{(i)}!} \right) + \ln\left(\frac{1}{5}\left(1 - \frac{\lambda}{10}\right) \right) \right)$$

$-\ln 50$.

$$\Rightarrow \ln P(\lambda|D) = \ln\lambda \left(\sum_{i=1}^{m} x^{(i)} \right) - m\lambda - \ln\left(\prod_{i=1}^{m} x^{(i)}! \right) + \ln(10 - \lambda)$$

$$\frac{\partial}{\partial\lambda} \ln P(\lambda|D) = \frac{1}{\lambda}\left(\sum_{i=1}^{m} x^{(i)} \right) - m + \frac{(-1)}{10 - \lambda} = 0.$$

Let $\sum_{i=1}^{m} x^{(i)} = s$

$$\Rightarrow \quad \frac{s}{\lambda} - m - \frac{1}{10-\lambda} = 0 \quad \rightarrow \quad \left(\frac{s - m\lambda}{\lambda} \right) = \frac{1}{10 - \lambda}$$

$$\left(\frac{s-m\lambda}{\lambda}\right) = \frac{1}{10-\lambda} \Rightarrow (s-m\lambda)(10-\lambda) = \lambda$$

$$\Rightarrow 10s - 10m\lambda + m\lambda^2 - \lambda s = \lambda$$

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

$$\lambda = \frac{+(10m+s+1) \pm \sqrt{(10m+s+1)^2 - 40ms}}{2m}$$

Pick a λ that is in $(0,10)$