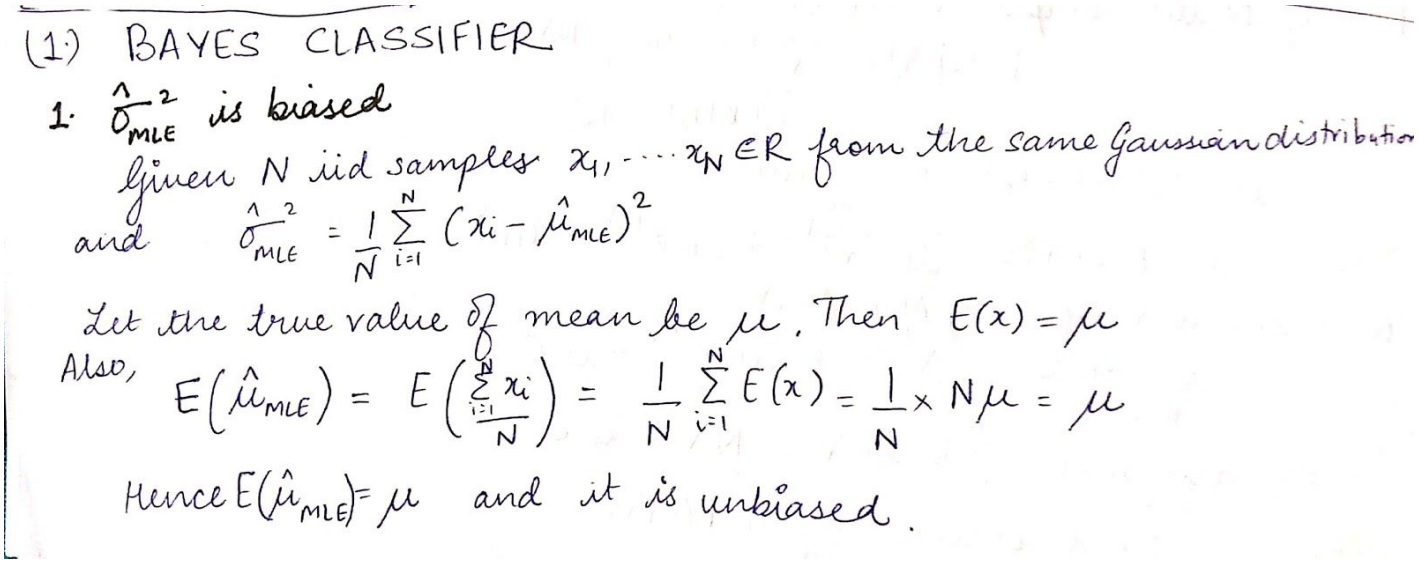
CSE-575: Statistical Machine Learning
Assignment 1

Name: Kunal Vinay Kumar Suthar
ASURite ID: 1215112535

Q1. Bayes Classifier

1. **Please prove that $\hat{\sigma}^2$ MLE is biased.**

   Solution 1:



(1.) BAYES CLASSIFIER

1. $\hat{\sigma}^2_{MLE}$ is biased

Given N iid samples $x_1, \dots x_N \in R$ from the same Gaussian distribution

and $\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$

Let the true value of mean be μ. Then $E(x) = \mu$

Also, $E(\hat{\mu}_{MLE}) = E\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = \frac{1}{N}\sum_{i=1}^{N} E(x) = \frac{1}{N} \times N\mu = \mu$

Hence $E(\hat{\mu}_{MLE}) = \mu$ and it is unbiased.

Now, solving for variance $\hat{\sigma}^2_{MLE}$

$$E\left(\hat{\sigma}^2_{MLE}\right) = E\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}_{MLE}\right)^2\right] \Rightarrow \frac{1}{N}E\left[\sum_{i=1}^{N}x_i^2 - 2\sum_{i=1}^{N}x_i\hat{\mu}_{MLE} + \sum_{i=1}^{N}\hat{\mu}^2_{MLE}\right]$$

Now, we know that $\sum_{i=1}^{N}x_i = N\hat{\mu}_{MLE}$ and $\sum_{i=1}^{N}\hat{\mu}^2_{MLE} = N\hat{\mu}^2_{MLE}$

$$\Rightarrow E\left(\hat{\sigma}^2_{MLE}\right) = \frac{1}{N}E\left(\sum_{i=1}^{N}x_i^2 - 2N\hat{\mu}^2_{MLE} + N\hat{\mu}^2_{MLE}\right)$$

$$\Rightarrow \frac{1}{N}E\left(\sum_{i=1}^{N}x_i^2 - N\hat{\mu}^2_{MLE}\right)$$

$$\Rightarrow \frac{1}{N}E\left(\sum_{i=1}^{N}x_i^2\right) - E\left(\hat{\mu}^2_{MLE}\right)$$

$$\Rightarrow E(x^2) - E\left(\hat{\mu}^2_{MLE}\right) \quad \text{— (1)}$$

Now, we know that $\sigma^2(x) = E(x^2) - (E(x))^2 = E(x^2) - \mu^2$ — (2)

Similarly, $\sigma^2(\hat{\mu}_{MLE}) = E(\hat{\mu}^2_{MLE}) - (E(\hat{\mu}_{MLE}))^2 = E(\hat{\mu}^2_{MLE}) - \mu^2$ — (3)

using (2) and (3) in (1), we get:

$$E\left(\hat{\sigma}^2_{MLE}\right) = \left(\sigma^2(x) + \mu^2\right) - \left(\sigma^2(\hat{\mu}_{MLE}) + \mu^2\right) = \sigma^2(x) - \sigma^2(\hat{\mu}_{MLE}) \quad \text{— (4)}$$

Now, $\sigma^2(\hat{\mu}_{MLE}) = VAR(\hat{\mu}_{MLE}) = VAR\left(\frac{\sum_{i=1}^{N}x_i}{N}\right) = \frac{1}{N^2}VAR\left(\sum_{i=1}^{N}x_i\right)$ — (5)

Because the samples are iid $VAR\left(\sum_{i=1}^{N}x_i\right) = \sum_{i=1}^{N}VAR(x) = \sum_{i=1}^{N}\sigma^2(x) = N\sigma^2(x)$

using the above derivation in (5), we get $\sigma^2(\hat{\mu}_{MLE}) = \frac{1}{N^2} \times N\sigma^2(x) = \frac{1}{N}\sigma^2(x)$

using the above derivation in (4), we get:

$$E\left(\hat{\sigma}^2_{MLE}\right) = \sigma^2(x) - \frac{1}{N}\sigma^2(x) = \left(\frac{N-1}{N}\right)\sigma^2(x) \neq \sigma^2(x)$$

Hence $E\left(\hat{\sigma}^2_{MLE}\right) \neq \sigma^2(x)$. Therefore it is biased.

2. If the prior distribution for mean follows μ ~ N (θ , λ), what is the MAP estimator $\hat{\mu}_{MAP}$ of μ?

Solution 2.

2. MAP estimator $\hat{\mu}_{MAP}$ of μ:

mean follows $N(\theta, \lambda)$

Posterior = $P(\mu|x) = P(x|\mu)\, P(\mu|\theta,\lambda)$

So, calculating log-likelihood:

$$\log P(\mu|x) = \log P(x|\mu) + \log P(\mu|\theta,\lambda)$$

$$\Rightarrow \log \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right) + \log \frac{1}{\sqrt{2\pi}\lambda} e^{\frac{-(\mu-\theta)^2}{2\lambda^2}}$$

$$\Rightarrow \sum_{i=1}^{N} \left(\frac{-(x_i-\mu)^2}{2\sigma^2} + \log \frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \frac{1}{\sqrt{2\pi}\lambda} - \frac{(\mu-\theta)^2}{2\lambda^2}$$

calculating $\arg\max_{\mu} \log P(\mu|x)$ by differentiating and equating it to zero, we get:

$$\Rightarrow \frac{d}{d\mu} \left(-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\theta)^2}{2\lambda^2} + C \right) = 0, \text{ where } C = \text{constant}$$

$$\Rightarrow \frac{2}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu) - \frac{2(\mu-\theta)}{2\lambda^2} + 0 = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{x_i - N\mu}{\sigma^2} + \frac{\theta - \mu}{\lambda^2} = 0$$

$$\Rightarrow \lambda^2 \sum_{i=1}^{N} x_i - N\mu\lambda^2 + \sigma^2\theta - \mu\sigma^2 = 0$$

$$\Rightarrow \lambda^2 \sum_{i=1}^{N} x_i + \sigma^2\theta = \mu(N\lambda^2 + \sigma^2)$$

$$\Rightarrow \hat{\mu}_{MAP} = \left(\frac{\lambda^2 \sum_{i=1}^{N} x_i + \sigma^2\theta}{N\lambda^2 + \sigma^2} \right)$$

Q2. Parameter Estimation

1. **Please provide the MLE estimator of λ.**

Solution 1.

(2.) PARAMETER ESTIMATION

Given N integers $K_1, K_2, \ldots, K_N \in Z$ which are iid samples from the following Poisson distribution:

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

(1) MLE ESTIMATOR OF λ:

$$\hat{\lambda}_{MLE} = \arg\max_{\lambda} P(K|\lambda)$$

We will use the log-likelihood function to compute the MLE for λ.

Prob. of N iid samples $K = \{K_1, K_2, \ldots K_N\}$:

$$P(K|\lambda) = \frac{\lambda^{K_1+K_2+\cdots+K_N} e^{-N\lambda}}{K_1! K_2! K_3! \cdots K_N!}$$

Let $K_1! K_2! \ldots K_N! = C$, a constant

then $\ln P(K|\lambda) = \ln \lambda^{\sum_{i=1}^{N} K_i} + \ln e^{-N\lambda} - \ln C$

Now calculating MLE for λ

$$\arg\max_{\lambda} \ln \lambda^{\sum_{i=1}^{N} K_i} + \ln e^{-N\lambda} - \ln C$$

$$\Rightarrow \arg\max_{\lambda} \sum_{i=1}^{N} K_i \ln\lambda - N\lambda - \ln C$$

$$\Rightarrow \text{Set derivative } \frac{d}{d\lambda} \ln P(K|\lambda) = 0$$

$$\Rightarrow \frac{d}{d\lambda}\left(\sum_{i=1}^{N} K_i \ln\lambda - N\lambda - \ln C\right) = 0$$

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

$$\Rightarrow \boxed{\lambda = \frac{\sum_{i=1}^{N} K_i}{N}} \quad \text{Answer}$$

2. Let X be a discrete random variable with the Poisson distribution, What is the expectation E[X]?

Solution 2:

(2.) What is $E(X)$, where X is a discrete random variable?

$$E(X) = \sum_{i=0}^{\infty} x_i P(x_i|\lambda)$$

$$\Rightarrow \sum_{i=0}^{\infty} x_i \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\Rightarrow \sum_{i=0}^{\infty} \lambda \cdot \frac{\lambda^{x_i-1} e^{-\lambda}}{(x_i-1)!}$$

$$\Rightarrow \lambda \cdot e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{x_i-1}}{(x_i-1)!}$$

$$\Rightarrow \lambda \cdot e^{-\lambda} \cdot e^{\lambda}$$

$$\Rightarrow \lambda \quad \underline{\text{Answer}}$$

Q3. Naïve Bayes Classifier

Given the training data set shown in Table 1, we train a Naı¨ve Bayes classifier with it. Each row refers to an apple, where the categorical features (size, color and shape) and the class label (whether one apple is good) are shown.

NOTE: We assume that the features size, color and shape are conditionally independent.

1. (5 points) How many independent parameters would be there for the Naı¨ve Bayes classifier trained with this data? What are they? Justify the your answers.

Solution 1:

3 NAÏVE BAYES CLASSIFIER

SOLUTION 1: There would be 7 independent parameters for the Naïve Bayes classifier trained with this data.

They are:

(1.) $P(Y = \text{Yes})$

(2.) $P(x_1 = \text{small} \mid Y = \text{Yes})$

(3.) $P(x_2 = \text{Green} \mid Y = \text{yes})$

(4.) $P(x_3 = \text{Irregular} \mid Y = \text{yes})$

(5.) $P(x_1 = \text{small} \mid Y = \text{No})$

(6.) $P(x_2 = \text{Green} \mid Y = \text{No})$

(7.) $P(x_3 = \text{Irregular} \mid Y = \text{No})$

where Y is the label which represents the class good_apple and x_1, x_2, x_3 are features representing size, color and shape respectively.

Since the class label output is binary, we need one parameter from the prior (1).

Size has 2 distinct values, so we need 1 parameter per class. (2)

color has 2 distinct values, so we need 1 parameter per class. (2)

shape has 2 distinct values, so we need 1 parameter per class. (2)

Hence, there would be 7 independent parameters in total.

2. (10 points) Using standard MLE, what are the estimated values for these parameters?

Solution 2:

SOLUTION 2:

Using standard MLE, the estimated values for these parameters are as follows:

(1) $P(Y = Yes) = 4/10$; Therefore $P(Y = No) = 6/10$

(2) $P(x_1 = Small \mid Y = Yes) = 1/4$; Therefore $P(x_1 = Large \mid Y = Yes) = 3/4$

(3) $P(x_2 = Green \mid Y = Yes) = 0/4$; Therefore $P(x_2 = Red \mid Y = Yes) = 1$ ie $(4/4)$

(4) $P(x_3 = Irregular \mid Y = Yes) = 1/4$; Therefore $P(x_3 = circle \mid Y = Yes) = 3/4$

(5) $P(x_1 = Small \mid Y = No) = 3/6 = 1/2$; Therefore $P(x_1 = Large \mid Y = No) = 1/2$

(6) $P(x_2 = Green \mid Y = No) = 5/6$; Therefore $P(x_2 = Red \mid Y = No) = 1/6$

(7) $P(x_3 = Irregular \mid Y = No) = 4/6 = 2/3$; Therefore $P(x_3 = circle \mid Y = No) = 1/3$

where Y = good_apple, x_1 = Size, x_2 = Color, x_3 = Shape are the parameters involved.

3. (5 points) Given a new apple with features x = (Small, Red, Circle), what is P (y = No | x)? Would the Naı¨ve Bayes classifier predict y = Yes or y = No for this apple?

Solution 3:

SOLUTION 3 :

Given a new apple with features $x = (\text{Small}, \text{Red}, \text{Circle})$, what is $P(y=\text{No}|x)$?

Using Bayes' Theorem :

$$\Rightarrow P(y=\text{No}|x_1=\text{Small}, x_2=\text{Red}, x_3=\text{Circle}) = \frac{P(y=\text{No})P(x_1=\text{Small}, x_2=\text{Red}, x_3=\text{Circle}|y=\text{No})}{[P(y=\text{No})P(x_1=\text{Small}, x_2=\text{Red}, x_3=\text{Circle}|y=\text{No}) + P(y=\text{Yes})P(x_1=\text{Small}, x_2=\text{Red}, x_3=\text{Circle}|y=\text{Yes})]}$$

Now as x_1, x_2, x_3 are independent features given class :

$$P(x_1, x_2, x_3|y=\text{No}) = P(x_1|y=\text{No})P(x_2|y=\text{No})P(x_3|y=\text{No})$$

Using the above property, we get :

$$\Rightarrow P(y=\text{No}|x_1=\text{Small}, x_2=\text{Red}, x_3=\text{Circle}) = \frac{P(y=\text{No})P(x_1|y=\text{No})P(x_2|y=\text{No})P(x_3|y=\text{No})}{P(y=\text{No})P(x_1|y=\text{No})P(x_2|y=\text{No})P(x_3|y=\text{No}) + P(y=\text{Yes})P(x_1|y=\text{Yes})P(x_2|y=\text{Yes})P(x_3|y=\text{Yes})}$$

$$\Rightarrow \frac{6/10 \times 1/2 \times 1/6 \times 1/3}{[(6/10 \times 1/2 \times 1/6 \times 1/3) + (4/10 \times 1/4 \times 1 \times 3/4)]}$$

$$P(y=\text{No}|x) \Rightarrow \frac{1/60}{1/60 + 3/40} \Rightarrow \frac{1/6}{22/24} = \frac{4}{22} = \frac{2}{11} = 0.182$$

Since, $P(y=\text{No}|x) = 0.182 < 0.5$, the Naïve Bayes classifier will predict Yes for this apple

Q4. Logistic Regression

LOGISTIC REGRESSION: ①

SOLUTION 4:

Positive Examples: $x_1 = (1,0,0)$, $x_2 = (0,0,1)$, $x_3 = (0,1,0)$

$y_1 = 1$, $y_2 = 1$, $y_3 = 1$

Negative Examples: $x_4 = (-1,0,0)$, $x_5 = (0,-1,0)$, $x_6 = (0,0,-1)$; $y_4 = 0$, $y_5 = 0$, $y_6 = 0$

Learning rate given = η

Initial weight vectors given: (1) $w_0 = (0,0,0,0)'$ (2) $w_0 = (0,0,1,0)'$

(1.) With $w_0^0 = (0,0,0,0)'$

1st Iteration: with $w_0 = (0,0,0,0)'$

$$P(y^j = 1 \mid x^j, w_0^{(t)}) = \frac{\exp\left(w_0^{(t)} + \sum_{i=1}^{3} w_i^{(t)} x_i\right)}{1 + \exp\left(w_0^{(t)} + \sum_{i=1}^{3} w_i^{(t)} x_i\right)}$$

Now, applying the above formula for $j = 1,2,3,4,5,6$.

$$P(y^1 = 1 \mid x^1, w^{(0)}) = \frac{\exp(0 + 0\times1 + 0\times0 + 0\times0)}{1 + \exp(0 + 0\times1 + 0\times0 + 0\times0)} = 0.5$$

$$P(y^2 = 1 \mid x^2, w^{(0)}) = \frac{\exp(0 + 0\times0 + 0\times0 + 0\times1)}{1 + \exp(0 + 0\times0 + 0\times0 + 0\times1)} = 0.5$$

$$P(y^3 = 1 \mid x^3, w^{(0)}) = \frac{\exp(0 + 0\times0 + 0\times1 + 0\times0)}{1 + \exp(0 + 0\times0 + 0\times1 + 0\times0)} = 0.5$$

$$P(y^4 = 1 \mid x^4, w^{(0)}) = \frac{\exp(0 + 0\times-1 + 0\times0 + 0\times0)}{1 + \exp(0 + 0\times-1 + 0\times0 + 0\times0)} = 0.5$$

$$P(y^5 = 1 \mid x^5, w^{(0)}) = \frac{\exp(0 + 0\times0 + 0\times-1 + 0\times0)}{1 + \exp(0 + 0\times0 + 0\times-1 + 0\times0)} = 0.5$$

$$P(y^6 = 1 \mid x^6, w^{(0)}) = \frac{\exp(0 + 0\times0 + 0\times0 + 0\times-1)}{1 + \exp(0 + 0\times0 + 0\times0 + 0\times-1)} = 0.5$$

We know that

$$w_0^{(t+1)} = w_0^{(t)} + \eta \sum_j \left[y^j - \hat{P}(y^j = 1 \mid x^j, w^{(t)}) \right]$$

Therefore, $w_0^{(1)} = w_0^{(0)} + \eta\left[(1-0.5)+(1-0.5)+(1-0.5)+(0-0.5)+(0-0.5)+(0-0.5)\right]$ ①

$\Rightarrow w_0^{(1)} = 0 + \eta(0) = 0$

We also know that: $w_i^{(t+1)} = w_i^{(t)} + \eta \sum_j x_i^j \left(y^j - \hat{P}\left(y^j = 1 \mid x^j, w^{(t)}\right)\right)$

So, for $i = 1, 2, 3$:

$w_1^{(1)} = w_1^{(0)} + \eta\left[(1)(1-0.5) + 0(1-0.5) + 0(1-0.5) + (-1)(0-0.5) + 0(0-0.5) + (0)(0-0.5)\right]$

$= 0 + \eta(0.5+0.5) = \eta$

$w_2^{(1)} = w_2^{(0)} + \eta\left[(0)(1-0.5) + 0(1-0.5) + 1(1-0.5) + 0(0-0.5) + (-1)(0-0.5) + (0)(0-0.5)\right]$

$= 0 + \eta(0.5+0.5) = \eta$

$w_3^{(1)} = w_3^{(0)} + \eta\left[(0)(1-0.5) + 1(1-0.5) + 0(1-0.5) + 0(0-0.5) + 0(0-0.5) + (-1)(0-0.5)\right]$

$= 0 + \eta(0.5+0.5) = \eta$

$\therefore w^{(1)} = (0, \eta, \eta, \eta)$

2nd Iteration: with $w^{(1)} = (0, \eta, \eta, \eta)'$

$P(y^1=1 \mid x^1, w^{(1)}) = \dfrac{\exp(0 + \eta\times1 + \eta\times0 + \eta\times0)}{1+\exp(0+\eta\times1+\eta\times0+\eta\times0)} = \dfrac{\exp(\eta)}{1+\exp(\eta)}$

$P(y^2=1 \mid x^2, w^{(1)}) = \dfrac{\exp(0 + \eta\times0 + \eta\times0 + \eta\times1)}{1+\exp(0+\eta\times0+\eta\times0+\eta\times1)} = \dfrac{\exp(\eta)}{1+\exp(\eta)}$

$P(y^3=1 \mid x^3, w^{(1)}) = \dfrac{\exp(0 + \eta\times0 + \eta\times1 + \eta\times0)}{1+\exp(0+\eta\times0+\eta\times1+\eta\times0)} = \dfrac{\exp(\eta)}{1+\exp(\eta)}$

$P(y^4=1 \mid x^4, w^{(1)}) = \dfrac{\exp(0 + \eta\times-1 + \eta\times0 + \eta\times0)}{1+\exp(0+\eta\times-1+\eta\times0+\eta\times0)} = \dfrac{\exp(-\eta)}{1+\exp(-\eta)} = \dfrac{1}{1+\exp(\eta)}$

$P(y^5=1 \mid x^5, w^{(1)}) = \dfrac{\exp(0 + \eta\times0 + \eta\times-1 + \eta\times0)}{1+\exp(0+\eta\times0+\eta\times-1+\eta\times0)} = \dfrac{\exp(-\eta)}{1+\exp(-\eta)} = \dfrac{1}{1+\exp(\eta)}$

$P(y^6=1 \mid x^6, w^{(1)}) = \dfrac{\exp(0 + \eta\times0 + \eta\times0 + \eta\times-1)}{1+\exp(0+\eta\times0+\eta\times0+\eta\times-1)} = \dfrac{\exp(-\eta)}{1+\exp(-\eta)} = \dfrac{1}{1+\exp(\eta)}$

$$w_0^{(2)} = w_0^{(1)} + \eta\left[\left(1 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) + \left(1 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) + \left(1 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) + \left(0 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) \times 3\right]$$

$$w_0^{(2)} = 0 + \eta\left(\frac{3}{1+\exp(\eta)} - \frac{3}{1+\exp(\eta)}\right) = 0$$

$$w_1^{(2)} = \eta + \eta\left[(1)\left(1 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) + 0\left(\frac{1}{1+\exp(\eta)}\right) + 0\left(\frac{1}{1+\exp(\eta)}\right) + (-1)\left(\frac{-1}{1+\exp(\eta)}\right) + 0\left(\frac{-1}{1+\exp(\eta)}\right) + 0\left(\frac{-1}{1+\exp(\eta)}\right)\right]$$

$$= \eta + \frac{2\eta}{1+\exp(\eta)}$$

$$w_2^{(2)} = \eta + \eta\left(0\left(\frac{1}{1+\exp(\eta)}\right) + 0\left(\frac{1}{1+\exp(\eta)}\right) + \frac{1}{1+\exp(\eta)} + 0\left(\frac{-1}{1+\exp(\eta)}\right) + (-1)\left(\frac{-1}{1+\exp(\eta)}\right) + 0\left(\frac{-1}{1+\exp(\eta)}\right)\right)$$

$$= \eta + \frac{2\eta}{1+\exp(\eta)}$$

$$w_3^{(2)} = \eta + \eta\left(0\left(\frac{1}{1+\exp(\eta)}\right) + 1\left(\frac{1}{1+\exp(\eta)}\right) + 0\left(\frac{1}{1+\exp(\eta)}\right) + 0\left(\frac{-1}{1+\exp(\eta)}\right) + 0\left(\frac{-1}{1+\exp(\eta)}\right) + (-1)\left(\frac{-1}{1+\exp(\eta)}\right)\right)$$

$$= \eta + \frac{2\eta}{1+\exp(\eta)}$$

$\therefore$ Therefore: $w^{(2)} = \left(0, \eta + \frac{2\eta}{1+\exp(\eta)}, \eta + \frac{2\eta}{1+\exp(\eta)}, \eta + \frac{2\eta}{1+\exp(\eta)}\right)$

If we continue training this model for many iterations, the final weight vector would be: $w^* = (0, \infty, \infty, \infty)$.

Therefore, $w^{(*)} = (0, \infty, \infty, \infty)'$ when $w^{(0)} = (0, 0, 0, 0)'$

(2.) With initial weight vector = $w_0 = (0,0,1,0)'$

1st iteration: $w = (0,0,1,0)'$

$$P(y^1=1 \mid x^1, w^{(0)}) = \frac{\exp(0+(0\times1)+(1\times0)+(0\times0))}{1+\exp(0+0+0+0)} = 0.5$$

$$P(y^2=1 \mid x^2, w^{(0)}) = \frac{\exp(0+(0\times0)+(1\times0)+(0\times1))}{1+\exp(0+0+0+0)} = 0.5$$

$$P(y^3=1 \mid x^3, w^{(0)}) = \frac{\exp(0+(0\times0)+(1\times1)+(0\times0))}{1+e} = \frac{e}{1+e}$$

$$P(y^4=1 \mid x^4, w^{(0)}) = \frac{\exp(0+(0\times-1)+(1\times0)+(0\times0))}{1+\exp(0)} = 0.5$$

$$P(y^5=1 \mid x^5, w^{(0)}) = \frac{\exp(0+(0\times0)+(1\times-1)+(0\times0))}{1+\exp(-1)} = \frac{1}{1+e}$$

$$P(y^6=1 \mid x^6, w^{(0)}) = \frac{\exp(0+(0\times0)+(0\times0)+(0\times-1))}{1+\exp(0)} = 0.5$$

$$w_0^{(1)} = w_0^{(0)} + \eta\left(0(1-0.5)+0(1-0.5)+0\left(1-\frac{e}{1+e}\right)+0(0.5)+0\left(\frac{-1}{1+e}\right)+0(0-0.5)\right)$$

$$= 0 + \eta(0) = 0$$

$$w_1^{(1)} = w_1^{(0)} + \eta\left(1(0.5)+0(0.5)+0\left(1-\frac{e}{1+e}\right)+(-1)(-0.5)+0\left(\frac{-1}{1+e}\right)+0(0-0.5)\right)$$

$$\Rightarrow 0 + \eta(1) = \eta$$

$$w_2^{(1)} = w_2^{(0)} + \eta\left(0(0.5)+0(0.5)+1\left(\frac{1}{1+e}\right)+0(-0.5)+(-1)\left(\frac{-1}{1+e}\right)+0(-0.5)\right)$$

$$= 1 + \frac{2\eta}{1+e}$$

$$w_3^{(1)} = w_3^{(0)} + \eta\left((0)(0.5)+1(0.5)+0\left(\frac{1}{1+e}\right)+0(-0.5)+(0)\left(\frac{-1}{1+e}\right)+(-1)(-0.5)\right)$$

$$\Rightarrow 0 + \eta(1) = \eta$$

$$\therefore w^{(1)} = \left(0, \eta, 1+\frac{2\eta}{1+e}, \eta\right)$$

2nd iteration with $w^{(1)} = \left(0, \eta, 1+\frac{2\eta}{1+e}, \eta\right)$

$$P\left(Y^1=1 \mid X^1, w^{(1)}\right) = \frac{\exp\left(0+(\eta\times 1)+\left(1+\frac{2\eta}{1+e}\right)\times 0+\eta\times 0\right)}{1+\exp(\eta)} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$P\left(Y^2=1 \mid X^2, w^{(1)}\right) = \frac{\exp\left(0+\eta\times 0+\left(1+\frac{2\eta}{1+e}\right)\times 0+\eta\times 1\right)}{1+\exp(\eta)} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$P\left(Y^3=1 \mid X^3, w^{(1)}\right) = \frac{\exp\left(0+\eta\times 0+\left(1+\frac{2\eta}{1+e}\right)\times 1+\eta\times 0\right)}{1+\exp\left(1+\frac{2\eta}{1+e}\right)} = \frac{\exp\left(1+\frac{2\eta}{1+e}\right)}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}$$

$$P\left(Y^4=1 \mid X^4, w^{(1)}\right) = \frac{\exp\left(0+\eta\times -1+\left(1+\frac{2\eta}{1+e}\right)\times 0+\eta\times 0\right)}{1+\exp(-\eta)} = \frac{1}{1+\exp(\eta)}$$

$$P\left(Y^5=1 \mid X^5, w^{(1)}\right) = \frac{\exp\left(0+\eta\times 0+\left(1+\frac{2\eta}{1+e}\right)\times -1+\eta\times 0\right)}{1+\exp\left(-\left(1+\frac{2\eta}{1+e}\right)\right)} = \frac{1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}$$

$$P\left(Y^6=1 \mid X^6, w^{(1)}\right) = \frac{\exp\left(0+\eta\times 0+\left(1+\frac{2\eta}{1+e}\right)\times 0+\eta\times -1\right)}{1+\exp(-\eta)} = \frac{1}{1+\exp(\eta)}$$

Now updating weights:

$$w_0^{(2)} = w_0^{(1)} + \eta\left(\left(1-\frac{\exp(\eta)}{1+\exp(\eta)}\right)+\left(1-\frac{\exp(\eta)}{1+\exp(\eta)}\right)+\left(1-\frac{\exp\left(1+\frac{2\eta}{1+e}\right)}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right)+\left(0-\frac{1}{1+\exp(\eta)}\right)+\left[0-\frac{1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right]+\left(0-\frac{1}{1+\exp(\eta)}\right)\right]$$

$$\Rightarrow 0+\eta\left(\frac{2}{1+\exp(\eta)}-\frac{2}{1+\exp(\eta)}+\frac{1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}-\frac{1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right]$$

$$\Rightarrow 0+\eta(0) = 0$$

$$w_1^{(2)} = \eta + \eta\left[(1)\left(\frac{1}{1+\exp(\eta)}\right) + (0)\left[\frac{1}{1+\exp(\eta)}\right] + (0)\left[\frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right]\right.$$

$$\left.(-1)\left[\frac{-1}{1+\exp(\eta)}\right] + (0)\left[\frac{-1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right] + (0)\left(\frac{-1}{1+\exp(\eta)}\right)\right]$$

$$\Rightarrow \eta + \frac{2\eta}{1+\exp(\eta)}$$

$$w_2^{(2)} = \left(1+\frac{2\eta}{1+\epsilon}\right) + \eta\left[(0)\left[\frac{1}{1+\exp(\eta)}\right] + (0)\left[\frac{1}{1+\exp(\eta)}\right] + (1)\left[\frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right]\right.$$

$$\left.(0)\left[\frac{-1}{1+\exp(\eta)}\right] + (-1)\left[\frac{-1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right] + (0)\left[\frac{-1}{1+\exp(\eta)}\right]\right)$$

$$\Rightarrow 1+\frac{2\eta}{1+\epsilon} + \frac{2\eta}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}$$

$$w_3^{(2)} = \eta + \eta\left[(0)\left[\frac{1}{1+\exp(\eta)}\right] + (1)\left[\frac{1}{1+\exp(\eta)}\right] + (0)\left[\frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right]\right.$$

$$\left.+ (0)\left[\frac{-1}{1+\exp(\eta)}\right] + (0)\left[\frac{-1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right] + (-1)\left[\frac{-1}{1+\exp(\eta)}\right]\right]$$

$$\Rightarrow \eta + \frac{2\eta}{1+\exp(\eta)}$$

$$\therefore w^{(2)} = \left(0,\ \eta+\frac{2\eta}{1+\exp(\eta)},\ 1+\frac{2\eta}{1+\epsilon}+\frac{2\eta}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)},\ \eta+\frac{2\eta}{1+\exp(\eta)}\right)'$$

If we continue training this model for many iterations, the final weight vector w^* would be $w^* = (0, \infty, \infty, \infty)'$ for $w^{(0)} = (0,0,1,0)'$

Hence, the final weight vector, $w^{(*)}$ will be the same for the two different initial weight vectors $w^{(0)} = [0,0,0,0]'$ and $w^{(0)} = [0,0,1,0]'$

Q5. Naïve Bayes Classifier and Logistic Regression

---

1. Gaussian Naı¨ve Bayes and Logistic Regression.

(5) NAÏVE BAYES CLASSIFIER AND LOGISTIC REGRESSION:

(1.) GAUSSIAN NAÏVE BAYES AND LOGISTIC REGRESSION:

— How many independent parameters are there in this Gaussian Naïve Bayes classifier? What are they?

Solution: We know that $P(Y=k|X_i) = \frac{P(X_i|Y=k)P(Y=k)}{P(X_i)}$, but the factors which count are the factors in the numerators $P(X_i|Y=k)$ and $P(Y=k)$, because the denominator is constant.

so $P(Y=k|X_i) \propto P(X|Y=k)P(Y=k)$

We know that X is d dimensional and all features are conditionally independent.

Therefore, $P(Y=k|X) \propto P(X_1|Y=k)P(X_2|Y=k)\ldots P(X_d|Y=k)P(Y=k)$

K entails 2 values: $\{0,1\}$. Therefore we will have $(2-1)=1$ independent parameter from the prior.

We also know that $P(X_i|Y=k) = N(\mu_{ik}, \sigma_{ik})$ i.e. we will have 2 parameters per likelihood function per class Y.

i.e.: set of μ's $= \{\mu_{10}, \mu_{20}, \mu_{30}, \ldots \mu_{d0}, \mu_{11}, \mu_{21}, \mu_{31}, \ldots, \mu_{d1}\}$

set of σ's $= \{\sigma_{10}, \sigma_{20}, \sigma_{30}, \ldots \sigma_{d0}, \sigma_{11}, \sigma_{21}, \sigma_{31}, \ldots, \sigma_{d1}\}$

Hence, total independent parameters in the Gaussian Naïve Bayes classifier: size of μ set + size of σ set + (parameters from prior)

$\Rightarrow \quad 2d + 2d + 1$

$\Rightarrow \quad 4d + 1$

— Can we translate w into parameters of Gaussian Naive Bayes without extra assumption?

Solution: No. We cannot translate w into parameters of Gaussian Naive Bayes without extra assumption. But, it can be translated by making the assumption that variance is the same for all the classes i.e. $\sigma_{ik} = \sigma_i$ for all classes.

PROOF: We know that:

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$\Rightarrow \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$\Rightarrow \frac{1}{1 + \exp\left(\ln\left(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)\right)}$$

$$\Rightarrow \frac{1}{1 + \exp\left(\left(\ln\frac{1-\theta}{\theta}\right) + \sum_i \left(\left(\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}\right)X_i + \left(\frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right)\right)\right)}$$

where $P(Y=1) = \theta$ and $P(Y=0) = 1-\theta$

Hence, the above equation is of the form: $\frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$

Therefore, the parameters w of logistic regression can be translated to Gaussian Naive Bayes parameters with the assumption that all classes of Y, have the same variance.

2. Implementation of Gaussian Naı¨ve Bayes and Logistic Regression

(i)

- **Pseudocode for Gaussian Naive Bayes:**

1. Perform 3-fold cross validation to split the data into training and testing sets for every fraction:

```
Data: X(a 1*4 array with 4 features) and Y (corresponding labels)

For every fraction of data: [0.01,0.02,0.05,0.1, 0.625,1.0]
    Randomly shuffle the data and split it in 3 parts.
    For every split, let one part be the test-set and the other two parts be
the training set.
    Randomly shuffle the data for 5 times and add it to the corresponding
fraction's dataset.
    Repeat the above process such that every split has been a test-set and
the other two the training-set.
```

2. Next step involves training the Gaussian Naive Bayes classifier:

We calculate parameters like priors, mean and variance for classes 0 and 1 from the training-set.

```
P(Y=1)= (# of 1's in the training space)/(total size of training space)
P(Y=0)= (# of 0's in the training space)/(total size of training space)

mean(Y=1) = average over all X's for which the corresponding label is 1
mean(Y=0) = average over all X's for which the corresponding label is 0

var(Y=1) = variance over all X's for which the corresponding label is 1
var(Y=0) = variance over all X's for which the corresponding label is 0
```

3. Predict the Labels on X_test using the trained parameters:

Our goal is now to compute the posteriors: $P(Y=1|X)$ and $P(Y=0|X)$.

We know that:

$$P(Y=1|X) = P(X|Y=1)*P(y=1)$$
$$P(Y=0|X) = P(X|Y=0)*P(y=0)$$

NOTE : The denominator(marginal probability P(X)) is not considered because it is the same for both $P(Y=1|X)$ and $P(Y=0|X)$.

We have assumed that the features of X(x1,x2,x3,x4)are conditionally independent and the likelihood function is a gaussian function. Therefore:

```
P(X|Y=1) = P(x1|Y=1)*P(x2|Y=1)*P(x3|Y=1)*P(x4|Y=1)
P(X|Y=0) = P(x1|Y=0)*P(x2|Y=0)*P(x3|Y=0)*P(x4|Y=0)
```

and P(xi|Y=j) is computed as follows:

```
P(xi|Y=j)= exp(-(xi-mean(Y=j)^2)/2*var(Y=j))/sqrt(2*pi*var(Y=j))
```

```
So, if P(Y=1|X)>= P(Y=0|X), we write 1 to the prediction set
    if P(Y=1|X)<  P(Y=0|X), we write 0 to the prediction set
```

4. Calculate the accuracy of the prediction set on the basis of Y_test

Accuracy=(# of matching values in the prediction set and Y_test)/(total size)

5. Average the accuracy for 5 different training and testing sets for the same fraction of data and plot the graph.

- ## Pseudocode for Logistic Regression:

1. Perform 3-fold cross validation to split the data into training and testing sets for every fraction:

```
Data: X(a 1*4 array with 4 features) and Y (corresponding labels)

For every fraction of data: [0.01,0.02,0.05,0.1, 0.625,1.0]
    Randomly shuffle the data and split it in 3 parts.
    For every split, let one part be the test-set and the other two parts be
the training set.
    Randomly shuffle the data for 5 times and add it to the corresponding
fraction's dataset.
    Repeat the above process such that every split has been a
test-set(X_test,Y_test) and the other two the training-set(X_train,Y_train).
```

2. Next step involves training the Logistic Regression classifier:

```
We decide on a learning rate(eta) and the number of iterations N.
We initialize the weight vector W to zeros of dimensions (4*1) and w0 to be 0.

Iterate N times:
    Calculate Z= dot(W.T,X_train) + w0
    We calculate P(Y|X) directly= sigmoid(Z)

    We then calculate the gradients:
    dw0= (sum(Y_train - P(Y|X)))/(No. Of training samples)
```

```
        dW = (dot(Y_train - P(Y|X), X_train.T))/(No. Of training samples)

        We then update the weights:
        W = W + eta*dW
        w0 = w0 + eta*dw0

After N iterations we return the weights.
```

3. Predict the Labels on X_test using the updated weights:

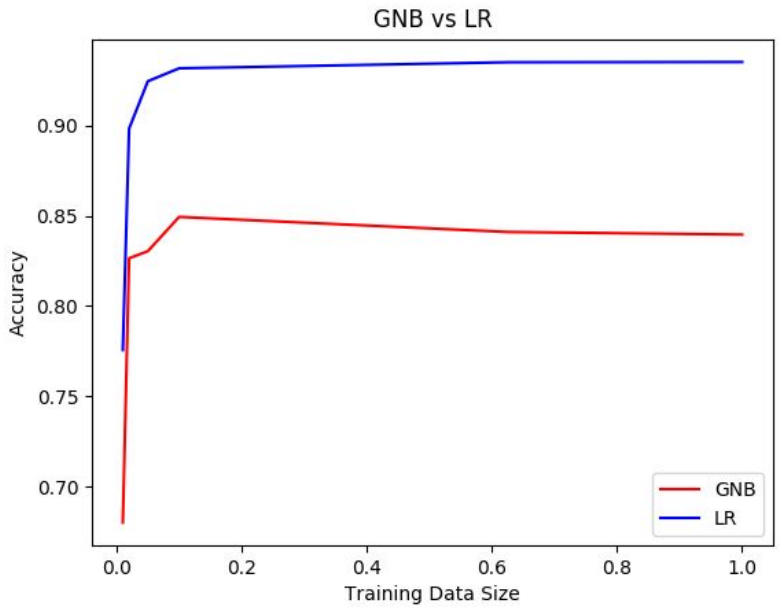
```
We compute P(Y=1|X=X_test) directly:
    Z= dot(W.T,X_test)+w0
    P(Y=1|X=X_test)= sigmoid(Z)

For every example in the X_test:
    if P(Y=1|X=X_test)>= 0.5, we write 1 to the prediction set
    if P(Y=1|X=X_test)< 0.5, we write 0 to the prediction set

We return the prediction set.
```

4. Calculate the accuracy of the prediction set on the basis of Y_test

```
Accuracy=(# of matching values in the prediction set and Y_test)/(total size)
```

5. Average the accuracy for 5 different training and testing sets for the same fraction of data and plot the graph.

(ii) Plot a learning curve:



(iii) Show the power of generative model:

```
('Fold=', 1)
('Training data mean=', array([-1.86733672, -0.99412709,  2.14739099, -1.25065034]))
('Generated data mean=', array([-1.95025607, -1.28903237,  2.20072703, -1.41661878]))
('Difference between means=', array([0.08291935, 0.29490529, 0.05333603, 0.16596845]))
('Training data variance=', array([ 3.53810202, 29.21258318, 27.68618191,4.27916742]))
('Generated data variance=', array([ 3.35099026, 30.95730264, 29.60706636,4.18542992]))
('Difference between variance=', array([0.18711176, 1.74471946, 1.92088445, 0.093737]))

('Fold=', 2)
('Training data mean=', array([-1.870034  , -0.93987612,  2.13177332, -1.22410755]))
('Generated data mean=', array([-1.96356534, -0.87560291,  2.39586256, -1.23813414]))
('Difference between means=', array([0.09353134, 0.06427322, 0.26408924, 0.01402659]))
('Training data variance=', array([ 3.65337257, 29.01119813, 27.72164037, 4.28211005]))
('Generated data variance=', array([ 4.08835828, 31.5482741 , 26.65189289,3.89760372]))
('Difference between variance=', array([0.43498571, 2.53707597,1.06974748,0.38450634]))

('Fold=', 3)
('Training data mean=', array([-1.85922714, -1.15723689,  2.19434676, -1.33045331]))
('Generated data mean=', array([-1.78855922, -1.78065952,  1.88935959, -1.40001651]))
('Difference between means=', array([0.07066792, 0.62342263, 0.30498717, 0.0695632 ]))
('Training data variance=', array([ 3.19144435, 29.78260957, 27.57663506,  4.26183345]))
('Generated data variance=', array([ 3.51018461, 27.12193768, 25.99142382,  4.62184534]))
('Difference between variance=', array([0.31874026, 2.66067188, 1.58521124, 0.36001189]))
```

From the above results, we can observe that there is very little absolute difference in the mean and variance of the training data and the generated data. This is because both datasets belong to the same Gaussian Distribution with the same mean and variance.
