CSE575: SML Assignment 1 — 02/15/2019

(Q1)

(1) Given:

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

here
$\hat{\sigma}^2$ – sample ~~S.D.~~ variance
$\hat{\mu}$ – sample mean.

The actual (unbiased) or populat^n standard deviation is always greater than sample standard deviat^n.

$\mu$ – populat^n mean
$\bar{x} = \hat{\mu}_{MLE}$ or Sample mean
$\sigma^2$ = populat^n ~~s.d~~ variance (unbiased)
$\hat{\sigma}^2$ – $\hat{\sigma}^2_{MLE}$ or Sample variance [biased]

$$\hat{\mu}_{MLE} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

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

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

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} \left( (x_i^2 - 2x_i\mu + \mu^2) - (x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2) \right)\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} (\mu^2 - \hat{\mu}^2 - 2x_i\mu + 2x_i\hat{\mu})\right]$$

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

$$= E\left(\mu^2 - \hat{\mu}^2 - 2(\mu - \hat{\mu})\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

$$= E\left(\mu^2 - \hat{\mu}^2 - 2(\mu - \hat{\mu})\hat{\mu}\right]$$

$$= E\left[\mu^2 - \hat{\mu}^2 - 2\mu\hat{\mu} + 2\hat{\mu}^2\right)$$

$$= E\left[\mu^2 + \hat{\mu}^2 - 2\mu\hat{\mu}\right)$$

$$= E\left[(\hat{\mu} - \mu)^2\right]$$

$$= \text{Var}(\hat{\mu})$$

$$= \frac{\sigma^2}{n}$$

$$\therefore \text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

$$\left( \text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum \text{Var}(x_i) \right.$$

$$\left. = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n} \right)$$

$$\therefore E[\hat{\sigma}^2] = \sigma^2 - \frac{\sigma^2}{n}$$

$$= \sigma^2\left[1 - \frac{1}{n}\right]$$

$$\therefore \hat{\sigma}^2 = \sigma^2\frac{(n-1)}{n}$$

Thus we see that

$$\hat{\sigma}^2 \neq \sigma^2$$

which means variance is biased ie the ~~un~~biased variance is slightly less than than the actual variance

∴ we divide with $(n-1)$ instead of $n$ while calculating standard deviation.

This correct term is also called as Bessels correction

(2) Given

Prior dist$^n$ for mean $\mu \sim N(0, 2)$

$\hat{\mu}_{MAP} = ?$

$$P(D|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$P(\mu|0, 2) = \frac{1}{\sigma\sqrt{2\pi}} - \frac{(\mu-0)^2}{2\cdot 2^2}$$

Now,

$$P(\mu|D) = P(D|\mu)\,P(\mu) \quad (1)$$

$P(\mu|D)$ is the Posterior which we have to maximise

$$P(D|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$P(\mu|\theta,\lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\mu-\theta)^2}{2\lambda^2}}$$

Now we take log & set derivate wrt to $\mu$ to 0.

$$\therefore \ln(\mu|D) = \ln[P(D|\mu,\sigma)\, P(\mu|\theta,\lambda)]$$

$$= \ln P(D|\mu,\sigma) + \ln P(\mu) \quad (2)$$

$$=$$

Now,

$$\ln P(D|\mu,\sigma) = \ln \frac{1}{\sigma\sqrt{2\pi}} + \ln\left[\pi e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= C_1 + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \quad (3)$$

$$\ln P(\mu) = \ln \frac{1}{\lambda\sqrt{2\pi}} + \frac{(\mu-\theta)^2}{2\lambda^2} \quad (4)$$

It is $\lambda$ not $\lambda^2$ since here $\lambda$ is the variance (5)

Taking derivative of (2)

$$\frac{\partial}{\partial u}\left[\ln P(D|u,\sigma) + \ln P(u)\right] = 0$$

From (3) & (4)

$$\frac{\partial}{\partial u}\left[-\frac{\sum(x_i-u)^2}{2\sigma^2} - \frac{(u-0)^2}{2\lambda^2}\right] = 0$$

$$\therefore \frac{\sum(x_i-u)}{\sigma^2} = \frac{(u-0)}{\lambda^2}$$

$$\therefore \frac{\sum x_i}{\sigma^2} - \frac{\sum u}{\sigma^2} = \frac{u}{\lambda^2} - \frac{0}{\lambda^2}$$

$$\therefore \frac{\sum x_i}{\sigma^2} + \frac{0}{\lambda^2} = u\left[\frac{1}{\lambda^2} + \frac{N}{\sigma^2}\right]$$

$$\therefore u = \frac{\frac{\sum x_i}{\sigma^2} + \frac{0}{\lambda^2}}{\frac{1}{\lambda^2} + \frac{N}{\sigma^2}}$$

$$\therefore \hat{u}_{MAP} = \frac{\frac{\sum x_i}{\sigma^2} + \frac{0}{\lambda^2}}{\frac{1}{\lambda^2} + \frac{N}{\sigma^2}}$$

Here, we can see,

the terms $\frac{\sigma^2}{\lambda^2}$ & $\frac{1}{\lambda^2}$ same

from prior, if we ignore

them

we get $\frac{\frac{\sum x_i}{\sigma^2}}{\frac{N}{\sigma^2}}$

$= \frac{1}{N}\sum x_i$

which is $\hat{\lambda}_{MLE}$

∴ we can see adding Prior to MLE gives MAP.

If $N \to \infty$, above terms would vanish (negligible) giving

MLE = MAP

i.e (Posterior = Likelihood)

(Q2)

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{\lfloor k}$$

$k_1, k_2 \ldots k_n \rightarrow$ Poisson samples.

$$P(\text{data}|\lambda) = \prod_{i=1}^{n} P(k_i|\lambda)$$

$$= \prod_{i=1}^{N} \frac{\lambda^{k_i} e^{-\lambda}}{\lfloor k_i}$$

Taking log.

$$\ln \prod \left( \frac{\lambda^{k_i} e^{-\lambda}}{\lfloor k_i} \right)$$

$$= \sum_{i=1}^{n} \ln \left( \frac{\lambda^{k_i} e^{-\lambda}}{\lfloor k_i} \right)$$

$$= \sum_{i=1}^{n} \left[ \ln \lambda^{k_i} - \lambda - \log \lfloor k_i \right]$$

Setting derivate wrt $\lambda = 0$

$$\frac{d}{d\lambda} \sum_{i=1}^{n} \left[ \ln \lambda^{k_i} - \lambda - \log \lfloor k_i \right] = 0$$

$$= \sum_{i=1}^{n} \frac{k_i}{\lambda} - \sum_{i=1}^{n} 1 - 0 = 0$$

$$\therefore \frac{\sum_{i=1}^{n} k_i}{\lambda} - n = 0$$

$$\therefore \sum k_i = n\lambda$$

$$\therefore \boxed{\lambda = \frac{1}{n}\sum_{i=1}^{n} k_i}$$

$$\therefore \hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

(2) Poisson

$$P(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{\lfloor k}$$

$$E(x) = \sum k\, P(k|\lambda)$$

$$= \sum k\, e^{-\lambda}\frac{\lambda^k}{\lfloor k}$$

$$= \sum e^{-\lambda}\frac{\lambda^k}{\lfloor k-1}$$

$$= e^{-\lambda}\sum \lambda\cdot\frac{\lambda^{k-1}}{\lfloor k-1}$$

$$= e^{-\lambda}\lambda\sum\frac{\lambda^{k-1}}{\lfloor k-1}$$

$$\therefore e^{-\lambda} \lambda \leq \frac{\lambda^{k-1}}{\lfloor k-1}$$

Now,

$$\leq \frac{\lambda^{k-1}}{\lfloor k-1} = e^{\lambda} \quad \text{for } k \geq 1$$

$$\therefore e^{-\lambda} \lambda e^{\lambda}$$

$$= \lambda$$

$$\therefore E(+) = \lambda.$$

(3)

(1) here, no. of features = 3
(size, color, shape)

without conditional independent assumptions, we have

$$P(Y|X) = P(X_1, X_2, X_3 | Y)\, P(Y)$$

So for $P(X_1, X_2, X_3 | Y)$

we have to estimate $2(2^n - 1) = 2^{n+1} - 2$ parameters, here $n = 3$

$2^n$ → Every $X$ has 2 values, This gives total combinations of values.

$2 \times 2^n$ → because $Y$ can have 2 values.

for $P(y)$ = 1 parameter.
since $y$ can have 2 values.

With conditional Independence, we have,

$$P(y|x) = P(x_1, x_2, x_3 | y) = P(x_1|y)\, P(x_2|y)\, P(x_3|y)$$

$x$ has 2 values for each feature.
where we only need to estimate
$P(x_1|y)$ $P(x_2|y)$ $P(x_3|y)$
for both values of $y$
which will be $3 \times 2 = 6$ parameters.
& $P(y)$ → 1 Parameter
∴ Totally $-6+1$
$= 7$ parameters.

where
$x_1$ = size
$x_2$ = shape
$x_3$ = color
$y$ = good_apple

| |
|---|
| $P(x_1 \mid y = yes)$ |
| $P(x_1 \mid y = No)$ |
| $P(x_2 \mid y = yes)$ |
| $P(x_2 \mid y = No)$ |
| $P(x_3 \mid y = yes)$ |
| $P(x_3 \mid y = No)$ & $P(y)$ |

[there since only 2 value of $x$ are there for every $x_i$ we can estimate other if one is known.]

(2) for MLE, we try to maximize $P(x|y)$ → likelihood.

$$P(x_1, x_2, x_3 | y) = \prod_{i=1}^{3} P(x_i | y_i)$$

Prior:

$P(y = yes) = 4/10$
$P(y = No) = 6/10$
} from the table

MLE estimate

$$\hat{\pi}_k = \hat{p}(y = y_k)$$

where ~~k = 0 or~~ $y_k \in \{yes, No\}$

$$0 \leq k \leq 1$$

$$\hat{\pi}_k = \frac{\#D\{y = y_k\}}{|D|}$$

where D is the data

$$\hat{\pi}_0 = \frac{\#D\{y = y_0\}}{|D|}$$ where $y_0 = yes$, $y_1 = No$

$$= \frac{4}{10}$$

$$\hat{\pi}_1 = \frac{\#D\{y = y_1\}}{|D|}$$

$$= \frac{6}{10}$$

Now, we estimate cond^ probabilities

$$\hat{\theta}_{ijk} = \hat{p}(x_i = x_{ij} \mid y = y_k)$$

$$= \frac{\#D\{x_i = x_{ij} \mid y = y_k\}}{\#D\{y = y_k\}}$$

Now, $i$ = Sample
$j$ = feature
$k$ = class

For Y = yes,

$$P(X_1 = small \mid Y = yes) = \frac{|x_i = small \wedge Y = yes|}{|Y = yes|}$$

$$= \frac{1}{4}$$

$$P(X_1 = large \mid Y = yes) = \frac{3}{4}$$

$$P(X_2 = Green \mid Y = yes) = 0$$

$$P(X_2 = Red \mid Y = yes) = \frac{4}{4} = 1$$

$$P(X_3 = Irregular \mid Y = yes) = \frac{1}{4}$$

~~$P(X_3 = Regular \mid Y = yes) = ?$~~

$$P(X_3 = Circle \mid Y = yes) = \frac{3}{4}$$

For Y = No,

$$P(X_1 = small \mid Y = No) = \frac{3}{6} = \frac{1}{2}$$

$$P(X_1 = large \mid Y = No) = \frac{3}{6} = \frac{1}{2}$$

$P(x_2 = \text{green} \mid Y = \text{No}) = \frac{5}{6} = \frac{5}{6}$

$P(x_2 = \text{Red} \mid Y = \text{No}) = \frac{1}{6}$

$P(x_3 = \text{Irregular} \mid Y = \text{No}) = \frac{4}{6} = \frac{2}{3}$

$P(x_3 = \text{Circle} \mid Y = \text{No}) = \frac{2}{6} = \frac{1}{3}$

~~& P(Y=Y)~~

(3) Given,

$x = (\text{Small}, \text{Red}, \text{Circle})$

$P(Y = \text{No} \mid x) = ?$

We know, the features are independant

$\therefore P(Y = \text{No} \mid x) \propto P(x \mid Y)\, P(Y)$

$\propto P(x_1, x_2, x_3 \mid Y)\, P(Y)$

where $x_i$ is feature $i$

$= P(x_1 \mid Y)\, P(x_2 \mid Y)\, P(x_3 \mid Y)\, P(Y)$

Since they are independant

$= P(x_1 = \text{Small} \mid Y = \text{No})$

$P(x_2 = \text{Red} \mid Y = \text{No})\, P(x_3 = \text{Circle} \mid Y = \text{No})\, P(\text{No})$

$= \frac{1}{2} \times \frac{1}{6} \times \frac{1}{3} \times \frac{6}{10} \quad \frac{1}{36}$

$6 \frac{1}{36} \times \frac{6}{10}$

$\Rightarrow \frac{1}{60} = 0.0167$

$P(Y = No|X) \propto C \times 0.0167$

where C is Normalisation constant

$= \frac{1}{P(X)}$

Now,

$P(Y = yes|X) = P(X_1 = Small | yes) \times P(X_2 = Red | yes) \times P(X_3 = Grade | yes) \times P(yes)$

$= \frac{1}{4} \times 1 \times \frac{3}{4} \times \frac{4}{10}$

$= \frac{3}{40} = 0.075$

Now

$P(Y = No|X) = 0.0167$

$P(Y = yes|X) = 0.075$

we see that $P(Y = yes|X) > P(Y = No|X)$

∴ we predict yes for this sample of * where * = (Small, red, circle)

(5)

① $P(x_i | y = k) = N(\mu_{ik}, \sigma_{ik})$

$k = 0, 1$

$d = 1 \ldots d$

So, here

~~for~~

~~it will have~~

There will $2 \times d$ different $\mu$

$\because 0 \leq i \leq 1$ (2 values)

$k \leq k \leq d$ (d values)

& $2 \times d$ different $\sigma$

since $\sigma$ is diff for classes & features ie. $\sigma_{ik}$

$0 \leq i \leq 1$

$1 \leq k \leq d$

$\therefore 2d + 2d = 4d$ for $\mu$ & $\sigma$ ie. Gaussian ie. likelihood fund$^n$ $P(x|y)$

& $P(y)$ (Prior) = Bernoulli with one Parameter $\theta$

∴ Total Parameters =

Parameters of Pivor + Parameters of likelihood

$= 1 + 4d$

$= 4d + 1$ parameters

where d are the features.

(2) Yes, we can translate w into GNB parameters with assumption.

Using Bayes's

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

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

Taking exp & log $\Rightarrow$ (doing nothing)

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

Now

Since $P(Y)$ is Bernoulli

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

Also Likelihood is Gaussian

& $P(X|Y=0) = \prod P(x_i|Y=0)$

$P(X|Y=1) = \prod P(x_i|Y=1)$

from (1)

$$\frac{1}{1+\exp\left(\sum \ln\frac{P(x_i|Y=0)}{P(x_i|Y=1)} + \ln\left(\frac{1-\theta}{\theta}\right)\right)} \quad (2)$$

Now Since $P(x_i|Y=Y_k)$ is Gaussian $\sim N(\mu_{ik}, \sigma_{ik})$

We have,

$$P(x_i = x \mid y = k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

$$\therefore P(x_i = x \mid y = 0) = \frac{1}{\sigma_{i0}\sqrt{2\pi}} e^{-\frac{(x-\mu_{i0})^2}{2\sigma_{ik}^2}}$$

& $$P(x_i = x \mid y = 1) = \frac{1}{\sigma_{i1}\sqrt{2\pi}} e^{\frac{(x-\mu_{i1})^2}{2\sigma_{i1}^2}}$$

$\therefore$ ~~from (2) we have~~

$$\therefore \ln\left(\frac{P(x_i = x \mid y = 0)}{P(x_i = x \mid y = 1)}\right) = \ln\left(\frac{\sigma_{i1}}{\sigma_{i0}} e^{-\frac{(x-\mu_{i0})^2}{2\sigma_{ik}^2} + \frac{(x-\mu_{i1})^2}{2\sigma_{i1}^2}}\right)$$

$$= \ln\left(\frac{\sigma_{i1}}{\sigma_{i0}}\right) - \frac{(x-\mu_{i0})^2}{2\sigma_{ik}^2} + \frac{(x-\mu_{i1})^2}{2\sigma_{i1}^2}$$

~~Plugging them in (2)~~

~~we get~~

$$\frac{1}{1+\exp\left(\ln\left(\frac{\sigma_{i1}}{\sigma_{i0}}\right) + \frac{(x-\mu_{i0})^2}{2\sigma_{ik}^2} + \frac{(x-\mu_{i1})^2}{2\sigma_{i1}^2} + \ln\left(\frac{1-\theta}{\theta}\right)\right)}$$

$$\sum_i \ln \frac{P(x_i|y=0)}{P(x_i|y=1)} = \sum_i \ln\left(\frac{\sigma_{i1}}{\sigma_{i0}}\right) + \sum_i \left( \frac{(x_i - \mu_{i1})^2}{2\sigma_{i1}^2} + \frac{(x_i - \mu_{i0})^2}{2\sigma_{i0}^2} \right)$$

$\Rightarrow$

Now,

$$\sum_i \left( \frac{(x_i - \mu_{i1})^2}{2\sigma_{i1}^2} - \frac{(x_i - \mu_{i0})^2}{2\sigma_{i0}^2} \right) \quad \text{(3)}$$

$= \sum_i (x_i$

Here, we assume that $\sigma_{i1} = \sigma_{i0} = \sigma_i$ which means variance is different for diff features but similar for features of diff classes.

ie it doesn't vary with class.

from (3)

$$\therefore \sum_i \left( \frac{(x_i - \mu_{i1})^2 - (x_i - \mu_{i0})^2}{2\sigma_i^2} \right)$$

$$= \sum_i \left( \frac{x_i^2 - 2x_i\mu_{i1} + \mu_{i1}^2 - x_i^2 + 2x_i\mu_{i0} + \mu_{i0}^2}{2\sigma_i^2} \right)$$

$$= \sum_i \left( \frac{-2x_i\mu_{i1} + 2x_i\mu_{i0} + \mu_{i1}^2 + \mu_{i0}^2}{2\sigma_i^2} \right)$$

$$= \sum_i \left( \frac{(\mu_{i0} - \mu_{i1})}{\sigma_i^2} x_i + \frac{\mu_{i1}^2 + \mu_{i0}^2}{2\sigma_i^2} \right) \quad \ldots (4)$$

Substituting (4) in (2)

we get

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

This is similar to logistic Regression function $\frac{1}{1+e^{-z}}$

where $z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$ for d features

∴ Comparing values we can see that

$w_0$ = constant with no x term

$$\therefore w_0 = \ln\left(\frac{1-\theta}{\theta}\right) + \sum \left( \frac{\mu_{i1}^2 + \mu_{i0}^2}{2\sigma_i^2} \right)$$

& $\sum_{i=1}^{d} w_i x_i$ from LR

can be equated with

$$\sum_{i} \frac{(\mu_{i0} - \mu_{i1}) x_i}{2\sigma_i^2}$$

where

$$w_i = \frac{\mu_{i0} - \mu_{i1}}{2\sigma_i^2}$$

∴ we can say that LR & GNB have same form with assumpt that variance doesn't among $ classes for given feature.

Here, we have to make this assumption because in order to convert GNB form to LR form, GNB has to be a linear classifier since LR is a linear classifier

ie for $\frac{1}{1+e^{z}}$ $z = w_0 x_1 + w_1 x_2 + w_d x_d -$

we have all $x_i$ as first degree terms

∴ assuming $\sigma_{ik} = \sigma_i$ helps eliminate $x^2$ terms, making GNB linear.

(1) ★ Pseudo code for logistic regression.

Load data into ~~memory array~~ by
Split X & Y [features & labels].
Divide the data into 3 parts.
for Repeat following 3 times, one for each
Subset of data:
Take $\frac{2}{3}^{rd}$ of data as train data
& remaining $\frac{1}{3}^{rd}$ as test data.

Initialise weight vector
Add a $0^{th}$ column to X
[so that we can do X.W
ie $\sum_{i=0}^{n} w_i x_i$ ]

For ~~iterations~~ every iterations, usually 10,000:
do following:

$z = X.w$
$h = \text{sigmoid}(z)$; i.e.
$\left[ h = \frac{1}{1+e^{-z}} \right]$

Calculate gradient
using formula

$\text{gradient}_j = \sum_j X_j (Y - h)$

i.e.

$\text{Gradient } w = X.(Y - h)$

Then update weights

OR $w_j = w_j + \eta \times \text{gradient}_j$

$w = w + \eta * \text{Gradient}$

After all iterations, we'll get final weight vector which we will use to predict remaining $\frac{1}{3}^{rd}$ (test) data

// Predict & Calculate loss.
Let $X_{test}$ be test data

$z$ = $X_{test}$ . final weight (W)
$h$ = sigmoid (z)
This h will give predicted value.
Find loss using fell funct$^n$
$-y \log h - (1-y) \log (1-h)$ .

Average the loss for all different test data sets used (3 here)

## ★ Pseudo code for Naive Bayes

Load data
Split train & test data
Separate
Split train values into the respective classes.
Now, for every class, for every feature calculate mean & Standard deviat$^n$

for i in class
  for j in features
    $\mu_{ij} = \frac{1}{m}\Sigma x_{ij}$

$$\sigma_{ij} = \frac{1}{m}\Sigma (x_{ij} - \mu_{ij})^2$$

where m is the no. of rows in training data.

This is the training part, Now, we have mean & Std for all classes & ith features, we can use this to Predict new values.

~~for feature in tot~~ Row

$$P(x_i | y = y_k) \quad \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}}$$

for $i^{th}$ feature & $k^{th}$ class

So

$$P(x_i | y = y_0) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_{i0})^2}{2\sigma_{i0}^2}\right)$$

$$P(x_i | y = y_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_{i1})^2}{2\sigma_{i1}^2}\right)$$

$$P(y|x) = \prod_{i=1}^{4} P(x_i | y)$$

~~i.e. for~~

$$P(y = 1 | x) = \prod_{i=1}^{4} P(x_i | y = 1)$$

if $P(Y=1|x) > P(Y=0|x)$
we predict 1
else we predict 0.

(3) We can see from the o/p of Q5 c) code that if we use trained model which has $\mu_{ik}$ & $\sigma_{ik}$ ($i^{th}$ feature & $k^{th}$ class) we can generate samples using these parameters ($\mu$ & $\sigma$)
After generating these samples, we saw that mean and standard deviation of these samples is similar to the mean & std of trained model which seems obvious. We generated the samples from the same model. The samples will be within 3 std of the model.

(2T)

(Q4) $x_1 = (1,0,0)$ $x_2 = (0,0,1)$ $x_3 = (0,1,0)$

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

→ +ve samples

$x_4 = (-1,0,0)$ $x_5 = (0,-1,0)$ $x_6 = (0,0,-1)$

$y_4 = 0$ $y_5 = 0$ $y_6 = 0$

→ -ve samples

$w_0 = (0,0,0,0)^t$ — 1st case (0)

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

Here, since $w^0 = (0,0,0,0)$

we get $w_0 + \sum w_i x_i = 0$

$\therefore P(y=1 \mid x^2, w^0) = P(y=1 \mid x^3, w^0) =$

$P(y=1 \mid x^4, w^0) = P(y=1 \mid x^5, w^0) = P(y=1 \mid x^6, w^0)$

$= 0.5$ (1)

Now,

$$w_0' = w_0^0 + \eta_0 \sum_j x_0^j (y^j - P(y=1 \mid x^j, w_0^0))$$

$= 0 + \eta [1 - 0.5 + 0 - 0.5 + 0$

$0 + 0 - 0.5 + 0 - 0.5]$ ✓

$w_0' = 0 + 0$

$\therefore w_0 = 0$

$$w_1' = w_1^0 + \eta \sum_j x_1^j (y^j - P(y=1 | x^j, w_1^0))$$

$$= 0 + \eta [1(1-0.5) + 0 + 0 + -1(0-0.5)]$$

$$= 0 + \eta [0.5 + 0.5]$$

$$= \eta$$

$$\therefore w_1' = \eta$$

$$w_2' = w_2^0 + \eta \sum_j x_2^j (y^j - P(y=1 | x^j, w_2^0))$$

$$= 0 + \eta [0 + 0 + 1(1-0.5) + 0 + -1(0-0.5)]$$

$$= \eta [0.5 + 0.5]$$

$$= \eta$$

$$w_3' = w_3^0 + \eta \sum_j x_3^j (y^j - P(y=1 | x^j, w_3^0))$$

$$= 0 + \eta [1(1-0.5) + 0 + 0 + 0 + 0 - 1(0-0.5)]$$

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

$$= \eta$$

$$\therefore w' = (0, \eta, \eta, \eta) \quad \ldots (2)$$

Now, we calculate Prob.

$$P(y=1 | x', w') = \frac{e^{w_0 + \sum w_i x_i}}{1 + e^{w_0 + \sum w_i x_i}}$$

$$e^{w_0 + \sum w_i x_i} = e^{0+n} = \frac{e^{n}}{1+e^{0+n}} = \frac{e^{n}}{1+e^{n}}$$

$$\therefore P(y=1 | x^1, w^1) = \frac{e^n}{1+e^n}$$

Similarly for other positive samples we get

$$P(y=1|x^2, w^1) = P(y=1|x^2, w^1) = P(y=1|x^3, w^1) = \frac{e^n}{1+e^n} \quad \text{--- (3)}$$

for negative samples

$$P(y=1|x^4, w^1) = \frac{e^{0+n}}{1+e^{0-n}} = \frac{e^{-n}}{1+e^{-n}}$$

Similarly,

$$P(y=1|x^4, w^1) = P(y=1|x^5, w^1) = P(y=1|x^6, w^1) = \frac{e^{-n}}{1+e^{-n}} \quad \text{--- (4)}$$

Now, we will calculate $w^2$ using Gradient Ascent

$$w_0^2 = w_0^1 + \eta \sum_j x_0^j \left( y^j - P(y=1|x^j, w_0^1) \right)$$

$$= 0 + \eta \left[ 3\left(1 - \frac{e^n}{1+e^n}\right) + 1(0 + 0 + 3\left(0 - \frac{e^{-n}}{1+e^{-n}}\right) \right]$$

Since $x_0^j = 1$

$$= 0 + \eta\left[\frac{1}{1+e^{\eta}} + \frac{e^{-\eta}}{1+e^{-\eta}}\right]$$

$$= 0 + \frac{2\eta}{1+e^{\eta}} \quad \ldots (5) \quad ①$$

$$w_1^2 = w_1^1 + \eta \sum_j x_j^0 \left(y^j - P(y=1 \mid x_j, w^1)\right)$$

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

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

$$w_1^3 = \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta$$

Similarly $w_2^2 = w_3^2 = \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta$

$$\therefore w^2 = \left(0, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta\right)$$

From ⓪, ① & ⑥ ... ⑧

we see

$w^0 = (0,0,0,0)$

$w^1 = (0,\eta,\eta,\eta)$

$w^2 = \left(0, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta, \left(\frac{3+e^{\eta}}{1+e^{\eta}}\right)\eta\right)$

where we can say that $w_0^0$ remain 0 $\forall i$ but $i \neq 3$

and $w_0^0 < w_i^1 < w_i^2$

which means $w$ is increasing constantly

$x_1 = (1,0,0)$ $x_2 = (0,1,0)$ $x_3 = (0,0,1)$ (24)

Now for $w = (0,0,1,0)$ ⟶ 2nd case

$w^0 = (0,0,1,0)$

$$P(y=1|x^1,w^0) = \frac{e^{0+\sum w_i^0 x_i^0}}{1+e^{0+\sum w_i^0 x_i}}$$

$$= \frac{e^{0+[w_2 x_2]}}{1+e^{0+w_2 x_2}}$$

$$= \frac{e^0}{1+e^0} = 0.5$$

$$P(y=1|x^2,w^0) = \frac{e^{w_2 x_2}}{1+e^{w_2 x_2}} = \frac{e}{1+e} = 0.73$$

$$P(y=1|x^3,w^0) = \frac{e^0}{1+e^0} = 0.5$$

$$P(y=1|x^4,w^0) = \frac{e^0}{1+e^0} = 0.5$$

$$P(y=1|x^5,w^0) = \frac{e^{0-1}}{1+e^{-1}} = 0.27$$

$$P(y=1|x^6,w^0) = \frac{e^0}{1+e^0} = 0.5$$

Now, we calculate $w_0^1$ using Gradient Ascent

$$w_0^1 = w_0^0 + \eta \sum_j x_j^0 (y_j^1 - P(y=1|x^j,w_0^0))$$

$$= 0 + \eta[0] = 0.$$

$$w_1^1 = w_1^0 + \eta \sum_j x_j^0 (y^j - P(y=1|x^j,w^0))$$

$$= 0 + \eta(1(1-0.5) - 1(0-0.5))$$
$$= \eta(0.5+0.5)$$
$$= \eta$$
$$\therefore w_1^1 = \eta$$

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

$$= 1 + \eta\left[\frac{1}{1+e} + \frac{1}{1+e}\right]$$

$$= 1 + \frac{2\eta}{1+e}$$
$$\therefore w_2^1 = 1 + \frac{2\eta}{1+e}$$

$$w_3^1 = w_3^0 + \eta[1(1-0.5) - 1(0-0.5))$$

$$= 0 + \eta$$
$$= \eta$$

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

Now we calculate Probabilities.

$$P(Y=1 \mid x, w') = \frac{e^{\eta}}{1+e^{\eta}}$$

$$P(Y=1 \mid x^2, w') = \frac{e^{1+\frac{2\eta}{1+e}}}{1+e^{1+\frac{2\eta}{1+e}}}$$

$$P(Y=1 \mid x^3, w') = \frac{e^{\eta}}{1+e^{\eta}}$$

$$P(y=1|x^4) = P(y=1|x^6) = \frac{e^{-n}}{1+e^{-n}}$$

$$P(y=1|x^5) = \frac{e^{-1-\frac{2n}{1+e}}}{1+e^{-1-\frac{2n}{1+e}}}$$

Now, we calculate $w^2$ using G.A.

$$w_0^2 = w_0^1 + n[0]$$
$$= 0$$

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

$$= n + n\left[\frac{1}{1+e^{n}} + \frac{1}{1+e^{n}}\right]$$

$$= n + \frac{2n}{1+e^{n}} n = n\left[\frac{3+e^{n}}{1+e^{n}}\right]$$

$$= n\left(\frac{3+e^{n}}{1+e^{n}}\right)$$

$$w_2^2 = w_2^1 + n\left[1\left(1-\frac{e^{+1+\frac{2n}{1+e}}}{1+e^{+1+\frac{2n}{1+e}}}\right) - \left(0-\frac{e^{-1-\frac{2n}{1+e}}}{1+e^{-1-\frac{2n}{1+e}}}\right)\right]$$

$$= 1+\frac{2n}{1+e} + n\left[\frac{1}{1+e^{z}} + \frac{1}{1+e^{z}}\right]$$

$$= 1+\frac{2n}{1+e} + \frac{2}{1+e^{z}} \qquad z = 1+\frac{2n}{1+e}$$

$$\therefore w_2^2 = 1+\frac{2n}{1+e} + \frac{2}{1+\exp\left(1+\frac{2n}{1+e}\right)}$$

$w_3^2 = w_1^2$

$= n + \frac{2n}{1+e^n} = \left(\frac{3+e^n}{1+e^n}\right) n$

∴ we have

$w^2 = \left(0, \left(\frac{3+e^n}{1+e^n}\right) n, 1 + \frac{2n}{1+e^n} + \frac{2n}{1+e^{\left(1+\frac{2n}{1+e}\right)}}\right)$

$w^0 = (0, 0, 1, 0)$

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

$w^2 = \left(0, \left(\frac{3+e^n}{1+e^n}\right) n, 1 + \frac{2n}{1+e} + \frac{2n}{1+\exp\left(1+\frac{2n}{1+e}\right)}, \left(\frac{3+e^n}{1+e^n}\right) n\right)$

∴ we see that $w_0^*$ is same for all values

$w_1^0 = w_3^0$ for all values &

$w_1^0 < w_1^1 < w_1^2$ also

$w_2^0 < w_2^1 < w_2^2$ &

$w_3^0 < w_3^1 < w_3^2$

which means w's are constantly increasing to attain a final value

$$w^{*} = \begin{bmatrix} 0 \\ \infty \\ \infty \\ \infty \end{bmatrix}$$

This is similar ~~who~~ in the first ~~case~~ as well.

ie. weights increase till they attain a certain max value.

I think for initial iterations final weight vector will be different for ~~from~~ ~~both~~ different weight initialisations but as we increase iterations they'll approach a max value where the difference will be negligible.

~~wt~~ I coded & observed that for $10^6$ iterations, the difference in the 2nd value ie. $w_2$ for the final vectors was nearly 0.01 only. (Q1 attached)

We can also see that, initially the ~~diff~~ ~~$w_1^0$~~ ~~$w_2^0$~~ =

$w_0^0 = (0,0,0,0)$ & $w_0^1 = (0,0,1,0)$

The difference was (0, 0, 1, 0)

with one iteration

$$|w_1^0 - w_1^1| = (0, \pi, \pi, \pi) - \left(0, \pi, 1 + \frac{2\pi}{1+e}, \pi\right)$$

(1st Iterat)

$$= (0, 0, \approx 0.99, 0)$$

Now the difference is 0.99 which is less than 1

$$|w_2^0 - w_2^1| = \left[0, \frac{(3+e^{\pi})}{1+e^{\pi}}\pi, \frac{(3+e^{\pi})}{1+e^{\pi}}\pi, \frac{(3+e^{\pi})}{1+e^{\pi}}\pi\right]$$

(2nd Iterat)

$$- \left[0, \frac{(3+e^{\pi})}{1+e^{\pi}}\pi, 1 + \frac{2\pi}{1+e} + \frac{2\pi}{1+e\left(1+\frac{2\pi}{1+e}\right)}, \left(\frac{3+e^{\pi}}{1+e^{\pi}}\right)\pi\right]$$

Now the difference is

$$1 + \frac{2\pi}{1+e} + \frac{2\pi}{1+e\left(1+\frac{2\pi}{1+e}\right)} - \left(\frac{3+e^{\pi}}{1+e^{\pi}}\right)\pi$$

$$\approx 0.27 - 0.019$$

$$\approx 0.25$$

which is less than 0.99

& so on,

The difference will keep on decreasing with iterations until a point that it won't matter much.
ie. both values will almost be similar