Name :- Prateek Mahajan

EESII : Homework 3

Problem 1 :-

a) I have read and understood the general instructions at the top of HW3 and I formally declare that all work I turn in for everything in this course will not contain or involve any cheating at all.

Problem 2 :-

a) Consider a binary naive Bayes classifier with multivariate class conditional distributions $p(x|y)$

let class variance be equal.

Bayes Assumption :- $p(x|y) = \prod_{j=1}^{m} p(x_j|y)$

∴ since they are gaussian :-

$$p(x|y) = \prod_{i=1}^{m} \frac{1}{(2\pi\sigma_{y,i}^2)^{1/2}} \cdot \exp\left(\frac{-1}{2}(x_i - \mu_{y,i})^2/\sigma_{y,i}^2\right)$$

for "m" univariate gaussians

As mentioned in the slides :-

~~Log odds ratio~~ for the binary case :-

Posterior probability, $p(y=y_1|x) = \frac{1}{1+\exp\left(-\log\left(\frac{p(x|y=y_1)\cdot p(y=y_1)}{p(x|y=y_0)\cdot p(y=y_0)}\right)\right)}$

$$\Rightarrow p(y=y_1|x) = g\left(\sum_{i=1}^{m} \log\left(\frac{p(x_i|y=1)}{p(x_i|y=0)}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)\right),$$

where the f^n inside g is the log ratio

Consider the term $\frac{p(x_i|y=y_1)}{p(x_i|y=y_0)}$

① $p(x_i|y=y_1) = \frac{1}{(2\pi\sigma^2_{y_1,i})^{1/2}} \cdot \exp\left(\frac{-1}{2}\frac{(x_i-\mu_{y_1,i})^2}{\sigma^2_{y_1,i}}\right)$

② $p(x_i|y=y_0) = \frac{1}{(2\pi\sigma^2_{y_0,i})^{1/2}} \cdot \exp\left(\frac{-1}{2}\frac{(x_i-\mu_{y_0,i})^2}{\sigma^2_{y_0,i}}\right)$

But class variances are equal, $\sigma^2_{y_0,i} = \sigma^2_{y_1,i} = \sigma^2_i \ \forall i$

$$\Rightarrow \frac{p(x_i|y=y_1)}{p(x_i|y=y_0)} = \exp\left(\frac{-1}{2\sigma_i^2}\left((x_i-\mu_{y_1,i})^2 - (x_i-\mu_{y_0,i})^2\right)\right)$$

$$= \exp\left(\frac{-1}{2\sigma_i^2}(2x_i - \mu_{y_1,i} - \mu_{y_0,i})(\mu_{y_0,i} - \mu_{y_1,i})\right)$$

$$= \exp\left(\frac{-1}{\sigma_i^2}\left(x_i - \frac{(\mu_{y_1,i}+\mu_{y_0,i})}{2}\right)(\mu_{y_0,i} - \mu_{y_1,i})\right)$$

$$\therefore \log\left(\frac{p(x_i|y=y_1)}{p(x_i|y=y_0)}\right) = \frac{(\mu_{y_1,i}-\mu_{y_0,i})}{\sigma_i^2}x_i + \frac{\mu^2_{y_0,i} - \mu^2_{y_1,i}}{2\sigma_i^2}$$

Now consider $\log\left(\frac{p(y=1)}{p(y=0)}\right)$

Clearly, the above term is a constant & independent of X

∴ the log ratio

$$\sum_{i=1}^{m} \log\left(\frac{p(x_i|y=1)}{p(x_i|y=0)}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)$$

$$= \sum_{i=1}^{m} \frac{(\mu_{y_1,i}-\mu_{y_0,i})}{\sigma_i^2}x_i + \frac{\mu^2_{y_0,i} - \mu^2_{y_1,i}}{2\sigma_i^2} + \log\left(\frac{p(y=1)}{p(y=0)}\right)$$

$$= b_i x_i + c_i \quad \text{where :—}$$

$$b_i = \frac{(\mu_{y_1,i} - \mu_{y_0,i})}{\sigma_i^2}$$

$$c_i = \frac{\mu^2_{y_0,i} - \mu^2_{y_1,i}}{2\sigma_i^2} + \log\left(\frac{p(y=1)}{p(y=0)}\right)$$

b) If the class variances are unequal, consider eq" ① & ② from 2a).
From ① & ② :—

$$\frac{p(x_i|y=y_1)}{p(x_i|y=y_0)} = \frac{\sigma_{y_0,i}}{\sigma_{y_1,i}} \cdot \exp\left(\frac{-1}{2}\left[\frac{1 \cdot 1}{\sigma_{y_1,i}^2 \cdot \sigma_{y_0,i}^2}\right]\right)$$

$$\exp\left[\sigma_{y_0,i}^2 (x_i^2 + \mu_{y_1,i}^2 - 2\mu_{y_1,i} x_i) + \sigma_{y_1,i}^2 (x_i^2 + \mu_{y_0,i}^2 - 2\mu_{y_0,i} x_i)\right]$$

$$\Rightarrow \frac{p(x_i|y=y_1)}{p(x_i|y=y_0)} = \frac{\sigma_{y_0,i}}{\sigma_{y_1,i}} \cdot \exp\left[\frac{-1 \cdot 1 \cdot 1}{2\sigma_{y_1,i}^2 \sigma_{y_0,i}^2} \cdot \left[x_i^2(\sigma_{y_0,i}^2 - \sigma_{y_1,i}^2) + \right.\right.$$

$$x_i\left[2\mu_{y_0,i}\sigma_{y_1,i}^2 - 2\mu_{y_1,i}\sigma_{y_0,i}^2\right] + \sigma_{y_0,i}^2 \cdot \mu_{y_1,i}^2$$

$$\left.\left. - \sigma_{y_1,i}^2 \cdot \mu_{y_0,i}^2\right]\right]$$

Taking log on both sides :—

$$\log\left(\frac{p(x_i|y=y_1)}{p(x_i|y=y_0)}\right) = \log\left(\frac{\sigma_{y_0,i}}{\sigma_{y_1,i}}\right) + \frac{1}{2\sigma_{y_1,i}^2 \sigma_{y_0,i}^2}\left[x_i^2(\sigma_{y_0,i}^2 - \sigma_{y_1,i}^2) + \right.$$

$$2x_i(\mu_{y_1,i}\sigma_{y_0,i}^2 - \mu_{y_0,i}\sigma_{y_1,i}^2)$$

$$\left. + \sigma_{y_1,i}^2 \mu_{y_0,i}^2 - \sigma_{y_0,i}^2 \mu_{y_1,i}^2\right]$$

Now consider $\log\left(\frac{p(y=1)}{p(y=0)}\right)$

Clearly, as in 2a, this is constant & independent of X.

∴ the log ratio =

$$\sum_{i=1}^{n} \log\left(\frac{p(x_i|y=1)}{p(x_i|y=0)}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)$$

$$= x_i^2\left(\frac{\sigma_{y_1,i}^2 - \sigma_{y_0,i}^2}{2\sigma_{y_1,i}^2\sigma_{y_0,i}^2}\right) + \left(\frac{\mu_{y_1,i}\sigma_{y_0,i}^2 - \mu_{y_0,i}\sigma_{y_1,i}^2}{\sigma_{y_1,i}^2\sigma_{y_0,i}^2}\right)x_i + \left(\frac{\sigma_{y_0,i}^2\mu_{y_1,i}^2}{2\sigma_{y_1,i}^2\cdot\sigma_{y_0,i}^2} - \frac{\sigma_{y_0,i}^2\mu_{y_1,i}^2}{2\sigma_{y_1,i}^2\cdot\sigma_{y_0,i}^2} + \log\left(\frac{\sigma_{y_0,i}}{\sigma_{y_1,i}}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)\right)$$

$$= a_i x_i^2 + b_i x_i + c_i,$$

where :—

$$a_i = \frac{\sigma_{y_1,i}^2 - \sigma_{y_0,i}^2}{2\sigma_{y_1,i}^2\cdot\sigma_{y_0,i}^2}$$

$$b_i = \frac{\mu_{y_1,i}\sigma_{y_0,i}^2 - \mu_{y_0,i}\sigma_{y_1,i}^2}{\sigma_{y_1,i}^2\cdot\sigma_{y_0,i}^2}$$

$$c_i = \frac{\sigma_{y_1,i}^2\mu_{y_0,i}^2 - \sigma_{y_0,i}^2\mu_{y_1,i}^2}{2\sigma_{y_1,i}^2\cdot\sigma_{y_0,i}^2} + \log\left(\frac{\sigma_{y_0,i}}{\sigma_{y_1,i}}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)$$

c) As mentioned in 2a :—

$$p(y=y_1|x) = \frac{1}{1+\exp\left(-\log\left(\frac{p(x|y=y_1)\cdot p(y=y_1)}{p(x|y=y_0)\cdot p(y=y_0)}\right)\right)}$$

$$= g\left(\sum_{i=1}^{m}\log\left(\frac{p(x_i|y=y_1)}{p(x_i|y=y_0)}\right) + \log\left(\frac{p(y=1)}{p(y=0)}\right)\right),$$

where $g = \frac{1}{1+e^{-z}}$ → logistic fn

& z → log ratio of gaussian class conditional distributions

∴ this classifier is essentially a logistic regression in the log ratio of the naive Bayes classifier.

For equal class variance (and assuming g is a logistic f^n): —

$p(y=1|x) = g(\theta^T x + \text{const})$ → linear logistic regression

For unequal class variance: —

$p(y=1|x) = g\left(\sum_{i=1}^{n} a_i x_i^2 + b_i x_i + c_i\right)$

↳ quadratic logistic regression, which can be made linear by consider the x_i^2 terms as new features

Problem 3: —

a) For $n=2$: —

$V_2(r) = \pi r^2$

$S_1(r) = 2\pi r$

For $n=3$: —

$V_3(r) = \frac{4}{3}\pi r^3$

$S_2(r) = 4\pi r^2$

b) In my opinion, the core intuition to remember here is that surface area represents the "boundary" or "skin" of a given volume. "dr" represents a small change in the radius of a volume.

∴ if you change the ~~outer surface of a volume~~ radius of a volume by "dr" the proportional change in volume, "dV" will be the surface area of the volume (especially since $dr \to 0$, which means surface area can be considered to be constant)

∴ $\frac{\text{change in volume (infinitely small)}}{\text{change in radius (infinitely small)}} = \text{Surface Area}$

$\Rightarrow \frac{d}{dr} V_m(r) = \alpha \cdot S_{m-1}(r)$ (eq as $dr \to 0$)

Now take $m = 2$:-

$V_2(r) = \pi r^2$

$\Rightarrow \frac{d}{dr} V_2(r) = 2\pi r = S_1(r)$

For $m = 3$:-

$V_3 = \frac{4}{3} \pi r^3$

$\Rightarrow \frac{d}{dr} V_3(r) = 4\pi r^2 = S_2(r)$

Clearly, this $\forall^n$ holds for $m = 2 \& 3$!!

c) $V_m(r) = K r^m$ (as mentioned in qsn.), where K is a constant

$\therefore S_{m-1}(r) = m \cdot K \cdot r^{m-1} = dV(r)/dr$

If $r = 1$:-

$S_{m-1}(r) = m \cdot K = \bar{S}_{m-1}$

$\Rightarrow K = \frac{\bar{S}_{m-1}}{m}$

$\therefore V_m(r) = \frac{\bar{S}_{m-1}}{m} \cdot r^m$

$\therefore S_{m-1}(r) = \bar{S}_{m-1} \cdot r^{m-1}$

d) $f_m(r) = \int_{x \in S_{m-1}(r)} p(x) \cdot dx$

$p(x) = \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-\|x\|_2^2}{2\sigma^2}\right)$

$\|x\|_2 = r$ & x in $S_{m-1}(r)$

$\therefore f_m(r)$ = probability density of sampled points lying on surface of $S_{m-1}(r)$

$$= \int_{x \in S_{m-1}(r)} p(x) \cdot dx$$

But $p(x) = \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-\|x\|_2^2}{2\sigma^2}\right)$

& since $\|x\|_2 = r$:-

$$p(x) = \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

note that this is because the gaussian is isotropic

$$\therefore f_m(r) = \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) \cdot \int_{x \in S_{m-1}(r)} dx$$

this is essentially $S_{m-1}(r)$

$$= \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) \cdot S_{m-1} \cdot r^{m-1}$$

$$= \frac{1}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) \cdot \overline{S_{m-1}} \cdot r^{m-1}$$

e) using the above $f_m(r)$,

set $\frac{d}{dr} f_m(r) = 0$

$$\frac{d}{dr} f_m(r) = \frac{d}{dr} \frac{\overline{S_{m-1}} \cdot r^{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

$$= \frac{\overline{S_{m-1}}}{(2\pi\sigma^2)^{m/2}} \cdot \left[(m-1) r^{m-2} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) - \frac{2r^m}{2\sigma^2} \exp\left(\frac{-r^2}{2\sigma^2}\right)\right]$$

$$= \frac{\overline{S_{m-1}}}{(2\pi\sigma^2)^{m/2}} \cdot r^{m-2} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) \cdot \left[(m-1) - \frac{r^2}{\sigma^2}\right]$$

Setting $\frac{d}{dr} f_m(r) = 0$

$\Rightarrow (m-1) - \frac{r^2}{\sigma^2} = 0 \Rightarrow r^2 = (m-1)\sigma^2$

$\Rightarrow r = \pm\sqrt{(m-1)\sigma^2}$

But $m >>> 1 \Rightarrow m-1 \simeq m$ (i.e. m is large)

$\Rightarrow r \simeq \pm\sqrt{m}\,\sigma$

Consider $\frac{d^2}{dr^2} f_m(r)$

$= \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right)\left[(m-2)r^{m-3}\cdot\left[(m-1) - \frac{r^2}{\sigma^2}\right] - 3r^{m-1}\frac{}{\sigma^2}\right]$

$= \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot \exp\left(\frac{-r^2}{\sigma^2}\right) r^{m-3}\left[(m-2)(m-1) - \frac{(m+1)r^2}{\sigma^2}\right]$

For $r = \sqrt{m}\,\sigma$, clearly $\frac{d^2}{dr^2} f_m(r) < 0$

Cos $m(m+1)\sigma^2 = m(m+1)$, ~~(as (m-2)) (as (m+1)) mσ²~~ / σ²

which is greater than $(m-2)(m-1)$]

$\therefore r = \sqrt{m}\,\sigma$ is a maximum value.

Clearly $\frac{d}{dr} f_m(r) = 0$ only at $r = \sqrt{m}\,\sigma$

$\Rightarrow f_m(r)$ has a single maximum value at $\hat{r} \simeq \sqrt{m}\sigma$

f) As proven above, $\hat{r} = \sqrt{m}.\sigma$

~~f(x)~~

$f(r) = \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} r^{m-1} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right)$

$\therefore f(\hat{r}+\epsilon) = \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot (\hat{r}+\epsilon)^{m-1} \cdot \exp\left(-\frac{(\hat{r}+\epsilon)^2}{2\sigma^2}\right)$

Now consider $(\hat{r}+\epsilon)^{m-1}$

Using multinomial expansion :—

$(\hat{r}+\epsilon)^{m-1} = {}^{m-1}C_0 \cdot (\hat{r})^{m-1} \cdot \epsilon^0 + {}^{m-1}C_1 \cdot (\hat{r})^{m-2} \cdot \epsilon + \ldots$

$+ {}^{m-1}C_{m-2}\, r^1 \cdot \epsilon^{m-2} + {}^{m-1}C_{m-1} \cdot (\hat{r})^0 \cdot \epsilon^{m-1}$

Clearly, since $\epsilon << \hat{r}$, all terms are much smaller than ${}^{m-1}C_0 \cdot (\hat{r})^{m-1} = (\hat{r})^{m-1}$

$$\therefore (\hat{r}+\epsilon)^{m-1} \simeq (\hat{r})^{m-1}$$

$$\therefore \rho(\hat{r}+\epsilon) \simeq \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot (\hat{r})^{m-1} \cdot \exp\left(\frac{-(\hat{r})^2 - \epsilon^2 - 2\hat{r}\epsilon}{2\sigma^2}\right)$$

$$\simeq \underbrace{\frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot (\hat{r})^{m-1} \cdot \exp\left(\frac{-(\hat{r})^2}{2\sigma^2}\right)}_{\text{this is } \rho(\hat{r})} \cdot \exp\left(\frac{-\epsilon^2 - 2\hat{r}\epsilon}{2\sigma^2}\right)$$

$$\simeq \rho(\hat{r}) \cdot \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\hat{r}\epsilon}{\sigma^2}\right)$$

$$\simeq \rho(\hat{r}) \cdot \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{\sqrt{m}\cdot\epsilon}{\sigma}\right)$$

Using Taylor series expansion:-

$$\exp\left(-\frac{\sqrt{m}\epsilon}{\sigma}\right) = 1 - \frac{\sqrt{m}\epsilon}{\sigma} + \frac{m\epsilon^2}{2!\sigma^2} + \frac{m^{3/2}\epsilon^3}{3!\sigma^3} + \dots$$

But $\epsilon << \hat{r} \Rightarrow \epsilon << \sqrt{m}$ (as σ is constant)

$$\Rightarrow \sqrt{m}\cdot\epsilon \simeq 0.$$

$$\therefore \exp\left(-\frac{\sqrt{m}\epsilon}{\sigma}\right) \simeq 1.$$

$$\Rightarrow \rho(\hat{r}+\epsilon) \simeq \rho(\hat{r}) \cdot \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right) \cdot 1$$

g) If we take a finite ~~sample~~ number of samples from a high dimensional gaussian distribution most of the points would reside at a distance / radius, $r = \sqrt{m}\cdot\sigma$

As shown in 3f), moving away from this point exponentially drops the mass density ρ

probability of sampling. This holds true whether we move from $r = \sqrt{m}\sigma$ towards both the origin & towards infinity.

For a low dimensional gaussian distribution, most points reside near $r_2 = (\sqrt{m-1}).\sigma$

=> For $m = 1$ they reside at the origin
For other low values of m as well, most points would be at small multiples of σ from the origin, and the drop would not be as dramatic as we move from the maxima

h) Probability density $f_m(r) = \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot r^{m-1} \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right)$

$\therefore f_m(0) = 0$

$f_m(r \pm S_{m-1}(\hat{r})) = f_m(\hat{r}) = \frac{S_{m-1}}{(2\pi\sigma^2)^{m/2}} \cdot (\sqrt{m}\sigma)^{m-1} \cdot \exp\left(\frac{-m\sigma^2}{2\sigma^2}\right)$

$= \frac{S_{m-1}}{(2\pi)^{m/2}} \cdot \frac{m^{(m-1/2)}}{\sigma} \cdot e^{-m/2}$

Clearly, probability density = 0 at the origin & is very very high at $r = \hat{r}$ (which is the curse of dimensionality)

i) Please refer to the end of the pdf for the graph & code.

Clearly, the mean of the radius increases as n increases, which is in line with our earlier observations, which show that the mass of the gaussian goes to a higher radius as the dimensionality increases (as depicted by this increase in mean).
Further, it was observed that even when the mass increased, the probability density at the maximum "r" dropped exponentially as you moved away from it in higher dimensions ⇒ the standard deviation should don't increase & be ≃ constant, as is observed in this graph.

Problem 5:-

a) Based on the scatter plot of the question the the decision boundary seems to be a non-linear & non-exponential one. Therefore I believe a low-dimensional linear SVM (note a higher dimensional may work) would not be a great choice which is also the case for logistic regression, as the decision boundary is not ~~ignorable~~ linear in nature.
∴ I would choose the gaussian kernel SVM.

(since it is probably better at representing a non-linear decision boundary, like the provided one)

b) Please refer to the end of this pdf.

c) Please refer to the end of this pdf for the accuracy answers.

The gaussian kernel SVM is better at representing non-linear decision boundaries than the others which are for linear boundaries. More so than anything else, this is a property of their mathematical formulation (which is "linear" in nature)

d) Logistic Regression: Linear Decision Boundary, as it is a linear mathematical formulation (i.e. it relies on a linear combination of input features)

Linear SVM: Linear Decision Boundary as it is a linear mathematical formulation (as it relies on a linear combination of input features)

Gaussian Kernel SVM:- Non linear decision boundary, as is a non-linear mathematical formulation

e) Please refer to the end of this pdf for visualisations. The experiment confirmed my hypothesis.

Problem 4:-

a) Please refer to the end of the pdf for the code & plots.
I took the soln for $\lambda = 25.0$

b) Please refer to the end of this pdf for this qn.

c) Please refer to the end of this pdf for this qn..

In the higher σ case, the initial weights & the change in weights is less consistent & more random.
As a consequence, the zero features take longer to converge to sparsity.
This occurs due to an increase in randomness in data and drop in consistency.

d) For n=50, m=75, λ = 25.0
For n=50, m=150, λ = 50.0
For n=50, m=1000, λ = 70.0
For n=100, m=75, λ = 25.0
For n=100, m=150, λ = 50.0.

I think n = O(m) is most the probable relationship here.

e) Please refer to the end of this pdf file for this qn.