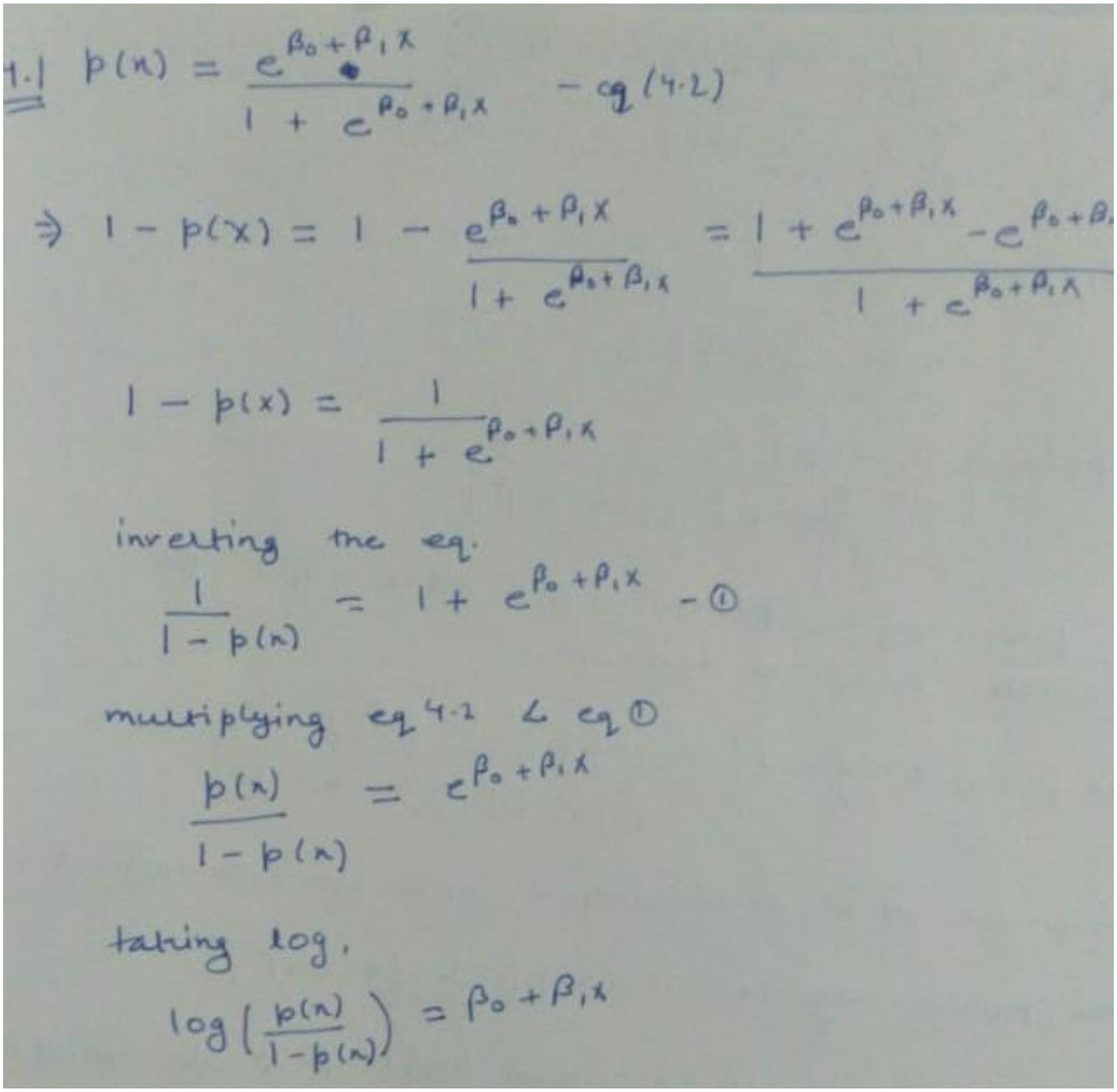
CONCEPTUAL QUESTIONS

1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.



1.1 $p(x) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ — eq (4.2)

$\Rightarrow 1 - p(x) = 1 - \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \dfrac{1+e^{\beta_0+\beta_1 x} - e^{\beta_0+\beta}}{1+e^{\beta_0+\beta_1 x}}$

$1 - p(x) = \dfrac{1}{1+e^{\beta_0+\beta_1 x}}$

inverting the eq.

$\dfrac{1}{1-p(x)} = 1 + e^{\beta_0+\beta_1 x}$ — ①

multiplying eq 4.2 & eq ①

$\dfrac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$

taking log,

$\log\left(\dfrac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x$

2. It was stated in the text that classifying an observation to the class for which (4.12) is largest is equivalent to classifying an observation to the class for which (4.13) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a N(µk, σ2) distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.

4.2

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$p_k(x) = \frac{\pi_k \cancel{\frac{1}{\sqrt{2\pi}\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} \sum_{l=1}^{k} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$p_k(x) = \frac{\pi_k \exp\left(\frac{-1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{k} \pi_l \exp\left(\frac{-1}{2\sigma^2}(x-\mu_l)^2\right)}$$

the func. $\arg\max_k(p_k(x))$ give k as output for which $p_k(x)$ is maximum.

$$\arg\max_k p_k(x) = \arg\max_k \left(\frac{\pi_k \exp\left(\frac{-1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{k} \left(\exp\left(\frac{-1}{2\sigma^2}(x-\mu_l)^2\right)\right)} \right)$$

Since for all k, denominator is same, we will neglect

$$\arg\max_k (p_k(x)) = \arg\max_k \left(\pi_k \exp\left(\frac{-1}{2\sigma^2}(x-\mu_k)^2 \right) \right)$$

{ if $f(x)$ is max at x_0, than $\log(f(x))$ is also max at

$$\arg\max_k (p_k(x)) = \arg\max_k \left(\log \left(\pi_k \cdot \exp\left(\frac{-1}{2\sigma^2}(x-\mu_k)^2 \right) \right) \right)$$

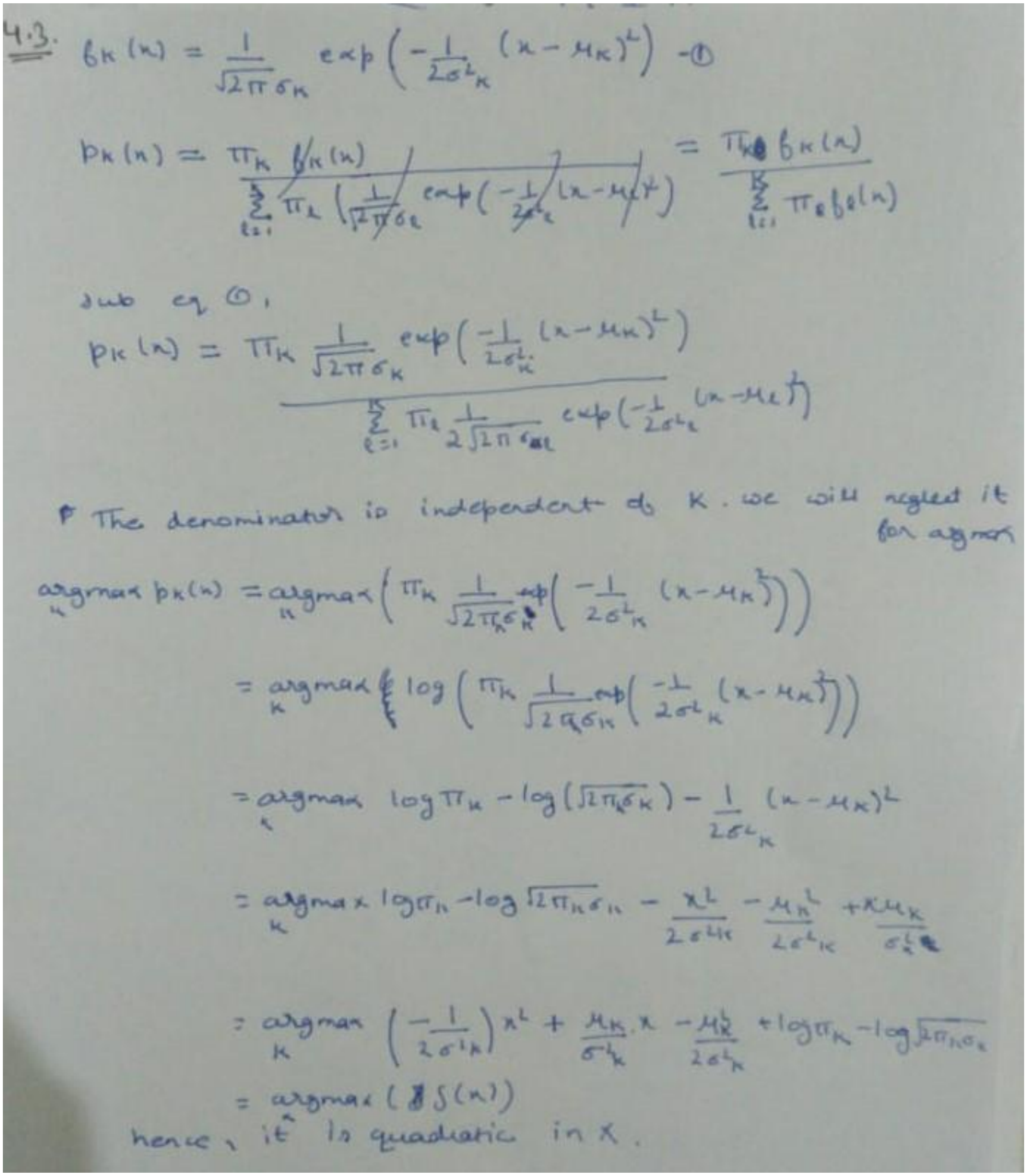
$$\arg\max_k (p_k(x)) = \arg\max \left(\log(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2 \right)$$

$$\arg\max_k (p_k(x)) = \arg\max \left(\log(\pi_k) - \frac{1}{2\sigma^2}\left(\overbrace{\boxed{x^2}}^{\text{neglected}} + \mu_k^2 - x\frac{\mu_k}{\sigma^2} \right) \right)$$

$$\arg\max (p_k(x) = \arg\max \left(x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \right)$$

$$= \arg\max(\delta(x))$$

3. This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a classspecific mean vector and a class specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature. Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, X ~ N(μk, σ2 k). Recall that the density function for the one-dimensional normal distribution is given in (4.11). Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.



4.3. $f_K(x) = \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(-\frac{1}{2\sigma^2_K}(x-\mu_K)^2\right)$ -①

$p_K(x) = \cancel{\frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l \left(\frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right)}} = \frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

sub eq ①,

$$p_K(x) = \frac{\pi_K \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(\frac{-1}{2\sigma^2_K}(x-\mu_K)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{2\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma^2_l}(x-\mu_l)^2\right)}$$

The denominator is independent of K. we will neglect it for argmax

$$\underset{k}{\arg\max}\, p_K(x) = \underset{k}{\arg\max}\left(\pi_K \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(\frac{-1}{2\sigma^2_K}(x-\mu_K)^2\right)\right)$$

$$= \underset{k}{\arg\max}\, \log\left(\pi_K \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left(\frac{-1}{2\sigma^2_K}(x-\mu_K)^2\right)\right)$$

$$= \underset{k}{\arg\max}\, \log \pi_K - \log(\sqrt{2\pi}\sigma_K) - \frac{1}{2\sigma^2_K}(x-\mu_K)^2$$

$$= \underset{k}{\arg\max}\, \log \pi_K - \log \sqrt{2\pi}\sigma_K - \frac{x^2}{2\sigma^2_K} - \frac{\mu_K^2}{2\sigma^2_K} + \frac{x\mu_K}{\sigma^2_K}$$

$$= \underset{k}{\arg\max}\left(-\frac{1}{2\sigma^2_K}\right)x^2 + \frac{\mu_K}{\sigma^2_K}.x - \frac{\mu_K^2}{2\sigma^2_K} + \log \pi_K - \log \sqrt{2\pi}\sigma_K$$

$$= \underset{k}{\arg\max}(\delta(x))$$

hence, it is quadratic in x.

4. When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

(a) Suppose that we have a set of observations, each with measurements on p = 1 feature, X. We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10 % of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0.6, 4.7 Exercises 169 we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?

We would be looking at 10% of the overall data.

(b) Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1, X2) are uniformly distributed on [0, 1] × [0, 1]. We wish to predict a test observation's response using only observations that are within 10 % of the range of X1 and within 10 % of the range of X2 closest to that test observation.

For X1, we would again be looking at the 10%of the data, and out of that 10%, for X2, we will take 10 % of the data.
Overall, the data that we considered for final prediction is 10% of 10%, that is 1% of the total observations.

(c) Now suppose that we have a set of observations on p = 100 features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10 % of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

The answer would be 10% of 10% of 10%.....100 times. Which is 0.10 *0.10*0.10.....100times. (0.10^100)%.

(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

When p is 1, if we consider 10% of the nearest observation for prediction, than we would be looking at 10% of the total observations. However when p is very large, i.e. a data with has very high dimensions, the percentage of observations considered decreases by power of p.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10 % of the training observations. For p = 1, 2, and 100, what is the length of each side of the hypercube? Comment on your answer.

For a given dimension p, a hypercube will be a p dimension figure which contains all the observations taken for consideration for prediction from all of the given observations. For p = 1, hypercube will be a 3 dimensional figure, ie a line. For p=2, it will be two dimensional figure, ie a square. For p = 3, it will be a cube.

For a given dimension p, let the side of hypercube is x, then the area of the hypercube is

$$\textbf{area}= x^p$$

This hypercube will contain 10% of the total observations.

The total volume = 1^p

Volume of hypercube = total volume * fraction of observations contained

For p=1,

$$x^1 = 1^1 * 0.10$$

x = 0.10

for p = 2,

$$x^2 = 1^2 * 0.10$$

$$x = (0.10)^{\frac{1}{2}}$$

$$x = 0.316$$

We can see that the side of the hypercube is 31% of the total length of all observations. X ~ [0,1]

For p = 100,

$$x^{100} = 1^{100} * 0.10$$

$$x = (0.10)^{\frac{1}{100}}$$

$$x = 0.977$$

We can see that as p is increasing , the side of hypercube is also increasing, and hence it is considering more observations for final prediction. Which is opposite to KNN, where, we considered less observations as p increases, and which led to curse of dimensionality.

5. We now examine the differences between LDA and QDA.

(a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

On training set, we expect QDA to have better performance as it is more flexible approach and better fits the training data. On test data, LDA is expected to perform better as the Bayesian decision boundary is linear, which means that QDA is overfitting the data.

(b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

On the training data QDA will perform better as It is more flexible model. On test data, since the Bayesian decision boundary is non linear, a model with high variance, QDA, will fit the non linear relationship better.

(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

For linear models –

For a given n, LDA will perform better on test data as compared to QDA, as QDA overfits the data due to high variance. But, as we increase n, the overfitting on the model decreases, and the performance of QDA increases on test data.

For non linear models –

For a given n, LDA will perform poorer as compared to QDA , as PDA will have high bias, and won't be able to map the non linear relationship. As we increase n, the QDA will better fit the data, and the performance will improve.

For both linear and non linear models, we expect performance of QDA to improve as the sample size n increases.

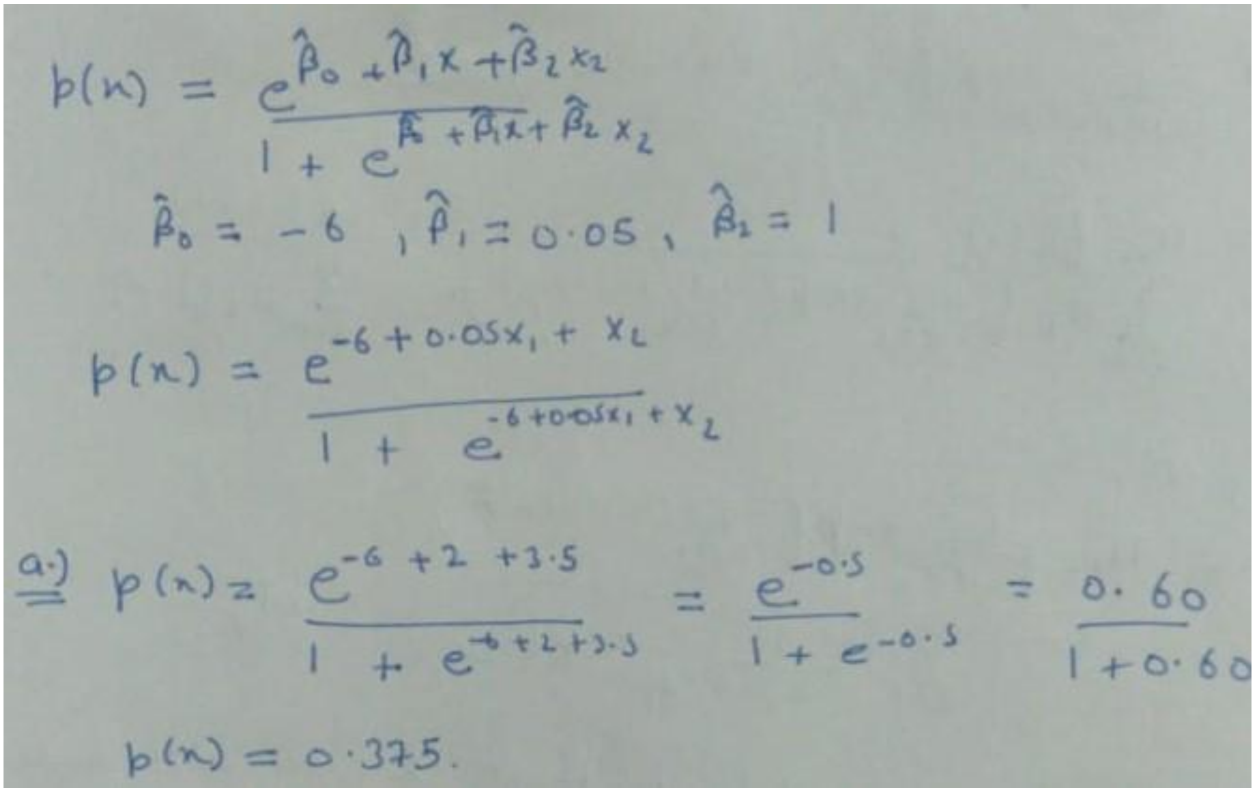
(d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

False.

Although, QDA is flexible enough to model a linear decision boundary, but due to more variance it looks for patterns in random noise and overfits the data. Therefore it will achieve less test accuracy than LDA.

6. Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, βˆ0 = −6, βˆ1 = 0.05, βˆ2 = 1.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.



$$p(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x_2}}$$

$$\hat{\beta}_0 = -6 \,,\ \hat{\beta}_1 = 0.05 \,,\ \hat{\beta}_2 = 1$$

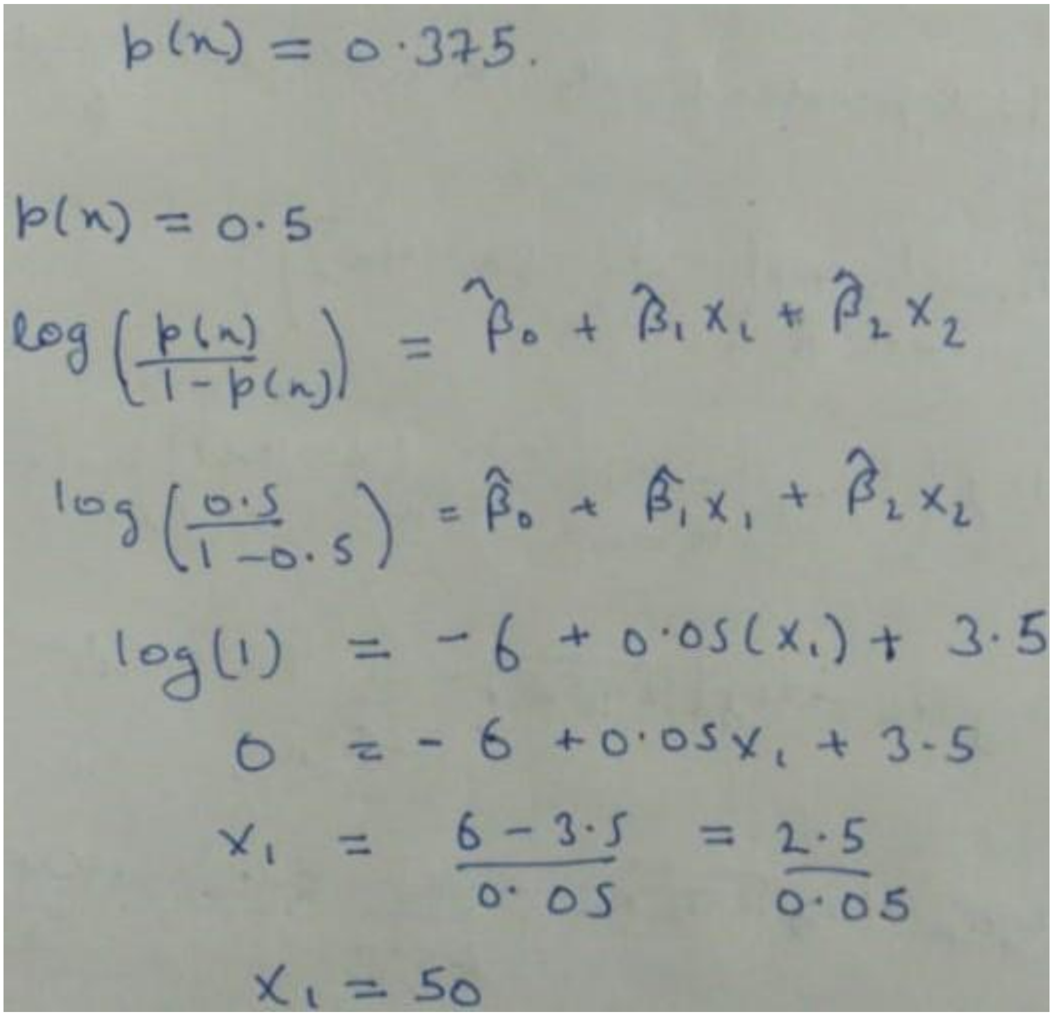
$$p(x) = \frac{e^{-6 + 0.05x_1 + x_2}}{1 + e^{-6 + 0.05x_1 + x_2}}$$

a.) $$p(x) = \frac{e^{-6 + 2 + 3.5}}{1 + e^{-6 + 2 + 3.5}} = \frac{e^{-0.5}}{1 + e^{-0.5}} = \frac{0.60}{1 + 0.60}$$

$$p(x) = 0.375.$$

A student who studies for 40 hours and has a gpa of 3.5 has 37.5% of getting an A.

(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?



$$p(n) = 0.375.$$

$$p(n) = 0.5$$

$$\log\left(\frac{p(n)}{1-p(n)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log\left(\frac{0.5}{1-0.5}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log(1) = -6 + 0.05(x_1) + 3.5$$

$$0 = -6 + 0.05x_1 + 3.5$$

$$x_1 = \frac{6-3.5}{0.05} = \frac{2.5}{0.05}$$

$$x_1 = 50$$

Therefore, student with 3.5 gpa, needs to study 50 hours for getting a 50% chance of getting an A.

7. Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on X, last year’s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was X¯ = 10, while the mean for those that didn’t was X¯ = 0. In addition, the variance of X for these two sets of companies was ˆσ2 = 36. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year.

class $y=1$, dividend.

$$P(y=1) = 0.80$$

$$P(x/y=1) = N(\mu, \sigma^2) = N(10, 36)$$

$$P(x/y=1) = \frac{1}{\sqrt{2\pi \cdot 36}} e^{-\frac{(x-10)^2}{2.36}}$$

class $y=0$,

$$P(y=0) = 0.2$$

$$P(x/y=0) = N(0, 36) = \frac{1}{\sqrt{2\pi . 36}} \times e^{-\frac{(x)^2}{2.36}}$$

By bayes thm,

$$P(y=1/x) = \frac{P(x/y=1)\,P(y=1)}{P(x/y=1)\,P(y=1) + P(x/y=0)\,P(y=0)}$$

$$P(y=1/x) = \frac{\frac{1}{\sqrt{72\pi . \cancel{36}}} e^{-\frac{(x-10)^2}{72}} \times 0.8}{\frac{1}{\sqrt{72\pi \cancel{36}}} e^{\frac{(x-40)^2}{72}} \times 0.8 + \frac{1}{\sqrt{72\pi}} e^{\frac{-(x^2)}{72}} \times 0.2}$$

$$P(y=1/x=4) = \frac{\cancel{\frac{0.8}{1}}\ \frac{e^{-\frac{1}{2}}}{15.033} \times 0.8}{\frac{e^{-\frac{1}{2}}}{15.035} \times 0.8 + \frac{e^{-\frac{2}{9}}}{15.033} \times 0.2}$$

$$P(y=1/x=4) = 0.751$$

Therefore for last year's percent profit = 4, there's a 75.1 % chance that company will issue a dividend this year.

8. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20 % on the training data and 30 % on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18 %. Based on these results, which method should we prefer to use for classification of new observations? Why?

Logistic regression –

Training error = 20%

Test error = 30%

For KNN, we are given the average error rate which is 12%. Since KNN is highly non linear algorithm and for each observation in the training data, it will output the value of the observation only, since it is the closest to itself. Therefore we can assume the training error to be 0.

Since the average error is 18%,

$$\frac{error_{test} + error_{training}}{2} = 18$$

$$error_{test} = 36$$

Test error for KNN is 36%.

Hence we would use Logistic Regression, since It has less test error.

9. This problem has to do with odds.

(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will be in fact default?

$$prob = \frac{odds}{1 + odds}$$

$$prob = \frac{0.37}{1 + 0.37}$$

prob = 0.27

(b) Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?

$$\boldsymbol{odds = \frac{prob}{1 - prob}}$$

$$\boldsymbol{odds = \frac{0.16}{1 - 0.16}}$$

odds = 0.19