classmate

Date ______

Page ______

Exercises - Classification chapter

4) About KNN and curse of dimensionality

when the no. of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction only using observations that are near the test observation for which a prediction must be made.

If X is a uniformly distributed [0,1] feature, p = 1 and we wish to predict a test observation's response using only observations that are within 10% of range closest to that test observation, then the fraction of available observations we'll use to predict is

10% of the range out of [0,1] → 10%

imagine it as a single line from 0 to 1. using only 10% would be the 10% of the length $= 1 \times \frac{10}{100} = 0.1$

extending it to 2D, X & Y axis. so 10% in each axis

$10\% \times 10\% = 0.01 = 1\%$

extending to 100 D, $10\% \times 10\% \times \dots = (10\%)^{100} = 10^{-98}\%$

drawback of KNN when p is large is that, we need to have larger area to cover the small no. of observations.

If we look at the above examples, the available observations near a test observation becomes very small as the dimensions increase. in other words, finding neighbours is difficult / there are no neighbours in a high dimension

length of hypercube to contain 10% of observations in 1D = 0.1

in 2D ⇒ $x^2 = 0.1$ ⇒ $x = \sqrt{0.1}$, in 3D ⇒ $x^3 = 0.1$ ⇒ $x = \sqrt[3]{0.1}$

in 100D ⇒ $x^{100} = 0.1$ ⇒ $x = \sqrt[100]{0.1} \approx 0.98$

when the p is large, it almost uses the entire space to find neighbours

5)
a) If Bayes decision boundary is linear, which performs better, QDA or LDA? in train & test set.

→ QDA performs better in train set, it will be more flexible. LDA performs better in test set, as the QDA might is more flexible & would be overfit

b) if decision boundary is non linear
→ QDA will be better in both train & test set

c) in general, what would happen if sample size n increases?
→ QDA would fit better in both linear & non linear cases as the overfitting would be handled because of the variance that comes with large sample size

d) True or False? even if the Baye's decision boundary is linear, we'll probably achieve superior test error rate using QDA than LDA because QDA is flexible enough to model a linear decision boundary.
→ False. it will overfit.
as the sample size increases, the overfitting is reduced. but in general we still expect LDA to do better since its unbiased & less prone to fit the noise

8) Suppose we take a dataset, divide into equally sized train, test sets. Logistic regression 20% train error rate, 30% test error rate. KNN with k=1, avg error rate of both train & test 18%. Which one to prefer for classifying new observation?
→ avg of both ⇒ train would be even zero. then the test error rate will be as high as 36% which is more than Logistic regression.
further knn with k=1 might be an overfit

classmate

Date

Page

9)

a) on avg what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

→ $odds = \frac{p}{1-p}$ $p = \frac{odds}{1+odds}$

$p = \frac{0.37}{1+0.37} = 0.27$

b) suppose the person has 16% chance of defaulting. what are the odds that she will default?

→ $odds = \frac{0.16}{1-0.16} = 0.19$

7) Prediction whether a given stock would give dividend that year based on company's performance of last year X% profit.

mean $\bar{X}$ of the companies that gave dividend $\bar{X} = 10$

didn't give - $\bar{X} = 0$

variance for both type of companies, $\sigma^2 = 36$

Finally 80% of companies issued dividends.

Assuming X follows a normal distribution, predict the prob that a company will issue dividend this year, given its last year's % profit was 4.

→ (density fn for a normal random variable is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$)

$Pr(X \mid Y = yes) = f_{yes}(x) = N(\mu = 10, \sigma^2 = 36)$

$Pr(X \mid Y = no) = f_{no}(x) = N(\mu = 0, \sigma^2 = 36)$

$Pr(Y = yes) = \pi_{yes} = 0.8$

$Pr(Y = no) = \pi_{no} = 0.2$

$$\pi_{yes}(x) = \frac{\pi_{yes} f_{yes}(x)}{\pi_{yes} f_{yes}(x) + \pi_{no} f_{no}(x)}$$

$$= \frac{0.8e^{\left(-\frac{1}{2\times 36}(4-10)^2\right)}}{0.8e^{\left(-\frac{1}{2\times 36}(4-10)^2\right)} + 0.2e^{\left(-\frac{1}{2\times 36}(4-0)^2\right)}}$$

$$= \underline{\underline{0.75}}$$