SML Homework 3.

Gowresh Gunupati
002647248.

3. $$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\Sigma \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_k)^2\right)}$$

$$\log(P_k(x)) = \frac{\log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + (-)\left(\frac{1}{2\sigma_k^2}\right)(x-\mu_k)^2}{\log\left(\Sigma \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(\frac{1}{-2\sigma_l^2}(x-\mu_l)^2\right)\right)}$$

$$\log(P_k(x)) \log\left(\Sigma \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) + -\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

$$f(x) = \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right) - \frac{1}{2\sigma_k^2}(x-\mu)^2$$

and thus $f(x)$ is quadratic function of x.

4. (a).

on average. 10%. for simplicity, ignoring cases when $X < 0.05$ and $X > 0.95$.

(b)

On average. 1%.

(c).

On average, $0.10^{100} * 100 = 10^{-98}\%$.

(d) As p increases linear, observations that are geometrically near decrease exponentially.

c.

$p=1, \; l=0.10$
$p=2, \; l=\sqrt{0.10}=0.32$
$p=3, \; l=0.10^{1/3}=0.46$

....

$p=N = 0.10^{1/N}.$

5. (a) if the Bayes decision boundary is linear, we expect QDA to outperform on the training set because its greater flexibility will result in a better fit. we expect LDA to outperform QDA on the test set because QDA may overfit the linearity of Bayes decision boundary.

(b) if the Bayes decision boundary is non-linear, we anticipate that QDA will outperform on both the training and test sets.

(c) we predict that the test prediction accuracy of QDA relative to LDA will improve as sample size n increases, because a more flexible method will yield a better fit as more samples can be fit and variance is offset by larger sample sizes

(d) False, with fewer sample points, the variance from using a more flexible method, such as QDA, would lead to overfit, yielding a higher test rate than LDA.

6.

$$p(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1+\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$

X_1 = hours studied, X_2 = undergrad GPA

$\beta_0 = -6, \; \beta_1 = 0.05, \; \beta_2 = 1$

a).

$$X = [40 \text{ hours}, 3.5 \text{ GPA}]$$

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

$$= \frac{\exp(-6 + 0.05 40 + 3.5)}{1 + \exp(-6 + 0.05 40 + 3.5)}$$

$$= \frac{\exp(-0.5)}{1 + \exp(-0.5)}$$

$$= 37.75\%.$$

b)

$$X = [x_1 \text{ hours}, 3.5 \text{ GPA}]$$

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

$$0.50 = \frac{\exp(-6 + 0.05 x_2 + 3.5)}{1 + \exp(-6 + 0.05 x_1 + 3.5)}$$

$$0.50(1 + \exp(-2.5 + 0.05 x_1)) = \exp(-2.5 + 0.05 x_1)$$

$$0.50 + 0.50 \exp(-2.5 + 0.05 x_1)) = \exp(-2.5 + 0.05 x_1)$$

$$0.50 = 0.50 \exp(-2.5 + 0.05 x_1)$$

$$\log(1) = -2.5 + 0.05 x_1.$$

$$x_1 = 2.5 / 0.05 = 50 \text{ hours}.$$

7.

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\Sigma\, \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$P_{yes}(x) = \frac{\pi_{yes} \exp\left(-1/2\sigma^2 (x-\mu_{yes})^2\right)}{\Sigma\, \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$= \frac{\pi_{yes} \exp\left(-1/2\sigma^2 (x-\mu_{yes})^2\right)}{\pi_{yes} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{yes})^2\right) + \pi_{no} \exp\left(\frac{-1}{2\sigma^2}(x-\mu_{no})^2\right)}$$

$$= \frac{0.80 \exp\left(-1/2*36\,(x-10)^2\right)}{0.80 \exp\left(-\frac{1}{2*36}(x-10)^2 + 0.20 \exp\left(-\frac{1}{2*36}x^2\right)\right.}.$$

$$P_{yes}(4) = \frac{0.80 \exp\left(-\frac{1}{2*36}(4-10)^2\right)}{0.80 \exp\left(-\frac{1}{2*36}(4-10)^2\right) + 0.20 \exp\left(-\frac{1}{2*36}4^2\right)} = 75.2\%.$$

8. Given,

logistic regression : 20% training error rate, 30% test error rate KNN (K=1) average error rate of 18%.

for KNN with K=1, the training error rate is 0%. because for any training observation, its nearest neighbour will be the response itself. So, KNN has a test error rate of 36%. I would choose logistic regression because of its lower test error rate of 30%.

9.

(a) $\dfrac{P(x)}{1-P(x)} = 0.37$

$P(x) = 0.37\,(1-P(x))$

$1.37\,P(x) = 0.37$

$P(x) = 0.37 / 1.37 = 27\%$

(b) $\text{odds} = \dfrac{P(x)}{1-P(x)} = 0.16 / 0.84 = 0.19$