(1) It is known that a natural law obeys the quadratic relationship $y = ax^2$. What is the best linear curve that can be used to model this data if all of the data points are drawn uniformly at random in the interval (0,1)?

Let $y = px + q$ be the equation of the linear curve

We need to minimize $L = \int_0^1 (ax^2 - px - q)^2 \, dx$

$$L = \frac{a^2}{5} + \frac{p^2}{3} + q^2 - \frac{2ap}{4} - \frac{2aq}{3} + \frac{2pq}{2}$$

$$\frac{\partial L}{\partial p} = 0 \quad \frac{\partial L}{\partial q} = 0 \Rightarrow \frac{2p}{3} + q = \frac{a}{2} ; \quad 2q + p = \frac{2a}{3}$$

$$\Rightarrow \boxed{p = a , \; q = -\frac{a}{6}}$$

Marking Scheme

Setting up the integral → 1 marks
Getting loss expression → 2 marks
Final solution → 2 marks

(2) Consider the following dataset for text classification where three training instances are given with corresponding classifications into the '+' or –'- category:

Hindi India India	+
India Kannada Hindi	+
Chinese Hindi India	-

Showing all intermediate calculations, find the appropriate classification for the test instance: Chinese Kannada Chinese using the Naïve Bayes text classification algorithm.

First we observe $P(+) = \frac{2}{3}$ and $P(-) = \frac{1}{3}$

$docs_+$ = Hindi India India India Kannada Hindi

$docs_-$ = Chinese Hindi India

vocabulary = {Hindi, India, Kannada, Chinese}

$P(Hindi/+) = \frac{2+1}{6+4} = \frac{3}{10}$ | $P\left(\frac{Hindi}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(India/+) = \frac{3+1}{6+4} = \frac{4}{10}$ | $P\left(\frac{India}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(Kannada/+) = \frac{1+1}{6+4} = \frac{2}{10}$ | $P\left(\frac{Kannada}{-}\right) = \frac{0+1}{3+4} = \frac{1}{7}$

$P(Chinese/+) = \frac{0+1}{6+4} = \frac{1}{10}$ | $P\left(\frac{Chinese}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

Decision $P(+)P\left(\frac{Chinese}{+}\right)P\left(\frac{Kannada}{+}\right)P\left(\frac{Chinese}{+}\right)$

vs

$P(-)\,P\left(\frac{Chinese}{-}\right)P\left(\frac{Kannada}{-}\right)P\left(\frac{Chinese}{-}\right)$

$\frac{2}{3} \times \frac{1}{10} \times \frac{2}{10} \times \frac{1}{10}$ vs $\frac{1}{3} \times \left(\frac{2}{7}\right)^2 \times \frac{1}{7}$

0.00132 vs 0.0038

Marking Scheme:

Positive Conditional Probabilities → 1.5 mark

Negative Conditional Probs → 1.5 marks

Decision - 2 marks

Q.3. There are two varieties of cucumbers – C_1 and C_2 which have different distributions of length. The joint probability density function of the length of the cucumber and category 1 is denoted by $p(x, C_1)$, and is a uniform distribution over the range (10cm, 30cm). Similarly $p(x, C_2)$ is a uniform distribution over the range (20cm, 50cm). What is the error of classification we will make if we assert that all cucumbers of length less than 25cm are of Variety 1 and all cucumbers of length greater than 25cm are of Variety 2?

[5 Marks]

The information in this question is <u>incompletely specified</u>

The functions given for $p(x, C_1)$ and $p(x, C_2)$ in the question are actually $p(x/C_1)$ and $p(x/C_2)$ respectively. To get $p(x, C_1)$ and $p(x, C_2)$ we need to multiply $p(x/C_1)$ and $p(x/C_2)$ by $p(C_1)$ and $p(C_2)$ respectively. <u>$p(C_1)$ and $p(C_2)$ are not specified in the question</u>, <u>so $p(\text{mistake})$ should be calculated in terms of $p(C_1)$ and $p(C_2)$</u>.

We have

$$p(\text{mistake}) = \int_{R_1} p(x, C_2)\,dx + \int_{R_2} p(x, C_1)\,dx$$

$$= p(C_2)\int_{R_1} p(x/C_2)\,dx + p(C_1)\int_{R_2} p(x, C_1)\,dx$$

$$= p(C_2)\int_{20}^{25} p(x/C_2)\,dx + p(C_1)\int_{25}^{30} p(x, C_1)\,dx$$

$$= p(c_2)\frac{5}{30} + p(c_1)\frac{5}{20}$$

$$\boxed{= p(c_2)\frac{1}{6} + p(c_1)\frac{1}{4}}$$

Marking Scheme

Formula for p(mistake) = 3 marks

Final Calculation = 2 marks

(3) Consider the standard set of Gaussian Naïve Bayes assumptions used in the derivation of the logistic regression expression, but with one modification – the class conditional density for each class has unique values for both the mean and variance, rather than a common value for the variance, i.e $P(X_i/Y = y_k) = N(\mu_{ik}, \sigma_{ik})$. Find the expression for $P(Y = 1/X_1, X_2, \dots X_n)$ in this case and find the decision boundary.

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

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

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

$$= \frac{1}{1 + \exp(\ (\ln \frac{1-\pi}{\pi}) + \sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)})}$$

Now $\ln \frac{P(X_i/Y=0)}{P(X_i/Y=1)} = -\frac{1}{2}\left(\frac{X_i - \mu_{i0}}{\sigma_{i0}}\right)^2 + \frac{1}{2}\left(\frac{X_i - \mu_{i1}}{\sigma_{i1}}\right)^2$

$$= \frac{1}{2}\frac{X_i^2}{\sigma_{i1}^2} - \frac{1}{2}\frac{X_i^2}{\sigma_{i0}^2} - \frac{\mu_{i1} X_i}{\sigma_{i1}} + \frac{\mu_{i0} X_i}{\sigma_{i0}}$$

$$+ \frac{\mu_{i0}^2}{\sigma_{i0}^2} - \frac{\mu_{i1}^2}{\sigma_{i1}^2}$$

This will give rise to an expression of the form $P(Y=1/X) = \frac{1}{1 + \exp(\omega_0 + \sum \omega_i X_i + \alpha X_i^2)}$

where $\alpha = \frac{1}{2}\left(\frac{1}{\sigma_{i1}^2} - \frac{1}{\sigma_{i0}^2}\right)$

The decision boundary is quadratic instead of linear

Marking Scheme

Derivation for the new expression = 4 marks

Decision boundary quadratic → 1 mark

-

(4) Consider the following dataset

price	maintenance	capacity	Safety measures	Beneficial
lowpriced	cheap	5	yes	yes
lowpriced	average	5	yes	yes
lowpriced	cheap	5	yes	no
lowpriced	excessive	3	no	no
fair	average	5	no	no
fair	average	5	no	yes
fair	excessive	3	yes	no
fair	excessive	6	yes	yes
overpriced	average	5	yes	yes
overpriced	excessive	3	yes	no
overpriced	excessive	6	yes	no

Classify the new instance given: "price = fair, maintenance = cheap, capacity = 5, safety measures = yes". Use Laplace smoothing only when needed for an attribute.

We have. $P(\text{Beneficial} = \text{Yes}) = \frac{5}{11}$

$P(\text{Beneficial} = \text{No}) = \frac{6}{11}$

$P\left(\frac{\text{Price} = \text{fair}}{\text{Yes}}\right) = \frac{2}{5}$ | $P\left(\frac{\text{Price} = \text{fair}}{\text{no}}\right) = \frac{2}{6}$

$P\left(\frac{\text{maint} = \text{cheap}}{\text{Yes}}\right) = \frac{1}{5}$ | $P\left(\frac{\text{maint} = \text{cheap}}{\text{no}}\right) = \frac{1}{6}$

$P\left(\frac{\text{capacity} = 5}{\text{Yes}}\right) = \frac{4}{5}$ | $P\left(\frac{\text{capacity} = 5}{\text{no}}\right) = \frac{2}{6}$

$P\left(\frac{\text{Safety} = \text{Yes}}{\text{Yes}}\right) = \frac{4}{5}$ | $P\left(\frac{\text{safety} = \text{Yes}}{\text{no}}\right) = \frac{4}{6}$

Compare $P(\text{Yes})\, P\left(\frac{\text{fair}}{\text{yes}}\right) P\left(\frac{\text{cheap}}{\text{yes}}\right) P\left(\frac{5}{\text{yes}}\right) P\left(\frac{\text{yes}}{\text{yes}}\right)$

with $P(\text{No})\, P\left(\frac{\text{fair}}{\text{No}}\right) P\left(\frac{\text{cheap}}{\text{no}}\right) P\left(\frac{5}{\text{no}}\right) P\left(\frac{\text{Yes}}{\text{no}}\right)$

$$\frac{5}{11} \times \frac{2}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{4}{5} \text{ vs } \frac{6}{11} \times \frac{2}{6} \times \frac{1}{6} \times \frac{2}{6} \times \frac{4}{6}$$

$\underline{0.023}$ vs 0.0067

The instance should be classified as <u>Yes</u>

Marking Scheme

$P(Yes)$, $P\left(\frac{X_i}{Yes}\right) \rightarrow$ 2 Marks

$P(No)$, $P(X_i / No) \rightarrow$ 2 Marks

final decision $\rightarrow$ 1 Mark