QBUS6810
Statistical Learning and Data Mining

Tutorial 13 (Written Exercises)

Question 1

Suppose that you want to train a Naive Bayes classifier for the following data.

Inputs		**Output**
Salary (X_1)	Salary of the spouse (X_2)	Happy (H)
80	60	Yes (1)
100	80	Yes (1)
70	80	No (0)
50	20	No (0)

The unit of measurement for Salary is \$1K.

(a) Assume that the distribution of Salary given the value of H is Gaussian, i.e. the class conditional densities for X_1 are:

$$p(x_1|H=h) = \frac{1}{\sqrt{2\pi\sigma_{1h}^2}} \exp\left(-\frac{(x_1-\mu_{1h})^2}{2\sigma_{1h}^2}\right), \quad \text{for } h=0 \text{ and } h=1.$$

Estimate (using the data) the means μ_{1h} and the variances σ_{1h}^2 for $h=0$ and $h=1$. Show your work.

(b) Assume that the distribution of the Salary of the spouse given the value of H is also Gaussian, i.e. the class conditional densities for X_2 are:

$$p(x_2|H=h) = \frac{1}{\sqrt{2\pi\sigma_{2h}^2}} \exp\left(-\frac{(x_1-\mu_{2h})^2}{2\sigma_{2h}^2}\right), \quad \text{for } h=0 \text{ and } h=1.$$

Estimate (using the data) the means μ_{2h} and the variances σ_{2h}^2 for $h=0$ and $h=1$. Show your work.

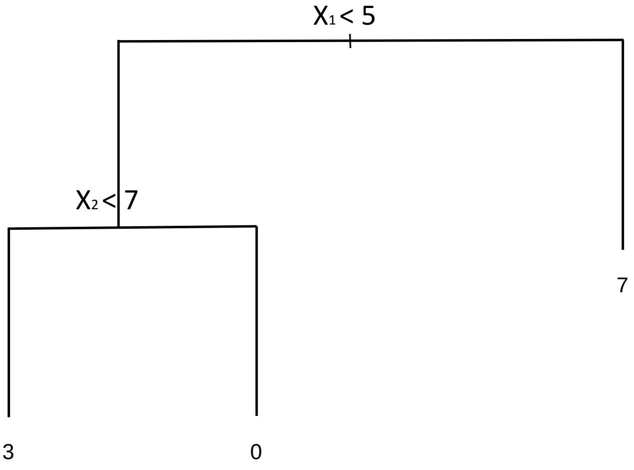
(c) Estimate (using the data) the class probabilities: $P(H=1)$ and $P(H=0)$.

(d) Derive the Naive Bayes classifier estimate of the probability $P(H=1|X_1=x_1, X_2=x_2)$, first using the general formulas, and then incorporating your specific results from parts (a), (b), and (c).

Question 2

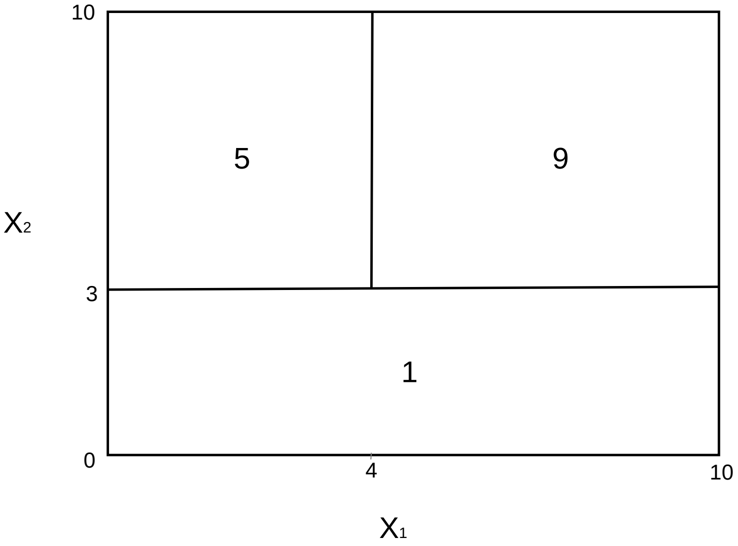
Suppose X_1 and X_2 take values in the interval $[0, 10]$.

(a) Consider the following regression tree. The numbers next to the terminal nodes give the corresponding predicted values.



Sketch the corresponding partition of the predictor space. Write the predicted values inside the corresponding regions of the partition.

(b) Now consider the following partition of the predictor space.



Create a diagram for the corresponding regression tree.

(c) For the tree in part (b), write down the estimated regression function, $\widehat{f}(x_1, x_2)$. Does this function correspond to a GAM (i.e. an additive model) or a model with interactions?