Homework 4

ST 697 / Fall 2017

Due: Tue Nov 7

Problem 4.1: Additive Bayes

Consider a binary classification setting with class labels $Y \in \{0, 1\}$ and predictors $X \in \mathbb{R}^p$. Use the following notation to answer the problems: $p(x) = \Pr(Y = 1|X = x)$, $\pi_k = \Pr(Y = k)$, $\gamma(x) = \log \frac{p(x)}{1-p(x)}$, and use $f_k(x)$ to denote the density function for class k.

a. Use Bayes theorem to write $\gamma(x)$ as a function of the class conditional pdfs.

For the next problems consider the general class of additive models:

$$\gamma(x) = \alpha + \sum_{j=1}^{p} g_j(x)$$

Specify α and $g_j(x)$ for the following specific models:

b. logistic regression

c. linear discriminant analysis (LDA)

d. naive Bayes

e. B-splines (one for each predictor variable)

Problem 4.2: Kernel Density Estimator

Note that R's `density()` does not allow you to specify the evaluation points. This exercise will rectify this problem. Write a function to calculate a kernel density estimate in 1 dimension.

- Inputs should be data, bandwidth, evaluation points
- Use a Gaussian kernel with bandwidth defined as the standard deviation
- The output should be a list including: the evaluation points, the density estimate at the evaluation points, and the bandwidth.
- Check out your function. Consider the observations `c(-1,0,.5,3)` and evaluation at `c(0,.5,1)`. Using a bandwidth of `1`, the kernel density estimate should be `0.249, 0.225, 0.176`.

Problem 4.3: Estimating the log density ratio: derivation

In the naive Bayes classification model, the multivariate log density ratio is estimated as a sum of univariate estimates. This problem will explore some different ways to estimate the log density ratio for a univariate predictor. Specifically, we will consider

$$\log \frac{\hat{f}_1(x)}{\hat{f}_0(x)}$$

where each component density $f_k, k = 0, 1$ is estimated independently and x is a scalar.

For each scenario: write out the form of the density function f_k, how to estimate the parameters, and show the form of the estimated log density ratio.

a. $x \in \mathbb{R}$. Model f_k as a Gaussian density and estimate the parameters using the maximum likelihood approach (MLE).

b. $x \in \mathbb{R}$. Use kernel density estimation (kde) to estimate the densities. Explain how you would select the bandwidth for each component.

c. $x \in \{A, B, C, D\}$ (categorical rv). What would you do if there are no observations in a category?

d. $x \in [0, \infty]$ (boundary). Suppose there is a boundary at zero (e.g., x measures income or height). Describe a strategy to estimate the log density ratio in this scenario or argue why it is not necessary to make any adjustments to your solutions for part (a) or (b). (You do not have to implement anything.)

Problem 4.4: Estimating the log density ratio: implementation

Now to implement the estimation models from Problem 4.3. Use the `Default` data in the `ISLR` R package to estimate the following log density ratios (use `default==Yes` as class 1). Produce plots of the component densities and the log density ratio.

a. Use the Gaussian model to estimate the log density ratio for the `balance` variable.

b. Use the kde model to estimate the log density ratio for the `balance` variable. Your choice of bandwidth.

c. Use the categorical data model to estimate the log density ratio for the `student` variable.

d. Convert your results from part (a) into an estimate of $p(x) = \Pr(Y = 1|X = x)$ and produce a plot.