Kernel density classification

Patrick Breheny

October 25

The classification problem

- In addition to providing estimates of density, kernel density methods may also be used for *classification*
- Suppose x is continuous, but that y is discrete, and can take values in K different categories
- Given a sample of n pairs of observations $\{x_i, y_i\}$, we would like to obtain an method for estimating $\mathbb{P}(y_i = j|x_i)$ in future observations for which x is observed but y is not

Kernel density classification

This can be accomplished in a straightforward fashion using kernel density estimation and Bayes' theorem:

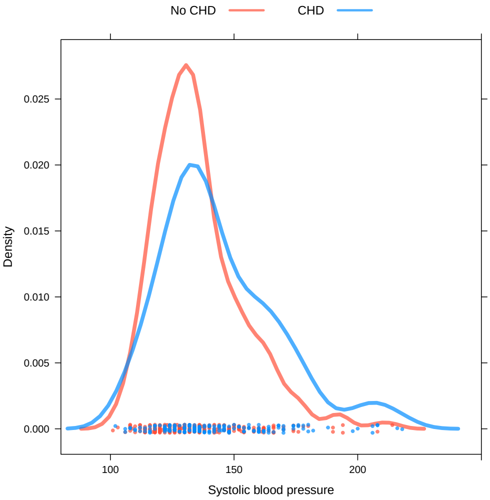
$$\hat{\mathbb{P}}(y = j|x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^K \hat{\pi}_k \hat{f}_k(x_0)}$$

- $\hat{\pi}_j$ is an estimate of the prior probability of class j; usually, $\hat{\pi}_j$ is the sample proportion falling into the jth category
- $\hat{f}_j(x_0)$ is the estimated density at x_0 based on a kernel density fit involving only observations from the jth class
- This is essentially the same idea as discriminant analysis, only instead of assuming normality, we are estimating the probability density of the classes using a nonparametric method

Coronary heart disease study

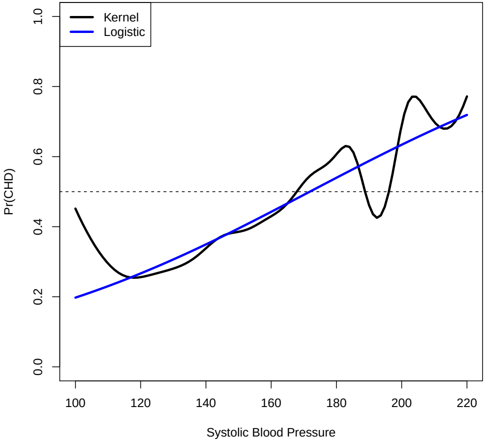
- Let us consider a study of coronary heart disease (CHD)
- The study looked at many potential risk factors for CHD, such as blood pressure, tobacco and alcohol consumption, age, family history, etc.
- One goal of the study is to try to asses the probability of developing coronary heart disease, given that a person has certain risk factors
- In this lecture, we will focus on systolic blood pressure as a risk factor

Kernel density estimates



Estimate of posterior probability

In the sample, $\hat{\pi}_{CHD} = .346$



Evaluation

- As we can see, the kernel density classifier is not restricted to a linear function, although it seems somewhat unstable in regions where there is little data
- As we have seen, there will be many regions with little data when we move to higher dimensions

The independence assumption

- Thus, the simplifying assumption of independence is often made:

$$\hat{f}_j(x) = \prod_{k=1}^{K} \hat{f}_{jk}(x_k),$$

  where $\hat{f}_{jk}$ is an estimate of the density of the jth class in the kth dimension
- This assumption is, generally speaking, not true
- However, it drastically simplifies the estimation and alleviates the curse of dimensionality by allowing the class-specific marginal densities f_{jk} to be estimated with one-dimensional kernel methods

The Naive Bayes Classifier

- This approach is called the *naive Bayes classifier*
- It is not necessarily a good way to estimate $\hat{f}_j(x)$, but in practice, it often performs well as a classifier
- The reason for this is that, although the estimator has considerable bias, the savings in variance are tremendous
- Furthermore, a bad estimate for f_j does not necessarily imply that the estimate $\mathbb{P}(y = j|x)$ is bad

Connection to additive models

- Finally, it is not hard to show that, for the naive Bayes classifier,

$$\text{logit}(y = 1|\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} g_k(x_k)$$

- Thus, the naive Bayes classifier is equivalent to a certain sort of additive (*i.e.*, no interactions) logistic regression model, with flexible functions g_k determining the impact of x_k on the log-odds that $y = 1$
- For the rest of the course, we will take a more direct role in this regression/classification problem by starting with the above model and estimating the functions g directly