Estimating a Probability Density

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

1. Reading
2. Estimating a Probability Density
3. Handling Continuous Predictors in Naive Bayes

Reading

The reading for density estimation is Hastie, Tibshirani, and Friedman Section 6.6 (2nd edition; latest printing at: http://statweb.stanford.edu/∼tibs/ElemStatLearn/)
Note: this reading uses slightly different notation than we use in class, but you should be able to follow it.

Estimating a Probability Density

- Say we have iid observations $X_i \overset{\text{iid}}{\sim} f$ for $i = 1, \ldots, n$ of a continuous random variable, from some unknown probability density function f defined on $\mathbb{R}$. Call the observed values $x_1, \ldots, x_n$.
- How would we estimate the pdf f?
- This problem is called **density estimation**.
- It is a type of unsupervised learning. Why?
-

Estimating a Probability Density

Uses:

- Predicting the distribution of a new observation X^*
    -
-

Estimating a Probability Density

- Let $\phi_{\mu,\sigma^2}(z)$ be the normal density with mean μ and standard deviation σ (so that, e.g., $\phi_{0,1}(z)$ is the standard normal density):

$$\phi_{\mu,\sigma^2}(z) =$$

- **Kernel density estimation**: for some fixed value of σ, estimate $f(\cdot)$ to be

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n}\phi_{x_i,\sigma^2}(z) \qquad \forall z \in \mathbb{R}$$

i.e.:

Estimating a Probability Density

σ is called the "bandwidth", and for now think of it as being a value that we have to specify.

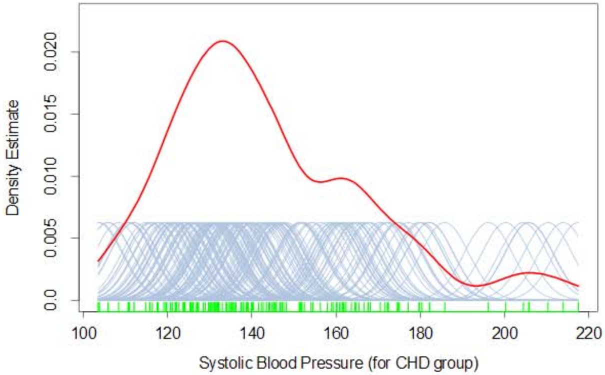
Sketch $\hat{f}$ when $\sigma = 1$ and the observations are $x_1 = -2.2$, $x_2 = -1.2$, $x_3 = 3.0$, and $x_4 = 5.1$:

Estimating a Probability Density

Idea: create a smooth function that is high in locations close to many observations, and low in locations that are not close to many observations

Estimating a Probability Density

Here is a density estimate of systolic blood pressure for individuals with coronary heart disease (HTF text):



Density Estimation in $p > 1$ dimensions

- Kernel density estimation generalizes naturally to the case where X_i has dimension $p \geq 1$.
- Still have $X_i \overset{\text{iid}}{\sim} f$; now $f(\cdot)$ is an unknown positive function in p dimensions that we want to estimate
- Still want to take an average over the data of some smooth function in p dimensions that is highest at x_i and symmetric about x_i. A natural choice is:

Density Estimation in $p > 1$ dimensions

So the estimate of f is

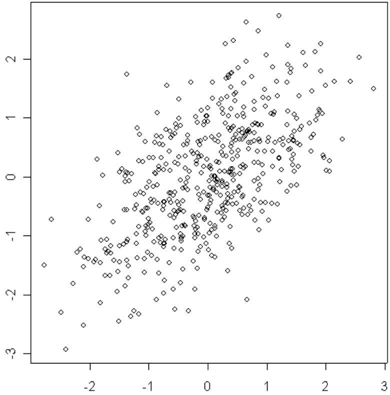
$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} \phi_{p,x_i,\sigma^2}(z) \qquad z \in \mathbb{R}^p.$$

Warning: kernel density estimation does not scale well with p, in the sense that

In practice it's only effective for very small p; most commonly used for $p \leq 3$.

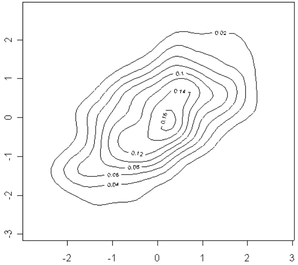
Continuous Predictors in Naive Bayes

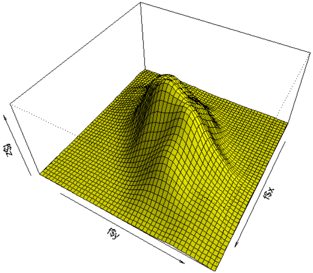
Example: say you have the following observations $\{x_i : i = 1, \ldots, n\}$ in $p = 2$ dimensions.



Continuous Predictors in Naive Bayes

Here is the density estimate, both a contour plot and a 3D plot:





Continuous Predictors in Naive Bayes

- So far we only know how to handle categorical predictors in naive Bayes.
- We estimate the marginal dist'n $\Pr(Y = y)$ and the conditional distribution $\Pr(X_k | Y = y)$ for each predictor k and value y.
- For a continuous predictor we still do this, but instead of estimating the probability $\Pr(X_k = x_k | Y = y)$ for each possible value x_k (this is now 0), we estimate **the probability density of X_k, conditional on** $Y = y$ for each y.

Continuous Predictors in Naive Bayes

- Let $i = 1, \ldots, n$ index the training samples $(x_{i1}, \ldots, x_{ip}, y_i)$.
- For some fixed value of σ, estimate the (one-dimensional) conditional density of X_k by the kernel density estimate

$$f_{X_k \mid Y=y}(z) =$$

where $n_y \triangleq \sum_{i=1}^{n} \mathbf{1}_{\{y_i = y\}}$.

Continuous Predictors in Naive Bayes

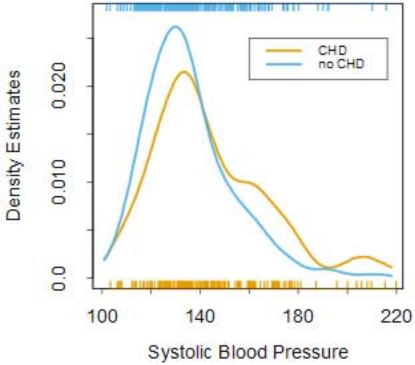
To do prediction with mixed continuous & discrete predictors, use the formula (same derivation as before)

$$\Pr(Y = y | X_1 = x_1, \ldots, X_k = x_k) =$$

where $f_{X_k|Y=y}(x_k)$ is the conditional density estimate if X_k is continuous, and the estimated conditional probability $\Pr(X_k = x_k | Y = y)$ if X_k is discrete.

Continuous Predictors in Naive Bayes

Example: Predicting Coronary Heart Disease (CHD: Yes / No) based on the systolic blood pressure.
Train N.B.: Estimate the marginal dist'n $\Pr(Y = \text{Yes})$, $\Pr(Y = \text{No})$ and the conditional densities $f_{X_1 \mid Y=\text{Yes}}(z)$ and $f_{X_1 \mid Y=\text{No}}(z)$ (figure from HTF):

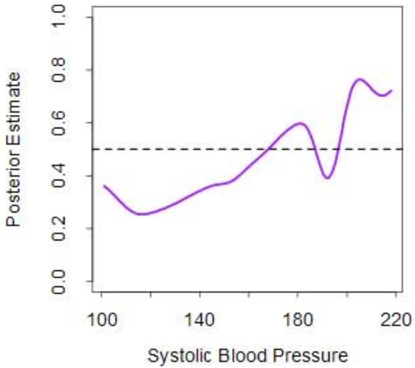


Continuous Predictors in Naive Bayes

N.B. prediction formula for CHD example:

$\Pr(Y = \text{Yes}|X_1 = x_1) =$

Plotting $\Pr(Y = \text{Yes}|X_1 = x_1)$ as a function of x_1 (figure from HTF):



Continuous Predictors in Naive Bayes

Does systolic blood pressure alone provide very certain prediction of heart disease status?

Continuous Predictors in Graphical Models

How to add an edge between two continuous predictors X_j and X_k, to relax the conditional independence assumption (creating a graphical model)? Estimate the conditional density of (X_j, X_k) given $Y = y$ for each y:

$$f_{X_j,X_k\,|\,Y=y}(w,z) = \frac{1}{n_y}\sum_{i=1}^{n} \phi_{2,(x_{ij},x_{ik}),\sigma^2}(w,z)\mathbf{1}_{\{y_i=y\}} \qquad w,z \in \mathbb{R}$$

Then the predicted probability that $Y = y$ for a new observation, where the predictors can be mixed continuous & discrete, is (by same calculation as before)