Density Estimation and Classification

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Density estimates and classifiers

The problem of describing structure for $\mathbf{x} \in \Re^p$ might be phrased in terms of estimating a pdf for the variable. So the problem:

$$\text{based on } \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \ \text{ iid with (unknown) pdf } \ q, \text{ estimate } q$$

is of independent interest. But of present importance is the fact that an optimal 0-1 loss classifier is for $\mathbf{x} \in \Re^p$ a k maximizing

$$\pi_k p(\mathbf{x}|k)$$

and if one can estimate each $p(\cdot|k)$ based on the part of a training sample with $y = k$ (and approximates each π_k with the fraction of the training sample with $y = k$) an approximately optimal classifier can potentially be made.

Parzen kernel density estimates

Temporarily suppose that $p = 1$. For $\phi(\cdot)$ the standard normal pdf (other choices of basic "kernel" are possible, but this is most common) and a "bandwidth" $\lambda > 0$,

$$\frac{1}{\lambda}\phi\left(\frac{\cdot - \theta}{\lambda}\right)$$

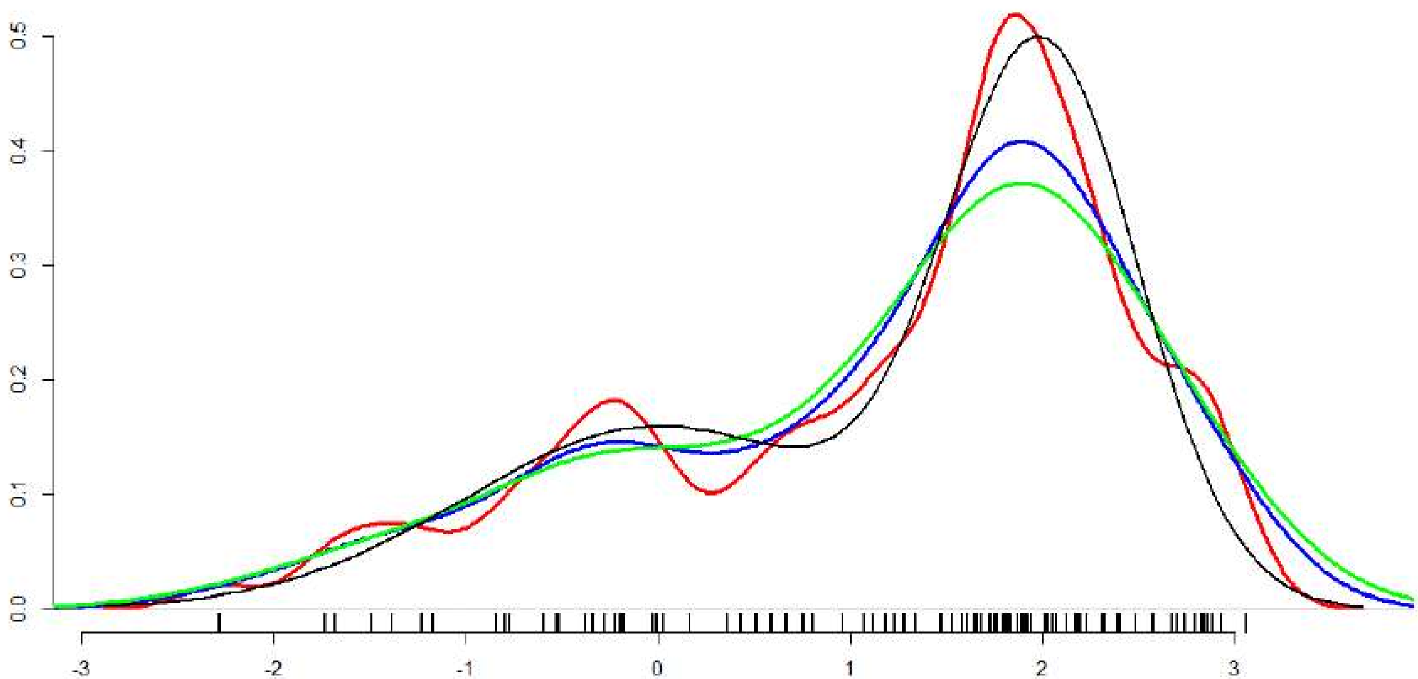
is the normal density for mean θ and standard deviation λ. The Parzen (kernel) estimate of a density at x, $q(x)$, is then

$$\hat{q}_\lambda(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda}\phi\left(\frac{x - x_i}{\lambda}\right)$$

an average of values of normal densities centered at the x_i in a training set.

A p=1 example

Below are plots of a pdf, q (in black), a sample of size $N = 100$ from the distribution and (normal kernel) density estimates made with bandwidths $\lambda = .2$ (red),.4 (blue), and .5 (green).



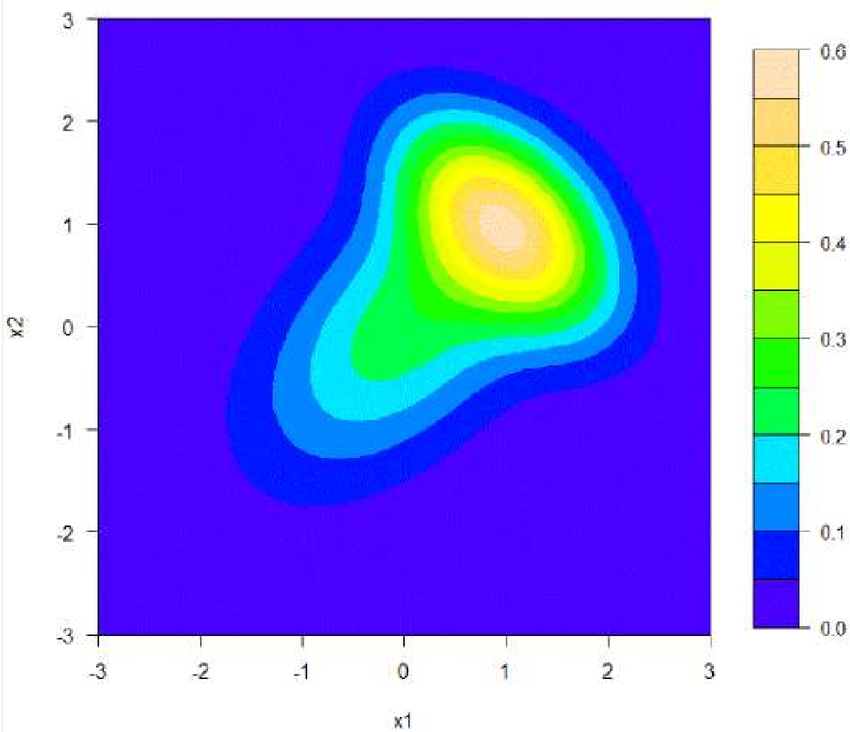
Normal kernels in density estimation

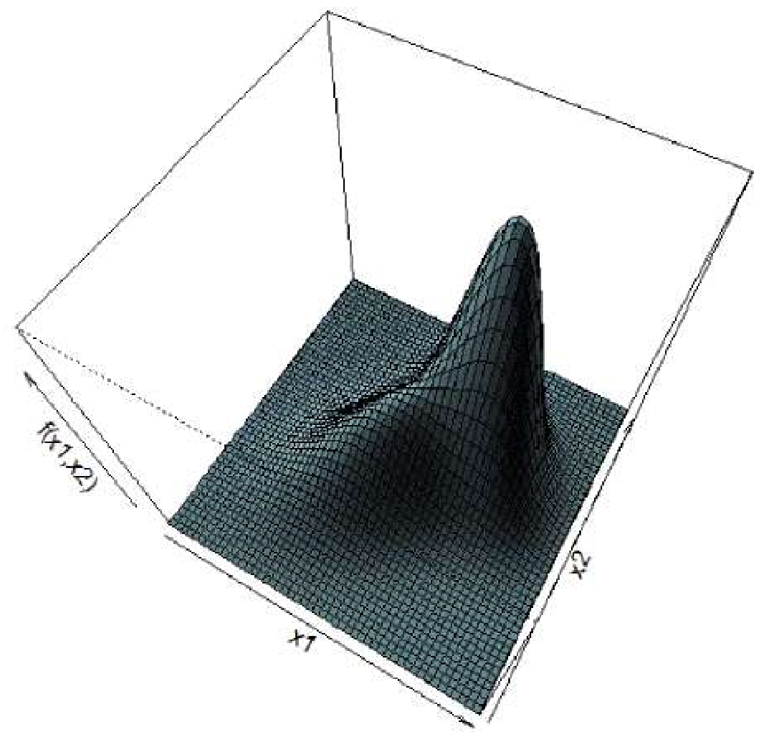
A density estimate that results from using a normal kernel represents the distribution of "a random choice from the training set perturbed by a mean 0 normal error with standard deviation equal to the bandwidth." If the bandwidth is extremely small, the density estimate will essentially consist of "spikes" at the x_i in the training set. If it is extremely large, the density estimate will essentially consist of a normal density centered around the mean of the x_i. Useful bandwidths will be neither extremely small nor extremely large.

A natural generalization of this to p dimensions is to let $\phi(\cdot)$ be a (mean $\mathbf{0}$) MVN_p density. One should expect that unless N is huge, this methodology will be reliable only for fairly small p (say 3 at most) as a means of estimating a general p-dimensional pdf.

An example p=2 density

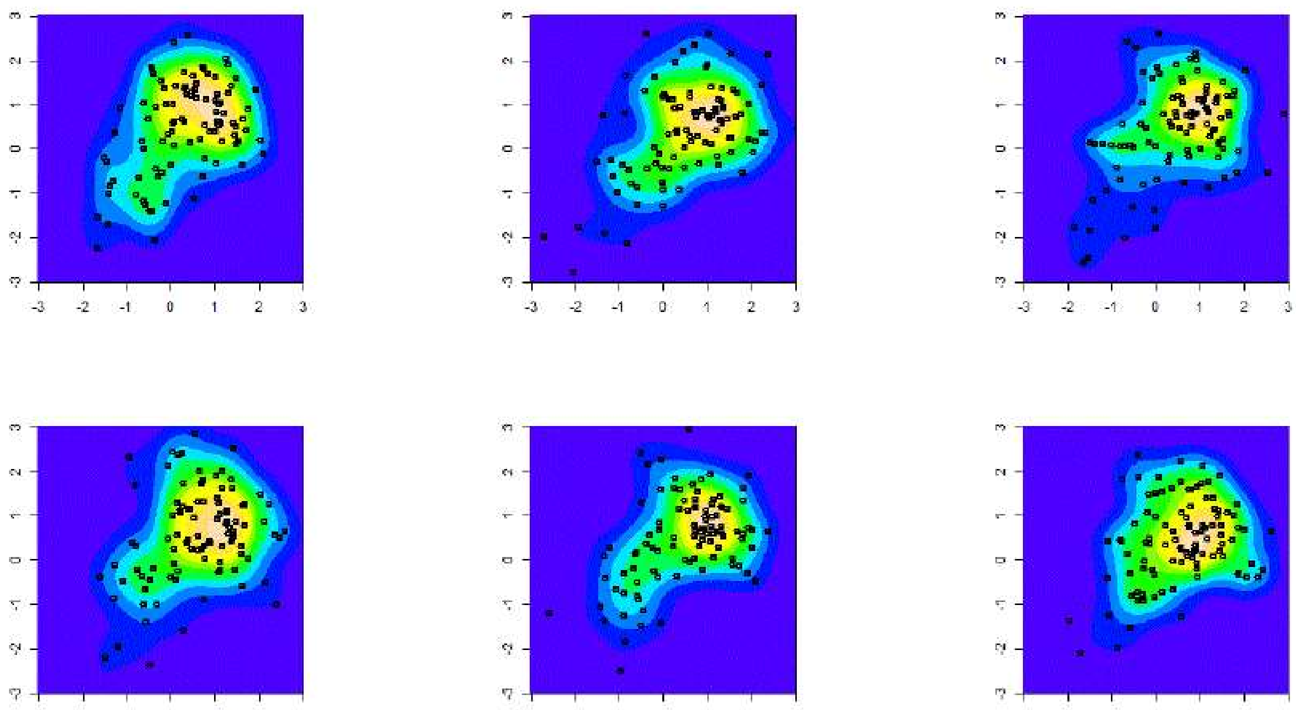
Below are two representations of a particular 2-D density (a mixture of two bivariate normals).





2-D density estimates

Below are 6 samples of $N = 100$ observations from the mixture density pictured on the previous slide and corresponding bivariate density estimates made using the `kde2d` function in the `MASS` package (and its default choice of "bandwidth" covariance matrix).



Direct approximation for optimal classifiers

Consider what form an estimated-density-approximately-optimal classifier

$$\hat{f}(\mathbf{x}) = \arg\max_k \widehat{P[y=k]}\,\widehat{p(\mathbf{x}|k)}$$

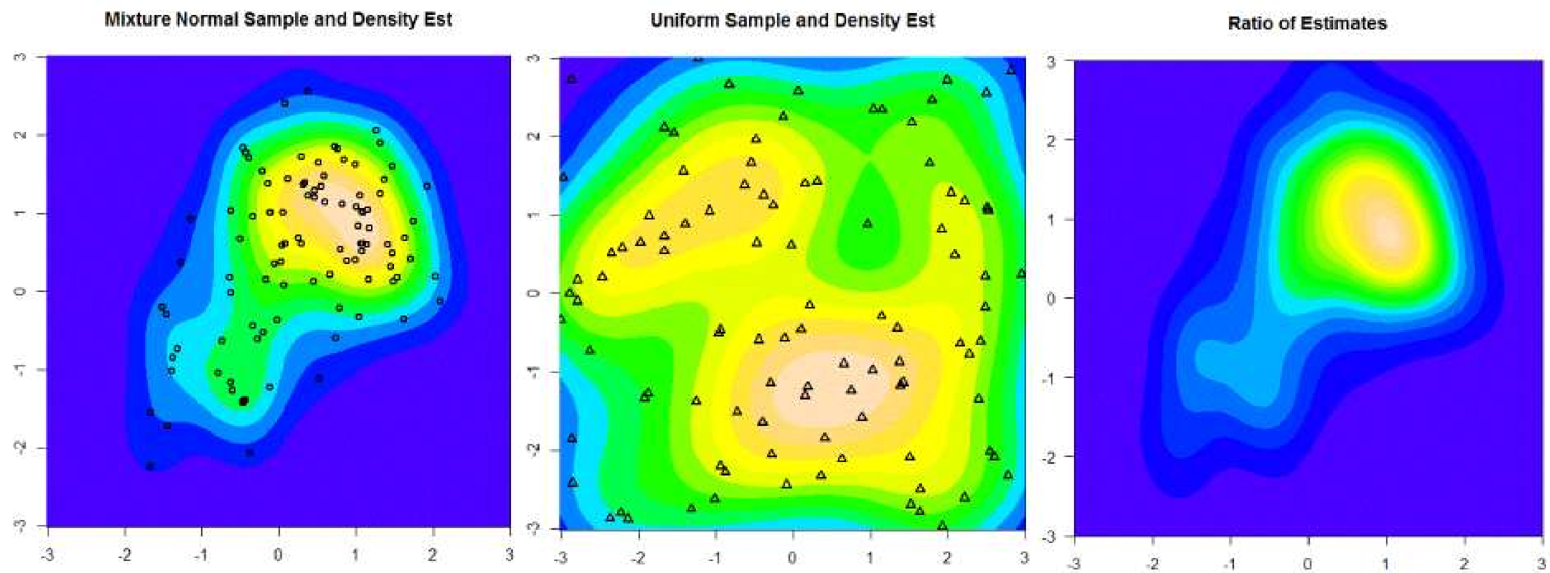
takes where symmetric Gaussian $\left(\mathrm{MVN}_p\left(\mathbf{0}, \lambda^2\mathbf{I}\right)\right)$ kernels are used to produce the $\widehat{p(\mathbf{x}|k)}$. A bit of algebra shows that estimating class-conditional densities based on the parts of the training set with $y = k$ and using training set relative frequencies to estimate class probabilities, an approximately Bayes classifier is

$$\hat{f}_\lambda(\mathbf{x}) = \arg\max_k \sum_{i \text{ s.t. } y_i = k} \exp\left(-\frac{1}{2\lambda^2}\|\mathbf{x}-\mathbf{x}_i\|^2\right)$$

This is a plausible form–classifying to class k when $\mathbf{x}$ is "close to" relatively many training inputs from class k–and bandwidth λ could be chosen by cross-validation.

An example of a ratio of 2-D density estimates

The possibility of using direct estimates $\widehat{P[y=k]}\,\widehat{p(\mathbf{x}|k)}$ to make approximately optimal classifiers basically depends upon how well likelihood ratios can be estimated. The graphic below shows for samples of $N = 100$ from the example bivariate density and from a uniform density on $[-3,3]^2$, and a **ratio of density estimates**.



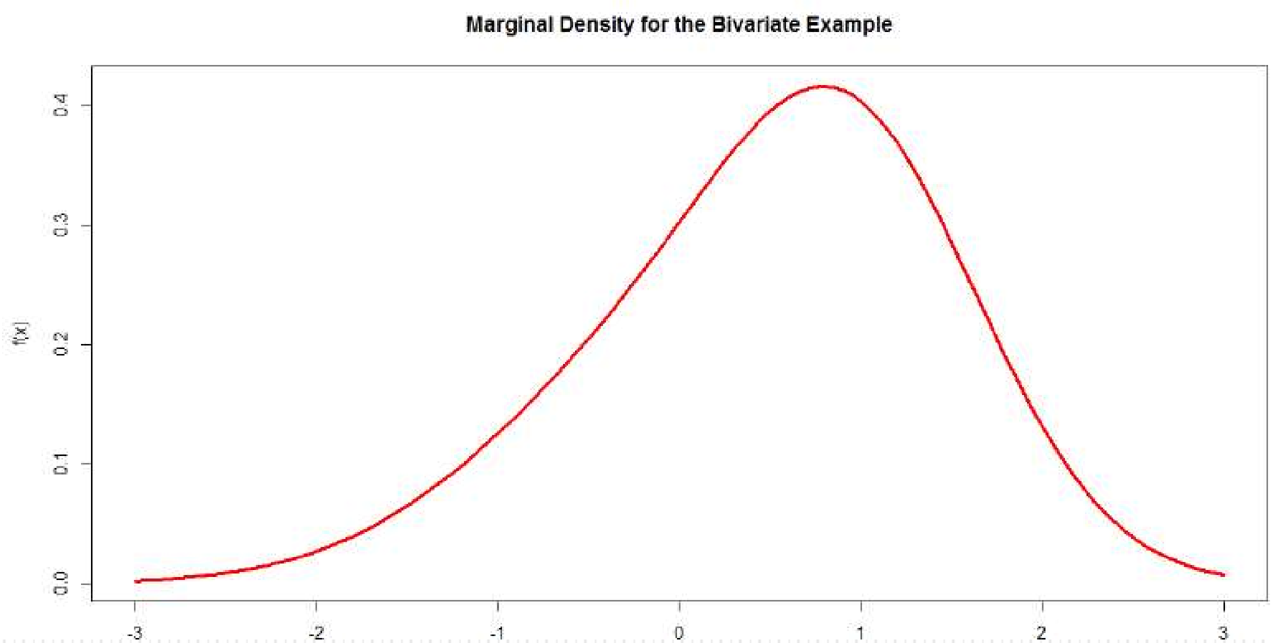
Large p and "naïve Bayes" classification

The $p = 2$ example used here looks reasonably hopeful, as the third graph on the previous slide is some approximation of the original example density (which is proportional to its ratio to a uniform/constant density). But the normal mixture and uniform densities are very simple and the curse of dimensionality makes density estimation for even moderate p (let alone estimation of ratios) problematic. So **direct approximation of optimal classifiers via density estimates also seems problematic for p at all large**.

One related idea that has proven to be of some use is that of estimating only low-dimensional (small p) marginals of a class-conditional density for $\mathbf{x}$ (for which density estimation is feasible) and making a product of them to substitute for an estimate of the joint density (effectively acting like the input $\mathbf{x}$ can be modeled as having independent pieces) in a classifier. This has been called a "**naive Bayes**" **classification** method.

Marginals for the 2-D density

It is easy to see that the naive Bayes idea can fail to be useful even for small p. The density below is the marginal density for both coordinates of $\mathbf{x}$ (both x_1 and x_2) in the bivariate example we have been using. The next panel contrasts the original bivariate density to a density of independence with this one as marginals.



Product of marginals for the example

The original density is clearly quite different from one of independence with the same marginals.