EEE 598C: Statistical Pattern Recognition
Lecture Note 6: Non-parametric Estimation

Darryl Morrell

October 1, 1996

1 Introduction

In this lecture note, we consider ways to estimate the density $f(\mathbf{x}|\omega_i)$ and the probability $P(\omega_i|\mathbf{x})$ from a training set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We will consider non-parametric approaches, in which the density functions are not parameterized by an unknown parameter.

Consider the estimation of a density function $f(\mathbf{x})$ from the set of training data. Let $\mathcal{R}$ be a region in the feature space with volume V, and let P be the probability that $\mathbf{X}$ falls in this region:

$$P = Prob(\mathbf{X} \in \mathcal{R}) \approx f(\mathbf{x})V$$

where $\mathbf{x} \in \mathcal{R}$. Let k be the number of elements of $\mathcal{X}$ that fall into region $\mathcal{R}$. Then an estimate of P is

$$\hat{P} = \frac{k}{n}$$

Combining the two above equations, we get an estimate of $f(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) = \frac{k}{nV} \tag{1}$$

As n goes to infinity, $\hat{P}$ converges (in probability) to P, and $\hat{f}(\mathbf{x})$ converges to the (spatial) average value of $f(\mathbf{x})$ over the region $\mathcal{R}$. In order to get an estimate of $f(\mathbf{x})$ for a given value of $\mathbf{x}$, we must let the volume of the region $\mathcal{R}$ go to zero in such a way that $\mathbf{x}$ is always contained in $\mathcal{R}$. If the number of samples n is fixed, then as V goes to zero, $\mathcal{R}$ will either contain no training points, in which case $\hat{f}(\mathbf{x}) = 0$, or it will contain one or more training points, in which case $\hat{f}(\mathbf{x}) = \infty$. This is not a particularly useful estimate.

Since in practice, we always have a finite number of training samples, we cannot let V approach zero. We will have to accept some variance in the estimate $\hat{P}$ (and thus in $\hat{f}(\mathbf{x})$), as well as some spatial averaging in $\hat{f}(\mathbf{x})$.

From a theoretical standpoint, we can consider the behavior of the estimate $\hat{f}(\mathbf{x})$ as the number of training samples goes to infinity. Ideally, we would like $\hat{f}(\mathbf{x})$ to converge to

$f(\mathbf{x})$. We consider the following way of creating this estimate at $\mathbf{x}$. We use a sequence of regions $\mathcal{R}_1, \mathcal{R}_2, \ldots$, each containing $\mathbf{x}$. $\mathcal{R}_1$ is to be used in estimating $f(\mathbf{x})$ with one training sample, $\mathcal{R}_2$ is to be used in estimating $f(\mathbf{x})$ with two training samples, etc. Let V_n be the volume of $\mathcal{R}_n$, k_n be the number of training samples falling in $\mathcal{R}_n$, and $\hat{f}_n(\mathbf{x})$ be the nth estimate of $f(\mathbf{x})$:

$$\hat{f}_n(\mathbf{x}) = \frac{k_n}{nV_n}$$

In order for $\hat{f}_n(\mathbf{x})$ to converge to $f(\mathbf{x})$, we need three conditions:

1. $\lim_{n\to\infty} V_n = 0$; this ensures that the ratio converges to $f(\mathbf{x})$ if f is continuous.
2. $\lim_{n\to\infty} k_n = \infty$; this ensures that the ratio converges in probability to P.
3. $\lim_{n\to\infty} k_n/n = 0$; this ensures that the ratio converges.

There are two ways that these conditions might be satisfied. One is to shrink an initial region by specifying the volume V_n as a decreasing function of n. The othe is to specify k_n as some function of n; the volume V_n is chosen to include k_n neighbors of $\mathbf{x}$.

2 Parzen Windows

Let $\varphi(\mathbf{u})$ be a function that satisfies the two following requirements:

$$\varphi(\mathbf{u}) \geq 0$$

$$\int \varphi(\mathbf{u}) d\mathbf{u} = 1$$

Define V_n as h_n^d, where h_n is a parameter that represents the width of the window. Define

$$\delta_n(\mathbf{x}) = \frac{1}{V_n} \varphi\left(\frac{\mathbf{x}}{h_n}\right)$$

δ_n is a *Parzen window*. We can use it to compute an estimate $\hat{f}_n(\mathbf{x})$ from n training samples $\mathbf{x}_1$ through $\mathbf{x}_n$:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta_n(\mathbf{x} - \mathbf{x}_i) = \frac{1}{nV_n} \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

This formula is essentially an implementation of Equation (1); the summation term can be interpreted as counting the number of training vectors that fall within a distance of $h_n/2$ to $\mathbf{x}$.

The parameter h_n affects both the magnitude and width of δ_n. If h_n is large, δ_n is broad and has a small amplitude; in this case, $\hat{f}_n(\mathbf{x})$ is a slowly changing function of $\mathbf{x}$, and is a highly smoothed version of $f(\mathbf{x})$. On the other hand, if h_n is small, δ_n is a narrow sharply peaked function (as h_n approaches zero, δ_n approaches a Dirac delta function), so $\hat{f}_n(\mathbf{x})$ is the superposition of n pulses centered at the training vectors. $\hat{f}_n(\mathbf{x})$ is a "noisy" estimate of $f(\mathbf{x})$. In practice, the choice of h_n determines the usefullness of the estimate $\hat{f}_n(\mathbf{x})$; if h_n is too big, the estiamte is too smooth, while if h_n is too small, the estimate is too noisy.

The estimate $\hat{f}_n(\mathbf{x})$ depends on the values of the n training vectors. Since these vectors are random vectors, the estimate is a random variable. Thus, we can consider its mean and variance:

$$\overline{\hat{f}_n(\mathbf{x})} = E_{\mathcal{X}}\left[\hat{f}_n(\mathbf{x})\right]$$

$$\sigma^2_{\hat{f}_n(\mathbf{x})} = Var\left[\hat{f}_n(\mathbf{x})\right]$$

Under certain conditions, the estimate $\hat{f}_n(\mathbf{x})$ can be shown to converge to $f(\mathbf{x})$ as n approaches infinity. These conditions are the following:

1. $f(\mathbf{x})$ must be continuous at $\mathbf{x}$.
2. $\varphi(\mathbf{u}) \geq 0$ and $\int \varphi(\mathbf{u})d\mathbf{u} = 1$
3. $\sup_{\mathbf{u}} \varphi(\mathbf{u}) < \infty$
4. $\lim_{||\mathbf{u}|| \to \infty} \varphi(\mathbf{u}) \prod_{i=1}^{d} u_i = 0$
5. $\lim_{n\to\infty} V_n = 0$
6. $\lim_{n\to\infty} nV_n = \infty$

If these conditions hold, then $\hat{f}_n(\mathbf{x})$ converges to $f(\mathbf{x})$ in a mean square sense:

$$\lim_{n\to\infty} \overline{\hat{f}_n(\mathbf{x})} = f(\mathbf{x})$$

$$\lim_{n\to\infty} \sigma^2_{\hat{f}_n(\mathbf{x})} = 0$$

To see convergence of the mean,

$$\begin{aligned}
\overline{\hat{f}_n(\mathbf{x})} &= E_{\mathcal{X}}\left[\hat{f}_n(\mathbf{x})\right] \\
&= E_{\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}}\left[\frac{1}{nV_n}\sum_{i=1}^{n}\varphi\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)\right] \\
&= E_{\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}}\left[\frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{X}_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{X}_i}\left[\delta_n(\mathbf{x}-\mathbf{X}_i)\right] \\
&= E_{\mathbf{X}}\left[\delta_n(\mathbf{x}-\mathbf{X})\right] \\
&= \int \delta_n(\mathbf{x}-\mathbf{v})f(\mathbf{v})d\mathbf{v}
\end{aligned}$$

Note that this is a convolution of the density $f(\mathbf{x})$ and $\delta_n(\mathbf{x})$; as V_n approaches zero, $\delta_n(\mathbf{x})$ approaches a delta function, and the convolution of a delta function with $f(\mathbf{x})$ is just $f(\mathbf{x})$. So $\lim_{n\to\infty} V_n = 0$ insures that $\overline{\hat{f}_n(\mathbf{x})} \to f(\mathbf{x})$.

To see that the variance of the estimator goes to zero, note that the estimator is a sum of functions of independent random vectors $\mathbf{X}_1$ through $\mathbf{X}_n$, so the variance of the sum is the sum of the variance of these functions:

$$\begin{aligned}
\sigma^2_{\hat{f}_n(\mathbf{x})} &= \sum_{i=1}^{n} Var\left[\frac{1}{nV_n}\varphi\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)\right] \\
&= n\left(E_{\mathbf{X}}\left[\frac{1}{n^2V_n^2}\varphi^2\left(\frac{\mathbf{x}-\mathbf{X}}{h_n}\right)\right] - \overline{\frac{1}{nV_n}\varphi\left(\frac{\mathbf{x}-\mathbf{X}}{h_n}\right)}^2\right) \\
&\leq \frac{1}{nV_n}E_{\mathbf{X}}\left[\frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{X}}{h_n}\right)\right] \\
&\leq \frac{\sup\varphi}{nV_n}\int\varphi\left(\frac{\mathbf{x}-\mathbf{v}}{h_n}\right)f(\mathbf{v})d\mathbf{v} \\
&= \frac{\sup\varphi\,\overline{\hat{f}_n(\mathbf{x})}}{nV_n}
\end{aligned}$$

For the variance to go to zero, we want $nV_n \to \infty$.