Non-parametric density Estimation.

Parametric estimation: Functional form assumed to the class-conditional density.

In Non-Parametric methods, no need to assume any functional form.

The basic idea: suppose $x \in R$.

Given x_i, $i = 1, \ldots, n$.

Problem: Estimate the density function $f(x)$.
with no form for f known.

One possible solution: Learn a piece-wise constant approximation to f.

Divide the x-axis into small intervals & build a function that is constant in each of these intervals.

i.e, $f(x) = K$ over an interval $[a, b]$ then by definition of PDF, we have

$$P\left[a \leq x \leq b\right] = K(b-a)$$
$$= f(x).(b-a)$$

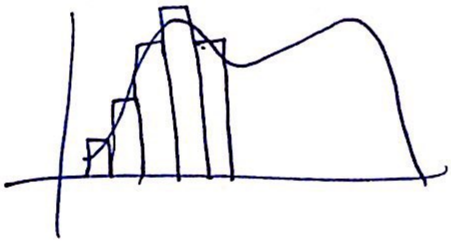
Also, this probability is well approximated by the mass [fraction of data points] falling in that invervals.

$$\therefore f(x) = \frac{P\left[a \leq x \leq b\right]}{b-a}.$$

This is the basic idea of the histogram - approximation.

Scanned with CamScanner

The approximation can be made better by having finer intervals.



This requires more memory.

Also may lead to many empty bins, more-so in high-dimensions.

This may be resolved by exacting bins only around training samples.

<u>Generalization of the histogram idea:</u>

Let $B(x)$ be a region [eg: ball of some radius] around x.

Let $\rho = \int_{B(x)} f(x')\,dx'$

Now if $f(x)$ is nearly constant over $B(x)$ then $\rho = f(x)\,V$, where V is the volume of $B(x)$. $\Rightarrow f(x) = \rho/V$.

Question: How to find out P & V?

We have $X_i \sim f(x)$, IID. $i=1,\dots n$

Suppose out of n IID samples, k samples fall in $B(x)$.

Then, k forms a Binomial distribution with parameter n & P.

$P(k) = {}^nC_k P^k (1-P)^{n-k}$

We also have that for very large n, binomial distribution sharply peaks at its mean (nP).

$\therefore\ k \simeq nP$ or $P \simeq \frac{k}{n}$.

$$\Rightarrow f(x) \simeq \frac{P}{V} \simeq \underline{\underline{\frac{k}{nV}}}$$

This is the basic idea of non-parametric density estimation.

At any x, take a small volume V around x & count the no of data points falling in that region. This gives $\hat{f}(x)$.

Choice of V affects the quality of approximation.

i.e, for $P \approx f(x) V$ to be true, we need V to be small.

But if V is very small, k may be zero most of the time.

$\therefore$ There is a trade-off b/w these two for the choice of V.

[let V_n - volume with n eg.
$f_n(x)$ & k_n be the corresponding values.

$$f_n(x) = \frac{k_n/n}{V_n}.$$

$\times$ for $f_n \to f$, as $n \to \infty$, we must have

$V_n \to 0$, $\frac{k_n}{n} \to 0 \to$ [If $f(x) \neq 0$, then $k_n \to \infty$]

$\downarrow$ to get correct estimates

also $\frac{k_n}{n} \to 0$ to get proper estimates]

Depending upon the way histograms are constructed there are two-approaches.

1) Fix V & Calculate K – Parzen window / Kernel density

2) Fix K & Calculate V – knn approach for density estimation.

We'll look at both one by one.

Parzen-window method.

Define a fn $\phi : R^d \to R$ by

$$\phi(u) = 1 \quad \text{if } |u_i| \leq 0.5, \; i = 1, \ldots, d$$
$$= 0 \quad \text{otherwise}$$

$$u = (u_1, u_2 \ldots u_d)^T$$

$\phi(u)$ is a unit hypercube in R^d, centered at origin. $\phi(u) = \phi(-u)$.

$\phi\left(\frac{u-u_0}{h}\right)$ - Hypercube of side h, centered at u_0.

As usual let $D = \{x_1, \ldots, x_n\}$ be the IID data samples.

Then, $\forall x$, $\phi\left(\frac{x-x_i}{h}\right) = 1$ iff x_i falls in a hypercube of side h centered at x.

$\therefore$ number of datapoints falling in a hypercube of side h centered at x is

$$k = \sum_{i=1}^{n} \phi\left(\frac{x-x_i}{h}\right).$$

Also, volume of the hypercube of side h in $\mathbb{R}^d$ is $\underline{h^d}$.

∴ The estimated density function at x of q could be written as

$$\hat{f}(x) = \underset{n.}{\frac{1}{n}} \sum_{i=1}^{n} \underset{v}{\frac{1}{h^d}} \underset{K}{\phi\left(\frac{x - x_i}{h}\right)}$$

The above is called the parzen window estimate.

∴ If we store all x_i, we can calculate $i=1$
$f(x)$ at any x.

The value of h determines the size of the volume element [and thus the quality of the estimate]

This choice of ϕ, however, leads to abrupt discontinuties as in the case of histogram

Thus, one might use 'smoother' versions of i

s.t

$$\phi(u) \geq 0 \;\forall u \;\&\; \int_{R^d} \phi(u)\,du = 1.$$

if the above happens, then

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V} \phi\left(\frac{x - x_i}{h}\right)$$

would be a density $\Rightarrow \hat{f}(x) \geq 0 \;\&\; \int \hat{f}(x) =$

ϕ is often called the Kernel & hence the name Kernel density estimation.

One popular choice of ϕ,

$$\phi(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^{d} \exp\left[-\frac{1}{n}\|u\|^2\right]$$

d-dim Gaussian : For this too,

$\underline{V = h^d}$.

Now, $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{h\sqrt{2\pi}}\right)^{d} \exp\left[-\frac{\|x - x_i\|^2}{2h^2}\right]$

Essentially, erecting a Gaussian centered around each data point & representing the unknown density as a mixture of these Gaussian. This gives a smoother estimat

One can show that a Kernel density estimate is consistent.

Knn- approach.

~~Now,~~

Kernel density estimates are essentially mixture densities

$$\hat{f}(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{V} \phi\left(\frac{x - x_i}{h}\right).$$

We end up storing all the training samples.

Consider a 2-class problem with n_1 & n_2 samples in each class.

Thus, with a gaussian kernel estimate at ever test x, we need to compute n Gaussians - Computationally expensive.

Also, the size of the volume element h is a critical hyperparameter.

In an alternative approach (knn) we do not choose h (volume element).

Instead we choose K & find V that encloses K nearest neighbours of x.

Then $\hat{f}(x) = \frac{k}{nV}$.

This takes us to the end of Density estimation and also ~~B~~. posterior prob based classification schemes. .

~~Some~~ unsupervised ~~techniques~~.

k-nearest neighbour classifier.

Consider a 2-class problem with priors P_i & class conditionals f_i, $i = 0, 1$.

$f(x) = P_0 f_0(x) + P_1 f_1(x)$ is the overal density of the feature.

Suppose there are n data samples with n_i being from class $i = 0, 1$.

Let's do a k-nn estimation of f. [as well as f_i].

Let there are k_i samples of class-i in this volume.