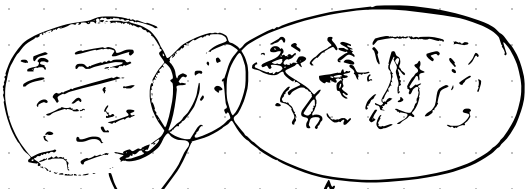
Density Estimation :-

→ are first estimated ID, & given ID, density is estimated.



→ ID = 2

→ 2 clusters, → high density

→ low density

→ data typically forms clusters

In molecular systems :-

→ high f → metastable states

→ low f → "barrier" regions

↓

regions b/w 2 stable states & are explored only rarely. Simulation of rare events.

⇒ in all density estimators we make a compromise in statistical error in estimating (ρ) & systematic error in (ρ) is not constant.

⇓

"Bias - Variance Tradeoff"

---

1. Bias - Variance Tradeoff :-

→ we have performed a preliminary dim. reduc^n & $\delta = ID = 1$

(let's say, as a simple eg.)

→ So we have only 1 variable.

→ We want to estimate Prob. Density as a func^n of 1 variable.

N.B. : Notation related :

(i) $\rho(x) \equiv$ func^n of all coords $\equiv$ Density of data pts

(ii) $P(x) \equiv$ Probability density

(i) & (ii) are same except for normalization.

$$\int dx\ \rho(x) = N \quad (\#\ of\ data)$$

$$P(x) = \frac{\rho(x)}{N}$$

$$\int dx\ P(x) = 1$$

→ Here we generally talk about estimating $\rho(x)$, NOT $P(x)$.

→ So density (ρ) is a func^{c} of only 1 variable (Y).

→ Data pts. correspond to a set of observations (y^i), $i = 1, \ldots, N$.

→ We want to estimate :-

$$\varphi^i \equiv \varphi(\rho^i)$$

φ evaluated at ρ^i

(i) Let's try a naive procedure to estimate φ^i :-

$n_i(\sigma) \rightarrow$ [# of data pts within a dist σ from a pt ρ_i]

data pts. in an interval $\left[\rho^i - \frac{\sigma}{2}, \rho^i + \frac{\sigma}{2}\right]$

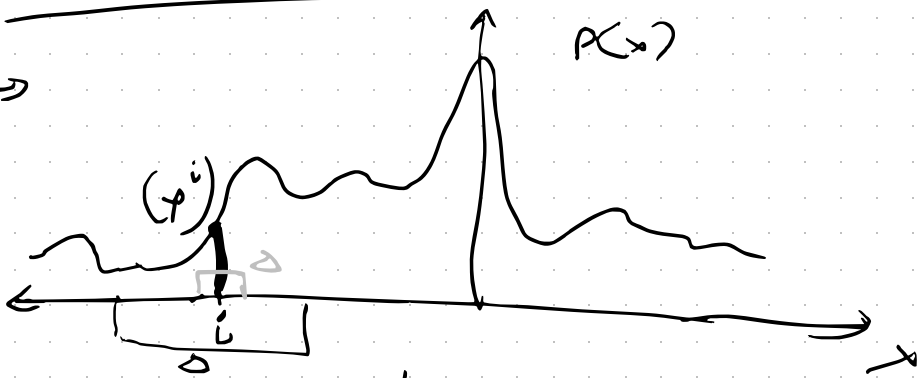
$$\varphi^i \simeq$$

Same procedure as when we make a histogram ; we count # of pts. in each interval.

→ This approx^{n} is correct (mathematically) if δ is small & N is large

(·) Why & how should δ be small?

→



(·) if we want to estimate prob. density at (i), we want to find (y^i), ∴ δ cannot be large (black range) & must instead be small (grey range).

→ more rigorously

$$\frac{n_i(\delta)}{\delta} \sim \frac{1}{\delta} \int_{y^i - \frac{\delta}{2}}^{y^i + \frac{\delta}{2}} dy \; P(y)$$

→ in the limit of large N this should be accurate.

$$\frac{n_i(0)}{\Delta} \sim \left[\frac{1}{\Delta} \int_{y^i - \frac{\Delta}{2}}^{y^i + \frac{\Delta}{2}} dy \; \rho(y) \right] \neq \rho(y^i)$$

$$= \frac{1}{\Delta} \int dy \Big[\rho(y^i) + (y - y^i)\rho'(y^i) + \frac{1}{2}(y - y^i)^2 \rho''(y^i) + \dots \Big]$$

∴ we have a correction factor.

So we Taylor expand the integrand

$\rho(y^i)$

$(y^i, 0)$

cancels out after integral due to parity

$(\sim \Delta^2)$

all odd power terms cancel out

$$\therefore \frac{n_i(\Delta)}{\Delta} \sim \rho^i + \frac{1}{24}\rho''(y^i)\Delta^2 + O(\Delta^4)$$

Σ: odd contributions cancel out

estimate of the "bias"

a systematic error induced by the fact that the density is NOT constant.

$\rightarrow$ $\rho''(y^i)$ is unknown. Only thing we know is to choose Δ small enough s.t. that

$$\frac{1}{24}\rho''\Delta^2 \ll \rho^i \quad \forall i$$

very difficult to impose this cond^n rigorously.

→ : we take (Δ) to be as small as possible. However Δ can't be infinitesimally small.

→ : if Δ is too small, N^i becomes very small (maybe even = 1) & then $\left(\frac{N^i}{\Delta}\right)$ becomes affected by large statistical error:

$$\frac{N^i}{\Delta} \sim p^i \pm \text{error}$$

↑ it's very large if Δ is small

: if Δ is small we have very few observations & if we have too few observations, the error on any estimate becomes very large.

② Estimating the Statistical Error :

→ "variance" of our estimator.

→ Recall : $n \sim \text{Poisson}(\rho V)$

n ↑ (n_i) ↑ # of data pts. in our interval

V ↓ in 1-D $V = \Delta$

$$\therefore P(n|\rho\Delta) = \frac{(\rho\Delta)^n}{n!} e^{-\rho\Delta}$$

Same as before

Swap the 2 members using Bayes' formula

(normalization factor)

$$\therefore P(\rho|n) = c\, \rho^n e^{-\rho\Delta}$$

$$\Rightarrow E(\rho) = \frac{n}{\Delta} \rightarrow \text{(expectation value of } \rho\text{)}$$

estimates density

$$\mathrm{Var}(\rho) = E(\rho^2) - \left(E(\rho)\right)^2$$

$$\therefore \mathrm{Var}(\rho) = \frac{n}{\Delta^2}$$ (simple algebra)

integrals of $P(\rho|n)$ func =

$$\therefore \varepsilon_\rho = \sqrt{\mathrm{Var}(\rho)} = \frac{\sqrt{n}}{\Delta}$$

(Statistical error on density)

→ ∵ density is not normalized relative error makes more sense. →

$$\therefore \frac{\varepsilon_\rho}{\rho} = \frac{\sqrt{n}}{\Delta \cdot \rho} = \frac{\sqrt{n}}{\Delta} \cdot \frac{n}{\Delta} = \frac{1}{\sqrt{n}}$$

$$\therefore \boxed{\frac{\varepsilon_\rho}{\rho} = \frac{1}{\sqrt{n}}} \Rightarrow \text{relative error}$$

→ So for Poisson statistics,

relative error $\propto \frac{1}{\sqrt{\text{no. of observations}}}$

⇓

To have small relative errors on the density estimate (f) we must have large (N).

But, to have large N, we must take large (δ). [$\because N \simeq f, \delta$]

↓

So, to reduce variance we need large δ

& as we saw before, to reduce bias, we need small δ.

⇒ So, we need to choose δ appropriately to compromise b/w statistical (variance) & systematic (bias) error.

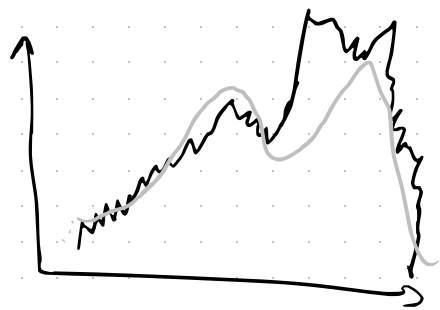
⇒ Same principle as what we do when making a histogram

↓

We don't want a very small interval,
∴ results are very noisy.

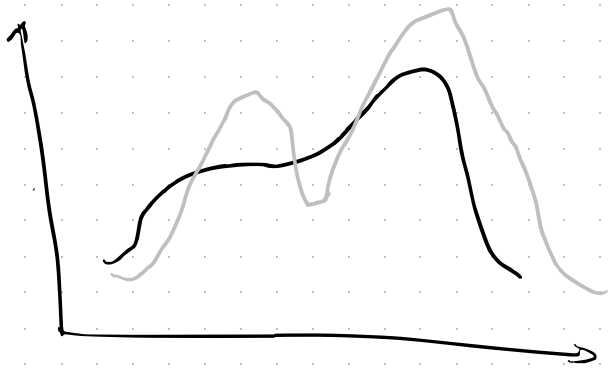
↓

We don't want a very large interval
∴ then we miss some features



our histogram

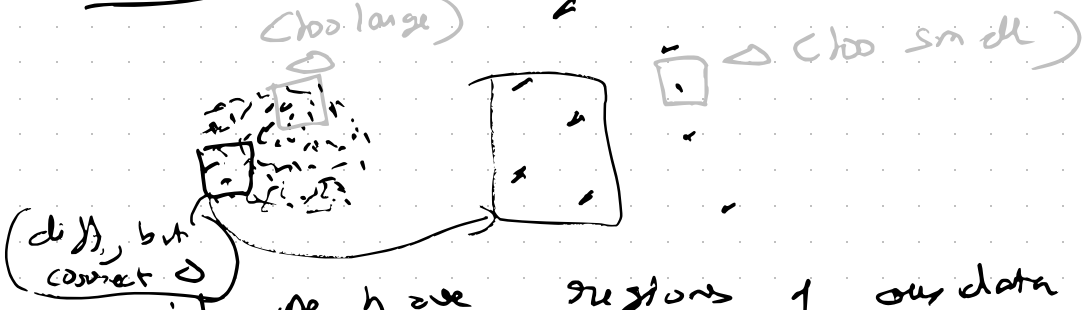
ground truth



→ This is the origin of the "bias – variance" tradeoff

- How to make our density estimate "adaptive"?

What is adaptive?



→ If we have regions of our data with widely varying densities. Eg of molecular systems in beginning of this lecture.

→ If we use a constant ϵ, for one region ϵ is too large, for one it's too small.

⇒ So, correct approach is to take
smaller ○ for high density region &
larger ○ " low " ".

⇒ ∴ The smaller the density, the larger
the ○. [the black boxes have
right choice ✓ ○]

⇒ In the black boxes, the same # of
data pts within ○ is taken,
which means that statistical error is
const

⇒ If ○ is uniform, & # of pts.
in it varies widely, some regions
have high statistical error & others
" low " ". This is
NOT a good choice, & it is much better
to have const. error.

→ This philosophy is at the basis of the k-NN density estimator.

K-NN Density estimator :-

1) choose $K = 10$ (say)

2) make D adaptive. choose

$D_i :=$ distance of k^{th} - N.N.

here $D_i = r^{(k)i} = r^{(10)i}$

$\Downarrow$

$\therefore D_i$ is now point-dependent

$\therefore$ kNN density estimator is adaptive.

(i) Assumption :- density is const. on the scale of D_i

→ if we take (k) as small we can better satisfy this assumption but then we'll have a larger statistical error.

→ Here we have "bias-variance" tradeoff. But the advantage is that "variance" is const. everywhere.

3) $f^i \simeq$ $\frac{K}{\Omega_{ID}\left(r^{(k)i}\right)^{ID}}$

(no of pts within)

$\left(\text{Vol}^m\right)$

Vol^m measured on the embedding manifold enclosed within Δ_i

$\Rightarrow$ ∴ This is why estimating ID is vital before computing ID.

∵ if we choose $k = 10$ & estimate ID = 30 K

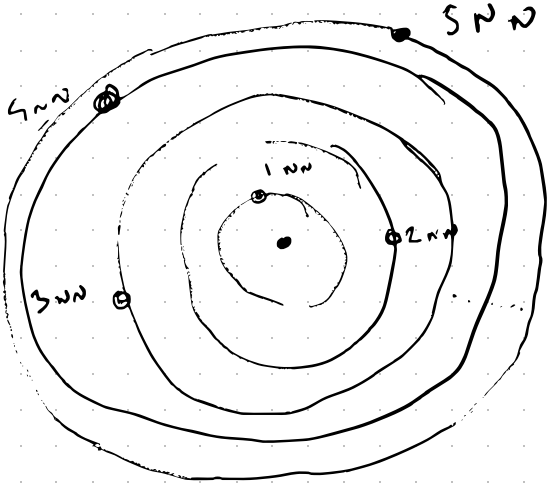
we have $10^{30,000}$ which is ridiculous.

→ (for numerical stability)

Q Derive for ψ^i for kNN estimators in the framework of MLE (maximum likelihood Estimation) :- (ID)

→ it should be similar to before

∵ the point is that all these concepts are closely connected.



the various neighbours & the vol^{m} from i^{th} central pt. to them

→ say, $k = 5$

→ each of the vol^{m}. of these hyperspheres, v_i are harvested from an exp. distrib^{n} of (γ) density parameter

$\therefore v_i \sim exp(\gamma)$

→ $P(v_i | \gamma) = \gamma \exp(-\gamma v_i)$

Also, $v_i = \Omega_{ID} \left[\left(r^{(i)}\right)^{ID} - \left(r^{(i-1)}\right)^{ID} \right]$

with $r^{(0)} = 0$ (by def^{n})

∴ DhA is prob of observing λ given successive val^n of observations $(v_1, v_2 \ldots v_k, \ldots, v_k)$

between (k) & $(k-1)^{th}$ no.

$$P(\lambda \mid v_1 \ldots v_k \ldots v_k)$$

$$= \prod_{i=1}^{k} \lambda e^{-\lambda v_i}$$

$$= \lambda^k e^{-\lambda \sum_{i=1}^{k} v_i}$$

$$= \lambda^k e^{-\lambda \Omega_{ID}(x^{(k)})^{ID}}$$

$\Rightarrow$ P depends on (λ) & (ID). So both can't be estimated simultaneously. However, we have already estimated the (ID).

$\therefore$ We do MLE to compute (ρ)

$$\rightarrow \rho: \quad \frac{\partial \log P}{\partial \rho} = 0$$

$$\rightarrow \log P = k \log \rho - \rho \sum_{ID} \left(n^{(k)}\right)^{ID}$$

$$\rightarrow \frac{\partial \log P}{\partial \rho} = \frac{k}{\rho} - \sum_{ID} \left(n^{(k)}\right)^{ID} = 0$$

$$\therefore \boxed{\rho = \frac{k}{\sum_{ID} \left(n^{(k)}\right)^{ID}}}$$

Thus we obtained the eqⁿ we wanted

N.B. $\rightarrow$ for simplicity's sake, the index

(i) from $n^{(k)i}$ has been

dropped to make the notation simpler.

→ Remember :-

⊙ everything is based on the Poisson distrib^{n}. In that distrib^{n} the 2 parameters that characterize the data distrib^{n} (ID & P) enter in an entangled manner. But it can be disentangled by first estimating ID (such as using 2NN) & then given ID we estimate (P).

⇓

This is particularly relevant for real world data which is high dimensional but where ID is relatively low.