Let $X \sim f(x|\theta)$.

$x = (x_1, \dots, x_n)^T$ — Data.

Can be thought of as one realization of $(X_1, \dots, X_n)^T$ where $X_i \sim$ IID $f(x|\theta)$.

Now, a statistic is any f^n of data.

$g(x_1, \dots, x_n)$

Estimator is a statistic: $\hat{\theta}(x_1, \dots, x_n)$.

Problem: Find an estimator for the density parameter.

Basics of Estimation theory:

unbiased if $\underset{\theta}{E}[\hat{\theta}] = \theta$.

E is w.r.t joint density of $(X_1 \dots X_n)$

$\hat{\Theta}$ is unbiased if for every density in the class of densities we are interested in, expected value is the true parameter value.

Eg: Let $f(x|\theta) \sim N(\theta, 1)$.

$$\hat{\theta}_n = \frac{1}{n}\sum_i x_i$$

Then $E[\hat{\theta}_n] = \theta \quad \forall n \quad \because E[x_i] = \theta$.

$\Rightarrow$ sample mean is an unbiased estimator of the true mean.

Let $\hat{\theta}'(x_1, \dots x_n) = 0.5(x_1 + x_2)$.

$\hat{\Theta}'$ is also unbiased

& consider $\hat{\Theta}'' = x_1$:- This is also unbiased.

Observation: unbiasedness is not enough!

One method : $\hat{\theta}$ is better than $\hat{\theta}'$ if

$$E_\theta\left[(\hat{\theta}-\theta)^2\right] \leq E_\theta\left[(\hat{\theta}'-\theta)^2\right] \; \forall \theta .$$

$$MSE_\theta(\hat{\theta}) = E_\theta\left[(\hat{\theta}-\theta)^2\right]$$

MSE is not easy to compute.

However : $MSE = V_\theta(\hat{\theta}) + \left[B_\theta(\hat{\theta})\right]^2 .$

when $V_\theta(\hat{\theta}) = E_\theta\left[\left(\hat{\theta} - E_\theta(\hat{\theta})\right)^2\right].$

$$B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta .$$

Variance: Extent to which a choice of $\hat{\theta}$ is sensitive to a particular choice of parameter / data

Bias : Extent to which the average of the parameter differs from the desired parameter ?

Ideally : one wants lower bias & lower variance

Proof:

$$MSE[\hat{\theta}] = E\left[(\hat{\theta} - \theta)^2\right]$$

$$= E\left[\left\{(\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta)\right\}^2\right]$$

$$= E\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + (E[\hat{\theta}] - \theta)^2 +$$

$$2E\left[(\hat{\theta} - E[\hat{\theta}]) \cdot (E[\hat{\theta}] - \theta)\right]$$

$$= V(\hat{\theta}) + \left[B(\hat{\theta})\right]^2 + 2\left(E[\hat{\theta}] - \theta\right) \times E\left[(\hat{\theta} - E[\hat{\theta}]\right].$$

$$= V(\hat{\theta}) + \left[B(\hat{\theta})\right]^2.$$

For unbiased estimators, low variance $\Rightarrow$ low MSE.

eg:

$$\hat{\theta}_n = \frac{1}{n} \Sigma x_i$$

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}.$$

for $\hat{\theta}_n' = 0.5(x_1 + x_2)$

$$= V_\theta(\hat{\theta}_n') = \frac{\sigma^2}{2}.$$

Hence $\hat{\theta}$ is better than $\hat{\theta}'$ under MSE.

Thus, unbiased estimators with low MSE are desirable.

$\hat{\theta}$ is said to be uniformly minimum variance unbiased estimator (UMVUE) if

1. $\hat{\theta}$ is unbiased &
2. $MSE_\theta(\hat{\theta}_n) \leq MSE_\theta(\hat{\theta}_n') \ \forall n, \theta$.

If we can get UMVUE, its best. often difficult to get.

Maximum likelihood estimation

Let $X = \{x_1, x_2 \ldots, x_n\}$ be the samples.

Likelihood fn $\triangleq L(\theta, X) = \prod_{j=1}^{n} f(x_j/\theta)$.

ML estimate of θ is

$$\theta^* = \arg\max L(\theta|X).$$

Thus, MLE is an optimization problem.

For convenience, we often take the log likelihood:,

$$l(\theta|X) = \log L(\theta|X) = \sum_{j=1}^{n} \log f(x_j/\theta)$$

For some densities it is analytically solved.

In general, numerical optimization is used.

Example 1 :

Let $f(x|\theta) \sim N(\mu, \sigma^2)$ with

$\theta_1 = \mu$ & $\theta_2 = \sigma$.

$$f(x|\theta) = \frac{1}{\theta_2\sqrt{2\pi}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right).$$

$$l(\theta|x) = \sum_{j=1}^{n}\left[-\log(\theta_2) - \frac{1}{2}\log 2\pi - \frac{(x_j-\theta_1)^2}{2\theta_2^2}\right]$$

$$= -n\log(\theta_2) - \frac{1}{2}n\log 2\pi - \sum_{j=1}^{n}\frac{(x_j-\theta_1)^2}{2\theta_2^2}$$

$$\frac{\partial l}{\partial \theta_1} = \sum_{j=1}^{n}(x_j-\theta_1) = 0$$

$$\frac{\partial l}{\partial \theta_2} = \frac{-n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j-\theta_1)^2 = 0.$$

$$\underline{\hat{\theta}_1 = \frac{1}{n}}\sum_{j=1}^{n}x_j, \qquad \underline{\hat{\theta}_2^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\theta}_1)^2}$$

$\longrightarrow$ unbiased $\qquad$ $\downarrow$ <u>not unbiased</u>

Discrete example:

Discrete RV, $Z \in \{a_1, \ldots, a_M\}$ with Probabilities $P_1, \ldots, P_M$.

Problem: Given IID realizations, estimate P_i.

We have: $P_i \geq 0$ & $\sum_i P_i = 1$.

Represent the Z by an M-D vector

$$X = [x^1, \ldots, x^M]^T, \quad x_i \in \{0, 1\}.$$

i^{th} compont X will be 1, if Z takes value a_i.

Eg: $X = [1, 0, \ldots 0]$ for $a_1 = Z$

one-hot representation (one of M).

Thus, $x^i \in \{0, 1\}$ & $\sum_i x^i = 1$.

also $P_i = \text{Prob}[x^i = 1]$.

PMF for x

$$f(x|p) = \prod_{i=1}^{M} p_i^{x^i}$$

$x = [x^1, \ldots, x^M]^T$, $x^i \in \{0,1\}$, $\sum_i x^i = 1$.

$P = (P_1, P_2 \ldots, P_M)^T$ is the parameter vector.

Problem: Estimate P_i, given data D.

$$D = \{x_1, x_2 \ldots x_n\}$$

$$x_i = [x_i^1, \ldots, x_i^M]^T, \quad x_i^j \in \{0,1\}$$

$$\& \quad \sum_j x_i^j = 1 \quad \forall i.$$

Log likelihood,

$$l(P|D) = \sum_{i=1}^{n} \ln\left(f(x_i|P)\right).$$

$$= \sum_{i=1}^{n} \ln\left(\prod_{j=1}^{M} P_j^{x_i^j}\right).$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(P_j).$$

Find P_i, $i = 1..M$, that maximizes l.

Ideally, P_i can be a large positive number, to maximize ln.

But we know that $\sum_i P_i = 1$.

$\therefore$ Can't make it an un-constrained optimization problem.

$$P_i^* = \arg\max_{P_i} \ell(P|D) = \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(P_j)$$

$$\text{subject to } \sum_{i=1}^{M} P_i = 1$$

Lagrangian. $L = \sum_{i=1}^{n} \sum_{s=1}^{M} x_i^s \ln(P_s) +$

$$\alpha\left(1 - \sum_{s=1}^{M} P_s\right).$$

$$\frac{\partial L}{\partial P_i} = 0$$

$$\sum_{i=1}^{n} \frac{x_i^j}{P_j} - \alpha = 0, \quad j = 1, \ldots, M$$

$$\Rightarrow P_j = \frac{1}{\alpha} \sum_{i=1}^{n} x_i^j, \; j = 1, \ldots, M.$$

Since, $\sum_j P_j = 1$,

$$\alpha = \underline{\sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j} = \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j = \underline{n}.$$

(4.3).

$\because \sum_j x_i^j = 1, \forall i.$

Thus, final estimate for P_j

$$P_j^* = \frac{1}{n} \sum_{i=1}^{n} x_i^j$$

— Fraction of times the j^{th} value occurs in the data.

The above is the general procedure for any DRV.

Eg: Suppose that in a problem feature takes only finitely many values.

Eg: Document classification, feature: word count.

Each document is a vector with i^{th} component — no of times i^{th} word in the dict occurs — Bag of words.

One can have many such features.

Each feature being a DRV whose marginal density can be estimated using the procedure described.

However, we need the joint density (Conditioned on Labels) for Bayes classifier.

One method: Assume independence b/w features & multiply the marginals.

Strong assumption: Called the naive Bayes.

$$f(x|y) = \prod_{j=1}^{d} f(x_j|y).$$

Each marginal has its own Parameters.

Simple examples: Doc classification

x_i - binary vector of dim eq to Doc size, representing whether or not i^{th} word appears in the Doc.

(45)

Then use MLE for Bernoulli parameter estimation

Finally, use Naive Bayes.

MLE can be used to estimate any other parameterized density parameters

Easy to obtain. However if sample sizes are small, ML estimates are very bad.

Doesn't allow to incorporate any knowledge one may have abt the parameters
[eg. of biased coin-toss).

Final estimation depends on data alone.

Soln: Bayesian Estimation.