Q 1. Let x be a d-dimensional binary vector with a multivariate Bernoulli distribution $P(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$

where $\theta = (\theta_1, \dots \theta_d)^t$ is an unknown parameter vector, θ_i being probability that $x_i = 1$. S.T. maximum likelihood estimate for θ is

$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Given $P(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$

The likelihood for n samples is given by.

$$P(D|\theta) = \prod_{k=1}^{n} \prod_{i=1}^{d} \theta_i^{x_{ki}} (1-\theta_i)^{(1-x_{ki})}$$ (assuming independence)

where $D = \{x_1, \dots x_n\}$

Taking log on both sides

$$\ln P(D|\theta) = \sum_{k=1}^{n} \sum_{i=1}^{d} x_{ki} \ln \theta_i + (1-x_{ki}) \ln(1-\theta_i)$$

To find max likelihood we differentiate w.r.t θ_i

i.e. $\frac{\partial}{\partial \theta_i} \ln P(D|\theta) = 0$ (evaluating each component where $i = (1, \dots d)$)

$$\Rightarrow \sum_{k=1}^{n} \sum_{i=1}^{d} \frac{x_{ki}}{\theta_i} + \left[-\frac{(1-x_{ki})}{1-\theta_i}\right] = 0$$

Scanned by CamScanner

$$\Rightarrow \sum_{k=1}^{n} \cancel{\sum_{i=1}^{d}} \frac{x_{ki}}{\hat{\theta}_i} - \frac{(1-x_{ki})}{1-\hat{\theta}_i} = 0$$

– Here we don't consider $\sum_{i=1}^{d}$ as this holds for any i.

$$\Rightarrow \sum_{k=1}^{n} \cancel{\sum_{i=1}^{d}} \frac{x_{ki} - x_{ki}\hat{\theta}_i - \hat{\theta}_i + x_{ki}\hat{\theta}_i}{\hat{\theta}_i(1-\hat{\theta}_i)} = 0$$

– Replacing with $\hat{\theta}_i$ to indicate MLE, the solution to $\frac{\partial}{\partial \theta_i} P(D|\theta_i) = 0$

$$\Rightarrow \sum_{k=1}^{n} \cancel{\sum_{i=1}^{d}} x_{ki} - \hat{\theta}_i = 0$$

$$\Rightarrow \sum_{k=1}^{n} x_{ki} - n\hat{\theta}_i = 0$$

$$\Rightarrow n\hat{\theta}_i = \sum_{k=1}^{n} x_{ki}$$

$$\Rightarrow \hat{\theta}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki}$$

The above result holds for all i. $i = \{1, \ldots, d\}$, so expressing in vector form.

$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

where $\hat{\theta}$ is max likelihood estimate which is same as sample mean in this case.

Scanned by CamScanner

2) Given ÷ $p(X|\omega_1) \sim N(0,1)$, assume $p(X|\omega_2) \sim N(\mu, 1)$
but true distribution is $p(x|\omega_2) \sim N(\mu=1, 10^6)$

a) what is value of maximum likelihood estimation μ_{ML} in poor model, given large data?

Answer ÷ $p(x|\theta) \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ (general form)

$$\ln p(x|\omega) = -\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}\ln p(x|\omega) = 0 - 0 - \frac{1}{2}\frac{(-2(x-\mu))}{\sigma^2}$$

$$P(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$\ln P(D|\theta) = \sum_{i=1}^{n} -\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}$$

To find max likelihood estimate $\frac{\partial}{\partial\mu}\ln P(D|\theta) = 0$

$$\Rightarrow \sum_{i=1}^{n} \frac{-1}{2\sigma^2}(-2(x_i-\mu)) = 0 \Rightarrow \sum_{i=1}^{n} x_i - \mu = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu$$

$\Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ i.e. $\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$

So μ_{ML} is same as sample mean.

So μ_{ML} for our poor model is same as sample mean of data in true dist. i.e. $\mu_{ML} = 1$.

Answer is $\boxed{\mu_{ML} = 1}$

b) $p(x|\omega_1) \sim N(0, 1)$ $p(x|\omega_2) \sim N(1, 10^6)$ - true dist

but in poor model $p(x|\omega_2) \sim N(\mu, 1)$

The question also mentions equally probable categories so $P(\omega_1) = P(\omega_2) = 0.5$. So to find decision boundary.

$$p(x|\omega_1)\, P(\omega_1) = p(x|\omega_2)\, P(\omega_2)$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}\,(1)^2}\, e^{\left\{-\frac{1}{2}\frac{(x-\mu_1)^2}{1}\right\}}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)\left(\frac{1}{\sqrt{2\pi}}\, e^{\left\{-\frac{1}{2}\frac{(x-\mu_2)^2}{(1)^2}\right\}}\right)$$

Cancelling & taking log on both sides

$$\Rightarrow \cancel{\frac{1}{\sqrt{2\pi}}}\left(-\frac{1}{2}(x-\mu_1)^2 = -\frac{1}{2}(x-\mu_2)^2\right.$$

$$\Rightarrow -2x\mu_1 + \mu_1^2 = -2x\mu_2 + \mu_2^2$$

$$\Rightarrow \mu_1^2 - \mu_2^2 = 2x(\mu_1 - \mu_2)$$

$$\Rightarrow (\mu_1 + \mu_2) = 2x$$

$$\Rightarrow x^* = \frac{\mu_1 + \mu_2}{2}$$ where x^* denotes decision boundary

So the solution for given distributions is

$$x^* = \frac{0+1}{2} = \frac{1}{2} = 0.5$$

So decision boundary for MLE in poor model is

$$\boxed{x^* = 0.5}.$$

c) $p(x|\omega_1) \sim \mathcal{N}(0,1)$ $\quad p(x|\omega_2) \sim \mathcal{N}(1, 10^6)$

To calculate decision boundary

$$p(x|\omega_1)\, P(\omega_1) = p(x|\omega_2)\, P(\omega_2)$$

$P(\omega_1) = P(\omega_2) = \frac{1}{2}$.

$$\Rightarrow \frac{1}{\sqrt{2\pi}\,1} \exp\left\{ -\frac{1}{2}(x^2) \right\} = \frac{1}{\sqrt{2\pi\, 10^6}} \exp\left\{ \frac{-1}{2} \frac{(x-1)^2}{10^6} \right\}$$

$$\Rightarrow e^{-\frac{1}{2}x^2} = 10^{-3} \exp\left\{ -\frac{1}{2} \frac{(x-1)^2}{(10)^6} \right\}$$

Taking ln on both sides

Scanned by CamScanner

$$\frac{-x^2}{2} = -3\ln 10 - \frac{1}{2}\frac{(x-1)^2}{(10)^6}$$

Solving for x. expanding $\left(-3\ln 10 - \frac{1}{2}\frac{(x-1)^2}{10^6}\right)$ gives

$$-0.0000005x^2 + 0.000001x - 6.90775 \text{ (used calculator)}$$

$$\Rightarrow \frac{-x^2}{2} = -0.0000005x^2 + 0.000001x - 6.90775$$

Add $\frac{x^2}{2}$ on both sides.

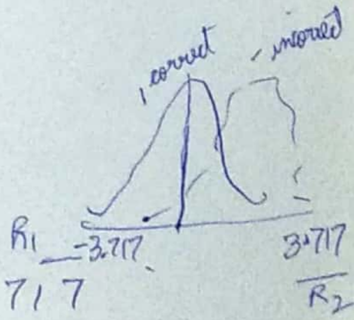
$$0.499x^2 + 0.000001x - 6.907 = 0$$

Solving using quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{-0.000001x \pm \sqrt{(0.000001)^2 - 4(0.499)(-6.907)}}{2 \times 0.499}$$

$$\Rightarrow x = \frac{-0.000001 \pm \sqrt{13.815}}{0.999}$$



$(\sqrt{13.81}) = 3.716$ so $x = \frac{\pm 3.716}{0.999} = \pm 3.717$

In (b) decision boundary was only a single point. But in above we see the regions & decision boundaries. And since b) is incorrect model we won't get minimum classification error.

Q 3) If we employ a novel method to estimate mean of data $D = x_1, \ldots x_n$; we assign mean to be value of first point

a) Show that method is unbiased.

Assuming $x_1, \ldots x_n$ are i.i.d's (identically and independently distributed) the $E(X_i)$ for any i will be μ as each x_i is identically distributed w.r.t others. So as per above problem statement we take mean to be value of first point i.e.

$$\mu_{MC} = x_1$$

Now $E[\mu_{MC}] = E[x_1] = \mu$ [as $E(X_i) = \mu$ for any i

provided that each of the data points are identically distributed.]

So $E(\mu_{MC}) = \mu$ i.e. estimate is unbiased as it is equal to true value.

b) State why this method is nevertheless highly undesirable.

The mean provides an unbiased estimate. But computing variance of μ_{ML} would give us an idea.

$$Var(\mu_{ML}) = Var(X_1) = \sigma^2 \quad \text{(as } X_i\text{'s are i.i.d)}$$

i.e. $\boxed{E[(X_1 - \mu)^2] = \sigma^2}$

But let us consider the usual case of when

$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$Var(\mu_{ML}) = E([\mu_{ML} - \mu]^2) = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2\right)$$

$$= \frac{1}{n^2} E\left(\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} E[(X_i - \mu)^2]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Above can also be written as $Var(\mu_{ML}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right)$

$$= \frac{1}{n^2}\left[Var(X_1) + \cdots Var(X_n)\right) \quad \text{as } X_i\text{'s are i.i.d}$$

$$= \frac{\sigma^2}{n}$$

So from above 2 comparisons we see that

the variance of estimated mean is much higher ~~than~~ when we assume μ_{ML} to be equal to value of first point. Higher variance of mean is undesirable because in usual assumptions $Var(\mu_{ML}) = \frac{\sigma^2}{n}$ but here $Var(\mu_{ML}) = \sigma^2$. In case of $Var(\mu_{M}) = \frac{\sigma^2}{n}$ as number of samples increases, we can get precise estimates of group means. But in this problem, variance of estimated mean $= \sigma^2$ has no dependence on n (number of samples). So this condition is undesirable.

Q 4) MLE for binomial distribution: $Bin(N, \mu) \sim P(X=m) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$

In general $P(X=x) = {}^{N}C_{m} \, \mu^{x} (1-\mu)^{x}$

$D = \{x_1 - - x_n\}$ - X are i.i.d's

$$P(D|\theta) = \prod_{i=1}^{m} {}^{N}C_{x_i} \, \mu^{x_i} (1-\mu)^{N-x_i}$$

$$\ln P(D|\theta) = \sum_{i=1}^{m} \ln N_{C_{x_i}} + x_i \ln \mu + (N-x_i)\ln(1-\mu)$$

$$\frac{\partial}{\partial \mu} \ln P(D|\theta) = 0 \Rightarrow \sum_{i=1}^{m} \frac{x_i}{\mu} + \left(-\frac{(N-x_i)}{1-\mu}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{m} \frac{x_i}{\mu} - \frac{(N-x_i)}{1-\mu} = 0 \Rightarrow \sum_{i=1}^{m} \frac{x_i(1-\mu) - (N-x_i)\mu}{\mu(1-\mu)} = 0$$

$$\Rightarrow \sum_{i=1}^{m} x_i - x_i\mu - N\mu + x_i\mu = 0 \Rightarrow \sum_{i=1}^{m} x_i - N\mu = 0$$

$$\Rightarrow \sum_{i=1}^{m} x_i = mN\mu \Rightarrow \mu = \frac{1}{mN}\sum_{i=1}^{m} x_i$$

$$\boxed{\hat{\mu} = \frac{1}{mN}\sum_{i=1}^{m} x_i}$$

Scanned by CamScanner