ML-As-2

Point Estimation

The Poisson distribution is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, packet arrival density is often modeled with the Poisson distribution. If X is Poisson distributed, i.e., $X\ Poisson(\lambda)$, its probability mass function takes the following form:

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

It can be shown that $\mathbb{E}(X) = \lambda$. Assume now we have n i.i.d. data points from $Poisson(\lambda) : \mathbb{D} = X_1, \ldots, X_n$. (For the purpose of this problem, you can only use the knowledge about the Poisson and Gamma distributions provided in this problem.)

(a)

Show that the sample mean $\hat{\lambda} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$ is the maximum likelihood estimate (MLE) of λ and it is unbiased ($\mathbb{E}\hat{\lambda} = \lambda$).

Finding the MLE

$$L(\lambda) = \prod_{i=1}^{n} P(X_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

$$\ln L(\lambda) = \sum_{i=1}^{n} (X_i \ln \lambda - \lambda - \ln(X_i!))$$

$$\frac{d}{d\lambda} \ln L(\lambda) = \sum_{i=1}^{n} \left(\frac{X_i}{\lambda} - 1 \right) = 0$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Unbiasedness

$$E(\hat{\lambda}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)$$

Since X_i are i.i.d., we can take the expectation inside the sum:

$$E(\hat{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \lambda = \frac{n\lambda}{n} = \lambda$$

Therefore, $E(\hat{\lambda}) = \lambda$, confirming that $\hat{\lambda}$ is an unbiased estimator of λ . of $\hat{\lambda}$

(b)

Now let's be Bayesian and put a prior distribution over λ. Assuming that λ follows a Gamma distribution with the parameters (α, β) , its probability density function:

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Where $\Gamma(\alpha) = (\alpha - 1)!$ (here we assume α is a positive integer). Compute the posterior distribution λ.

$$P(\theta|\lambda) = \frac{P(X|\lambda)P(\lambda|\alpha, \beta)}{P(X)}$$

$$P(\theta|\lambda) \propto P(X|\lambda)P(\lambda|\alpha, \beta)$$

$$= \frac{\lambda^X e^{-\lambda}}{X!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$P(\theta|\lambda) \propto \lambda^{X+\alpha-1}e^{-\lambda(\beta+1)}$$

Let $\alpha' = X + \alpha$, $\beta' = \beta + 1$
Then the distribution is still a Gamma distribution

(c)

Derive an analytic expression for the maximum a posterior (MAP) of λ under $Gamma(\alpha, \beta)$ prior.

$$MAP(\lambda) = \prod_{i=1}^{n} P(X_i|\lambda) = \frac{\prod_{i=1}^{n} P(X_i|\lambda)P(\lambda)}{P(X)} \propto \prod_{i=1}^{n} P(X_i|\lambda)P(\lambda)$$

$$\prod_{i=1}^{n} P(X_i|\lambda)P(\lambda) \propto log\prod_{i=1}^{n} P(X_i|\lambda)P(\lambda)$$

$$logP(\lambda \mid X) \propto log(\prod_{n}^{i=1} P(X_i \mid \lambda)P(\lambda)) \propto \sum_{i=1}^{n} logP(X_i \mid \lambda) + logP(\lambda)$$

$$\sum_{i=1}^{n} logP(X_i \mid \lambda) + logP(\lambda)$$

Prior Distribution $P(\lambda)$

$$P(\lambda|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$logP(\lambda|\alpha, \beta) \propto (\alpha - 1)log\lambda - \beta\lambda$$

Likelihood function $P(X_i|\lambda)$

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

$$logP(X_i \mid \lambda) \propto X_i log\lambda - \lambda$$

$$\sum_{i=1}^{n} logP(X_i \mid \lambda) + logP(\lambda)$$

$$MAP(\lambda) \propto \sum_{i=1}^{n} X_i log\lambda - n\lambda + (\alpha - 1)log\lambda - \beta\lambda$$

$$= log\lambda(\sum_{i=1}^{n} X_i + \alpha - 1) - \lambda(n + \beta)$$

$$\frac{d}{d\lambda}logP(\lambda|X) = \frac{\sum_{i=1}^{n} X_i + \alpha - 1}{\lambda} - (n + \beta) = 0$$

$$\lambda_{MAP} = \frac{\sum_{i=1}^{n} X_i + \alpha - 1}{n + \beta}$$

Source of Error: Part 1

(a)

The bias of an estimator is defined as $E[\hat{\mu}] - \mu$

The bias is $1 - \mu$

The variance of an estimator is defined as $Var(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}^2])]$

$\therefore Var(\hat{\mu}) = 0$

This is not a good estimator, since the bias is large when the true value of μ is not 1. Usually we don't have any information about the true value of μ, so it is unreasonable to assume it is equal to 1.

(b)

$E(\hat{\mu}) = \mu$ the bias is 0.
This is an unbiased estimator.
The variance of this estimator is $Var(\hat{\mu}) = Var(y_1) = 1$

This is not a good estimator since its variability does not decrease with the sample size.

(c)

$$-2\sum_{i=1}^{n}(y_i - \mu) + 2\lambda\mu = 0$$

$$\hat{\mu} = \frac{1}{n+\lambda}\sum_i y_i = \frac{n}{n+\lambda}\bar{y}$$

$$E[\hat{\mu}] = \frac{1}{n+\lambda}E\left[\sum_i y_i\right] = \frac{n}{n+\lambda}\mu$$

Bias of the estimator :

$$\text{bias} = \frac{-\lambda\mu}{n+\lambda}$$

Variance of the estimator :

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{n+\lambda}\sum_i y_i\right) = \frac{1}{(n+\lambda)^2}\sum_i \text{Var}(y_i) = \frac{n}{(n+\lambda)^2}\sigma^2$$

Source of Error: Part 2

(a)

(b)

The error is equal to 0.

Because $p(X|Y=0)$ and $p(X|Y=1)$ do not overlap.

Just check whether it is in the interval [-4,-1] or in the interval [1,4]

(c)

$$P[\text{error}] = P[x \in [0,1]] \times P[\text{error}|x \in [0,1]]$$
$$= (P[x \in [0,1]|y=0]P[y=0] + P[x \in [0,1]|y=1]P[y=1]) \times P[\text{error}|x \in [0,1]]$$
$$= \left(\frac{1}{4}\times\frac{1}{2} + \frac{1}{4}\times\frac{1}{2}\right)\times\frac{1}{2} = \frac{1}{8}$$

(d)

- $E[X|Y=0] = -2.5$ and $\text{Var}[X|Y=0] = \frac{3}{4}$ (using the variance formula for the uniform distribution),
- $E[X|Y=1] = 2.5$ and $\text{Var}[X|Y=1] = \frac{3}{4}$.

Since we are approximating $p(X|Y)$ using a normal distribution, we have:

- $\hat{p}(X|Y=0) = N(-2.5, 0.75)$,
- $\hat{p}(X|Y=1) = N(2.5, 0.75)$.

Using these, for $x < 0$, we find $\hat{p}(X|Y=0) > \hat{p}(X|Y=1)$, and for $x > 0$, $\hat{p}(X|Y=0) < \hat{p}(X|Y=1)$. Therefore, the classifier will make no error in classifying new points.

(e)

Given a finite amount of data, we will not learn the mean and variance of $p(X|Y)$ perfectly. Therefore, the classifier's error will increase due to the limited data. In this scenario, we would have both bias and error in our model.

Gaussian (Naïve) Bayes and Logistic Regression

No, the new $P(Y|X)$ is no longer the form used by logistic regression.

$$P(Y=1|\mathbf{X}) = \frac{P(Y=1)P(\mathbf{X}|Y=1)}{P(Y=1)P(\mathbf{X}|Y=1)+P(Y=0)P(\mathbf{X}|Y=0)}$$

$$= \frac{1}{1+\frac{P(Y=0)P(\mathbf{X}|Y=0)}{P(Y=1)P(\mathbf{X}|Y=1)}}$$

$$= \frac{1}{1+\exp\left(\ln\frac{P(Y=0)P(\mathbf{X}|Y=0)}{P(Y=1)P(\mathbf{X}|Y=1)}\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\ln\frac{P(\mathbf{X}|Y=0)}{P(\mathbf{X}|Y=1)}\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

The log ratio of class-conditional probabilities:

$$\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \sum_i \ln\frac{\frac{1}{\sqrt{2\pi}\sigma_{i0}}\exp\left(-\frac{(X_i-\mu_{i0})^2}{2\sigma_{i0}^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_{i1}}\exp\left(-\frac{(X_i-\mu_{i1})^2}{2\sigma_{i1}^2}\right)}$$

Simplifies to:

$$= \sum_i \ln\frac{\sigma_{i1}}{\sigma_{i0}} + \sum_i\left(\frac{(X_i-\mu_{i1})^2}{2\sigma_{i1}^2}-\frac{(X_i-\mu_{i0})^2}{2\sigma_{i0}^2}\right)$$

$$= \sum_i \ln\frac{\sigma_{i1}}{\sigma_{i0}} + \sum_i \frac{\sigma_{i0}^2-\sigma_{i1}^2}{2\sigma_{i0}^2\sigma_{i1}^2}X_i^2 + 2\left(\frac{\mu_{i0}\sigma_{i1}^2-\mu_{i1}\sigma_{i0}^2}{\sigma_{i0}^2\sigma_{i1}^2}\right)X_i + \frac{\mu_{i1}^2\sigma_{i0}^2-\mu_{i0}^2\sigma_{i1}^2}{2\sigma_{i0}^2\sigma_{i1}^2}$$

Probability of $P(Y=1|X)$:

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

Simplifies to:

$$P(Y=1|X) = \frac{1}{1+\exp\left(w_0+\sum_i w_iX_i+\sum_i v_iX_i^2\right)}$$

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i\left(\ln\frac{\sigma_{i1}}{\sigma_{i0}} + \frac{\mu_{i1}^2\sigma_{i0}^2-\mu_{i0}^2\sigma_{i1}^2}{2\sigma_{i0}^2\sigma_{i1}^2}\right)$$

$$w_i = \frac{\mu_{i0}\sigma_{i1}^2-\mu_{i1}\sigma_{i0}^2}{\sigma_{i0}^2\sigma_{i1}^2}$$

$$v_i = \frac{\sigma_{i0}^2-\sigma_{i1}^2}{2\sigma_{i0}^2\sigma_{i1}^2}$$