Homework 1 Solutions

Solution 1:- Beta updating from censored likelihood and MLE for uniform distribution

A. In this question, we have to infer a posterior for the parameter θbased on limited information. In the usual situation,we would have the results (i.e. number of heads) of the n=5 tosses. If this were the case, the result of the inference would be a posterior with a Beta distribution. The corresponding parameters would be a combination based on the parameters of the likelihood and the prior, as discussed in the class. However, we are not working with this simple situation. The only information that we have about the result is that the number of head is less than three ($X < 3$).So we have an event which has more than one possible outcome. Fortunately, it is simple to decompose this particular event in individual results of the sample space. Being more specific,

$$p(X < 3) = p(X = 0) + p(X = 1) + p(X = 2)$$

Therefore, our problem is nothing more than a combination of the simpler case mentioned above and hence, we should expect a similar combination to appear in the result.

The posterior of the parameters θ is given by Bayes' rule:

$$p(\theta|X < 3) = \frac{p(X<3|\theta)p(\theta)}{p(X<3)}$$

Since we want to derive an expression proportional to the posterior, we can ignore the denominator. So, what is left is to calculate the likelihood, calculate the prior and combine both of them. We shall start with the likelihood:

$$p(X < 3|\theta) = p(X = 0|\theta) + p(X = 1|\theta) + p(X = 2|\theta)$$

$$= \binom{5}{0}\theta^0\theta^5+\binom{5}{1}\theta^1\theta^4+\binom{5}{2}\theta^2\theta^3$$

$$=Bin(X = 0|\theta, 5) + Bin(X = 1|\theta, 5) + Bin(X = 2|\theta, 5)$$

We see that the likelihood is a mixture distribution composed by 3 Binomial distributions. Now, let's focus our attention in the prior $p(\theta)$. According to the question (and considering that $\theta \epsilon [0, 1]$ as the probability of getting heads in a coin toss), the prior is given by:

$$p(\theta) = Beta(\theta|1, 1) = \frac{1}{B(1,1)}\theta^0(1 - \theta)^0 = 1 = U(0, 1)$$

Thus, the prior is a uniform distribution over the interval [0,1]. Therefore, it is now easy to calculate the posterior (up to a normalization constant)

$$p(\theta|X < 3) = \infty p(X < 3|\theta)\, p(\theta)$$

$$= Bin(X = 0|\theta, 5) + Bin(X = 1|\theta, 5) + Bin(X = 2|\theta, 5)$$

$$= \binom{5}{0}\theta^0\theta^5 + \binom{5}{1}\theta^1\theta^4 + \binom{5}{2}\theta^2\theta^3 \qquad (1)$$

Eq. (1) is the posterior of the parameter θ.

B. **(a)** Let's calculate the maximum likelihood estimator of the uniform distribution on the dataset $D = \{x_1, \ldots\ldots\ldots, x_n\}$.

$$p(D|a) = \prod_{i=1}^{n} p(x_i|a)$$

$$= \prod_{i=1}^{n} \frac{1}{2a} \mathrm{I}(x_i \epsilon [- a, a])$$

$$= (\frac{1}{2a})^n \prod_{i=1}^{n} \mathrm{I}(x_i \epsilon [- a, a])$$

We see that the likelihood is a combination of two terms. The first term $\frac{1}{(2a)^N}$ is a monotonic decreasing function in a.Thus as we decrease a,we get bigger likelihoods. The second term, $\prod_{i=1}^{n} \mathrm{I}(x_i \epsilon [- a, a])$is equal to 0 if any of the points of the dataset are outside the interval $[-a, a]$, and equal to 1 otherwise. So, in order to have the maximum likelihood, we have to take the minimum value $\hat{a}$such that $D \subset [-\widehat{a, a}]$.Let's define $D_{abs} = |x_1|, \ldots\ldots |x_n|$.Thus, the MLE is given as

$$\hat{a} = sup(D_{abs}) = max(|x_i|) \tag{2}$$

Eq. (2) is the MLE for the parameter a.

(b) Given a new datapoint x_{n+1},the model would assign it the following probability

$$p(x_{n+1}) = \frac{1}{2\hat{a}}\mathbf{I}(x_{n+1}\epsilon[-\widehat{a, a}])$$

(c) The problem exposed in part (b) is the zero count problem, where we have an unreliable MLE, given the nature of our dataset. The MLE is unreliable because it gives 0 probability to every datapoint bigger than max($|x_i|\epsilon D$). This becomes even more unreliable when we have small datasets. A better approach would be using Bayesian analysis.

Solution 2. The solution procedure goes as follows.

- **Univariate normal distribution**:

  The density is $p(y|\mu, \sigma^2)=\frac{1}{\sqrt{2\pi\sigma}}exp(-\frac{(y-\mu)^2}{2\sigma^2})$, which can be re-written as

$$p(y|\mu, \sigma^2)=\frac{1}{\sqrt{2\pi}}exp(-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - log\sigma)$$

  It is clear that $h(y)=\frac{1}{\sqrt{2\pi}}$, $\theta = [\theta_1, \theta_2] = [\mu/\sigma^2, -1/2\sigma^2]$, $R(y)=[y, y^2]$,

  $\Psi(\theta)= \mu^2/2\sigma^2+ log\sigma= -\theta_1^2/4\theta_2-\frac{1}{2}log(-2\theta_2)$.

- **Binomial distribution**:

  The density is $p(y|n, p) = \binom{n}{y} p^y(1-p)^{n-y}$, which can be re-written as

$$p(\mathrm{y}|n,p) = \binom{n}{y}\ exp\left(\ y\ log\ \frac{p}{1-p}\ +\ n\ log\ (1\ -\ p)\right)$$

It is clear that $h(y)=\binom{n}{y}$,$\theta\ = log\ (\frac{p}{1-p})$, $R(y) = y$, $\Psi(\theta) = -\ n\ log\ (1\ -\ p) =$ $n\ log\ (1\ +\ e^{\theta})$

- **Geometric distribution**:

The density of the distribution (corresponding to the number of failures before a success) can be written as $p(\mathrm{y}|p) = p\ \ (1\ -\ p)^{y}$. For such case

$$p(\mathrm{y}|p) = exp\ \ (\ y\ log(1\ -\ p)\ +\ log\ p\ )$$

It is clear that $h(y)=1$, $\theta\ = log\ (1\ -\ p)$, $R(y) = y$, $\Psi(\theta) = -\ \ log\ p = -\ \ log\ (1\ -\ e^{\theta})$

- **Poisson distribution**:

The density is $p(\mathrm{y}|\lambda) = \dfrac{\lambda^{y} exp\ (-\lambda y)}{y!}$, which can be re-written as

$$p(\mathrm{y}|\lambda) = \frac{1}{y!}\ exp(y\ log\ \lambda\ \ -\ \lambda)$$

It is clear that $h(y)=\dfrac{1}{y!}$, $\theta\ = log\ \lambda$, $R(y) = y$, $\Psi(\theta) = \lambda = exp(\theta)$

- **Exponential distribution**:

The density is $p(\mathrm{y}|\lambda) = \lambda\ exp(-\ \ \lambda y)$, which can be re-written as

$$p(\mathrm{y}|\lambda) = \lambda\ exp(-\ \ \lambda y)\ =\ exp(-\ \lambda y\ +\ log(\lambda))$$

It is clear that $h(y)=1$, $\theta\ = \lambda$, $R(y) = -\ y$, $\Psi(\theta) = -\ log(\lambda) = -\ log(\theta)$

Solution 3. Integrate the density over the whole sample space $y\ \in\ Y$, if $f_{\theta}(y)$is a probability density, we can derive that

$$\int_y f_\theta(y)\mathrm{dy} = \int_y h(y)\, exp\big(\theta \cdot R(y) - \psi(\theta)\big)\mathrm{dy}$$

$$1 = \int_y h(y)\, exp[\theta \cdot R(y)]/[exp\, \psi(\theta)]\mathrm{dy}$$

$$exp\, \psi(\theta) = \int_y h(y)\, exp(\theta \cdot R(y))\mathrm{dy}$$

Therefore, $\psi(\theta) = log \int h(y)\, exp(\theta \cdot R(y))dy$. Taking its derivative w.r.t. θ we can get

$$\frac{d}{d\theta}\psi(\theta) = \frac{1}{\int h(y)\, exp(\theta \cdot R(y))dy} \cdot \int h(y)\, exp\,(\theta \cdot R(y))\, R(y)\, dy$$

$$= \frac{1}{\int h(y)\, exp(\theta \cdot R(y) - \psi(\theta))dy} \cdot \int$$

$$h(y)\, exp\,(\theta \cdot R(y) - \psi(\theta))\, R(y)\, dy$$

$$= \int R(y)h(y)\, exp\,(\theta \cdot R(y) - \psi(\theta))\, dy \; = E_{p(y|\theta)}R(y)$$

Solution 4.

(Codes)