CS/DSC/AI 391L: Machine Learning

Homework 4 - Theory - Solutions

Lecture: Prof. Qiang Liu

1. (a) To ensure that $\Pr(X = i)$ is a valid probability mass function, we should set θ to ensure that the probabilities are non-negative and sum to one:

$$\sum_{i=1}^{3} \Pr(X = i) = \theta_1 + 2\theta_1 + \theta_2 = 1 \qquad \Rightarrow \qquad 3\theta_1 + \theta_2 = 1$$

$$\Pr(X = i) \geq 0 \qquad \Rightarrow \quad \theta_1 \geq 0, \quad \theta_2 \geq 0.$$

(b) Write down the joint probability $\Pr(D \mid \theta)$ and the log probability $\Pr(D \mid \theta)$.
Joint probability:
Recall that s_1, s_2, s_3 are the numbers of observation of 1, 2, 3, respectively. Then

$$\Pr(D \mid \theta) = \theta_1^{s_1}(2\theta_1)^{s_2}\theta_2^{s_3} = 2^{s_2}\theta_1^{s_1+s_2}\theta_2^{s_3}$$

Log probability:

$$\log \Pr(D \mid \theta) = s_2 \log 2 + (s_1 + s_2) \log \theta_1 + s_3 \log \theta_2.$$

(c) Calculate the maximum likelihood estimation $\hat{\theta}$ based on D;
Using $3\theta_1 + \theta_2 = 1$, we have

$$\begin{aligned}\log \Pr(D \mid \theta) &= s_2 \log 2 + (s_1 + s_2) \log \theta_1 + s_3 \log \theta_2 \\ &= s_2 \log 2 + (s_1 + s_2) \log \theta_1 + s_3 \log(1 - 3\theta_1).\end{aligned}$$

Taking gradient w.r.t. θ_1:

$$\frac{\partial \log \Pr(D \mid \theta)}{\partial \theta_1} = \frac{s_1 + s_2}{\theta_1} - \frac{3s_3}{1 - 3\theta_1}$$

The maximum likelihood estimator $\hat{\theta}$ should have zero gradient:

$$\begin{aligned}&\frac{s_1 + s_2}{\hat{\theta}_1} - \frac{3s_3}{1 - 3\hat{\theta}_1} = 0 \\ &\Rightarrow (s_1 + s_2)(1 - 3\hat{\theta}_1) - 3s_3\hat{\theta}_1 = 0 \\ &\Rightarrow \hat{\theta}_1 = \frac{s_1 + s_2}{3(s_1 + s_2 + s_3)} \\ &\Rightarrow \hat{\theta}_2 = 1 - 3\hat{\theta}_1 = \frac{s_3}{s_1 + s_2 + s_3}.\end{aligned}$$

2. The log-likelihood function of the exponential distribution is

$$\ell(\beta) = \sum_{i=1}^{n} \log f(x^{(i)} \mid \beta) = \sum_{i=1}^{n} \left(-\frac{x^{(i)}}{\beta} - \log \beta \right) = -\frac{S_n}{\beta} - n \log \beta,$$

where we define $S_n = \sum_{i=1}^n x^{(i)}$. The optimal point $\hat{\beta}$ should have zero gradient:

$$\nabla \ell(\hat{\beta}) = \frac{S_n}{\hat{\beta}^2} - \frac{n}{\hat{\beta}} = 0.$$

Solving this gives

$$\hat{\beta} = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n x^{(i)}.$$

Therefore, $\hat{\beta}$ equals the sample mean.

3. (a) Recall the Bayes rule:

$$P(\theta \mid D) \propto P(D \mid \theta) P_0(\theta),$$

where D is the observed data (which are the 30 items in this case); $P_0(\theta)$ is the prior distribution of θ, and $P(D \mid \theta)$ is the likelihood.

Since we assume the prior is a uniform distribution $U([0,1])$, we have that $P_0(\theta)$ is constant for $\theta \in [0,1]$.

Because we observe 5 defective items out of a sample of size 30, the likelihood is

$$P(D \mid \theta) = \theta^5 (1-\theta)^{25}, \qquad \forall \theta \in [0,1].$$

Therefore, the posterior is

$$P(\theta \mid D) \propto P(D \mid \theta) P_0(\theta) \propto P(D \mid \theta) = \theta^5 (1-\theta)^{25}, \qquad \forall \theta \in [0,1].$$

Additional remark Note that although $P(\theta \mid D)$ is proportional to $P(D \mid \theta)$, they have very different meaning: $P(D \mid \theta)$ is a probability over the observation D, while the posterior $P(\theta \mid D)$ is a distribution over θ. If we explicitly write down the normalization of $P(\theta \mid D)$, we have

$$P(\theta \mid D) = \frac{1}{Z} \theta^5 (1-\theta)^{25}, \qquad Z = \int_0^1 \theta^5 (1-\theta)^{25} d\theta.$$

This is an instance of Beta distribution (see Wikipedia).

(b) Recall the Bayes rule:

$$p(\mu \mid D) \propto p(D \mid \mu) p_0(\mu).$$

Since the prior is assumed to be $\mathcal{N}(0,1)$, we have

$$p_0(\mu) \propto \exp\left(-\frac{1}{2}\mu^2\right).$$

Since $x^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(\mu, 1)$, we have

$$p(x^{(i)} \mid \mu) \propto \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^2\right).$$

Hence

$$\begin{aligned} p(D \mid \mu) &= \prod_{i=1}^{n} p(x^{(i)} \mid \mu) \\ &\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^2\right) \\ &= \exp\left(\sum_{i=1}^{n} -\frac{1}{2}(x^{(i)} - \mu)^2\right). \end{aligned}$$

Now,

$$\begin{aligned} p(\mu \mid D) &\propto p(D \mid \mu) \times p_0(\mu) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^2 - \frac{1}{2}\mu^2\right) \end{aligned}$$

This shows that $p(\mu \mid D)$ is a Gaussian distribution, because the density is proportional to an exponential of the a quadratic form. We now need to find out the mean and variance of $p(\mu \mid D)$, by completing the square.

$$\begin{aligned} p(\mu \mid D) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^2 - \frac{1}{2}\mu^2\right) \\ &= \exp\left(-\frac{1}{2}\left(a\mu^2 + b\mu + c\right)\right) \end{aligned}$$

where a, b, c are the coefficients.

$$a = n + 1, \qquad b = -2\sum_{i=1}^{n} x^{(i)}, \qquad c = \sum_{i=1}^{n}(x^{(i)})^2.$$

Recall the complete square formula (see Wiki):

$$a\mu^2 + b\mu + c = a(x - h)^2 + k, \qquad \text{where} \qquad h = -\frac{b}{2a}, \qquad k = c - \frac{b^2}{4a}.$$

We have

$$\begin{aligned} p(\mu \mid D) &\propto \exp\left(-\frac{1}{2}\left(a\mu^2 + b\mu + c\right)\right) \\ &\propto \exp\left(-\frac{1}{2}a(\mu - h)^2\right) \qquad //k \text{ is dropped since it is a constant.} \end{aligned}$$

Thus, the posterior $p(\mu \mid D)$ is a Gaussian distribution with mean

$$\mu_{posterior} = h = \frac{\sum_{i=1}^{n} x^{(i)}}{n + 1},$$

and variance

$$\sigma^2_{posterior} = a^{-1} = \frac{1}{n + 1}.$$