CAP 5638: Assignment #3

Due on Wednesday, Oct 7, 2015

XiuWen Liu 10:10am

Jian Wang

Contents

Problem 1

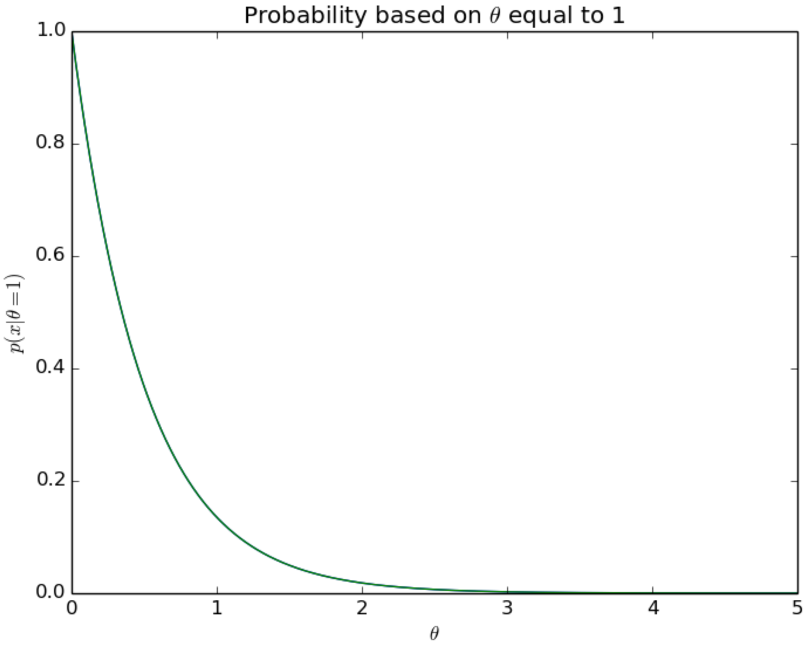
Problem 1, Chapter 3 of the textbook
Let x have an exponential density:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

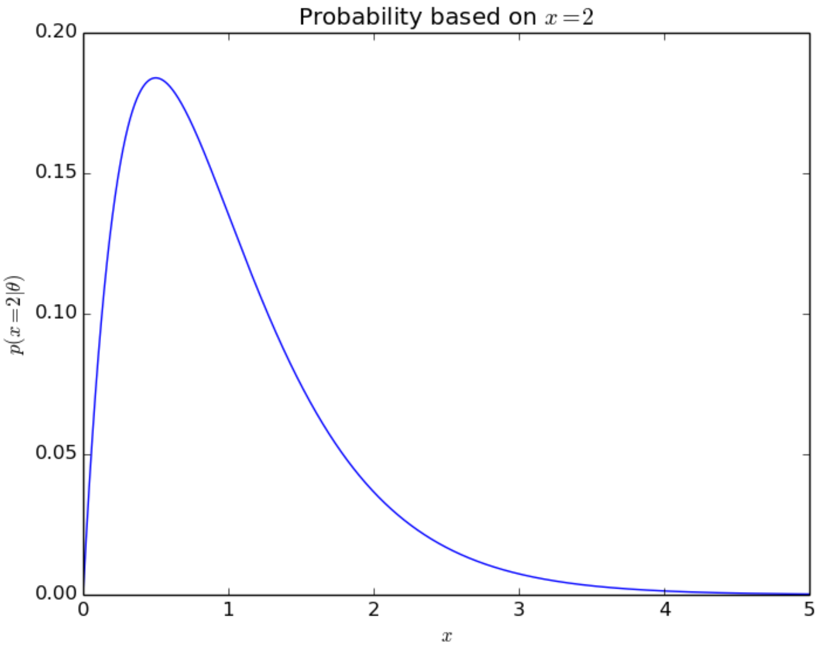
(a) Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, for $x = 2$.
(b) Suppose that n samples $x_1, ..., x_n$ are drawn independently according to $p(x|\theta)$.
Show that the maximum likelihood estimate fo θ is given by:

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

Answer:
(a)we draw the chart as follows, the first chart is $p(x|\theta)$ versus x for $\theta = 1$ and the second chart is $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, for $x = 2$



Problem 1 continued on next page. . .



(b)The log-likelihood function is:

$$l(\theta)=\sum_{k=1}^{n} lnp(x_k|\theta)=\sum_{k=1}^{n}[ln\theta-\theta x_k]=nln\theta-\theta\sum_{k=1}^{n}x_k$$

To find the maximum of the likelihood function, we take the first derivative of the above equation. The result is as follows:

$$\nabla_\theta l(\theta)=\frac{\partial}{\partial\theta}[nln\theta-\theta\sum_{k=1}^{n}x_k]$$
$$=\frac{n}{\theta}-\sum_{k=1}^{n}x_k=0$$

So the maximum -likelihood solution is:

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

Problem 2

Problem 3, Chapter 3 of the textbook.
Maximum likelihood methods apply to estimates of prior probabilities as well. Let samples be drawn by successive, independent selections of a state of nature ω_i, with unknown probability $P(\omega_i)$. Let $z_{ik}=1$, if the state of nature for the kth sample is ω_i and $z_{ik}=0$ otherwise.

(a) Show that

$$P(z_{i1},...,z_{in}|P(\omega_i)) = \prod_{k=1}^{n} P(\omega_i)^{z_{ik}}(1-P(\omega_i))^{1-z_{ik}}$$

(b) Show that the maximum likelihood estimate for $P(\omega_i)$ is

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} z_{ik}.$$

Interpret your result in words.

Answer:

We denote that

$$z_{ik} = \begin{cases} 1 & if\ the\ state\ of\ nature\ for\ the\ k^{th}\ sample\ is\ \omega_i \\ 0 & otherwise \end{cases}$$

(a) The sample are drawn by successive independently selection of a state of nature ω_i with probability $P(\omega_i)$. We have then :

$$Pr[z_{ik}=1|P(\omega_i)] = P(\omega_i)$$

and:

$$Pr[z_{ik}=0|P(\omega_i)] = 1-P(\omega_i)$$

we can rewrite the above equation as:

$$Pr[z_{ik}|P(\omega_i)] = [P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{1-z_{ik}}$$

By the independence of the successive selections, we have :

$$\begin{aligned} P(z_{i1},...,z_{in}|P(\omega_i)) &= \prod_{k=1}^{n} P(z_{ik}|P(\omega_i)) \\ &= \prod_{k=1}^{n} [P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{1-z_{ik}} \end{aligned}$$

(b) The log-likelihood as a function of $P(\omega_i)$ is:

$$\begin{aligned} l(P(\omega_i)) &= lnP(z_{i1},...,z_{in}|P(\omega_i)) \\ &= ln[\prod_{k=1}^{n}[P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{(1-z_{ik})}] \\ &= \sum_{k=1}^{n}[z_{ik}lnP(\omega_i) + (1-z_{ik})ln(1-P(\omega_i))] \end{aligned}$$

Therefore, the maximum-likelihood values for the $P(\omega_i)$ must satisfy:

$$\nabla_{P(\omega_i)} l(P(\omega_i)) = \frac{1}{P(\omega_i)}\sum_{k=1}^{n} z_{ik} - \frac{1}{1-P(\omega_i)}\sum_{k=1}^{n}(1-z_{ik}) = 0$$

We solve this equation and find :

$$(1-\hat{P}(\omega_i))\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n}(1-z_{ik})$$

which can be rewritten as:

$$\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik} + n\hat{P}(\omega_i) - \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik}$$

So the final solution is then:

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} z_{ik}$$

That is, the estimate of the probability of category ω_i is merely the probability of obtaining its indicatory value in the training data, just as we would expected.

Problem 3

Problem 7 , Chapter 3 od the textbook
Show that if our model is poor, the maximum likelihood classifier we derive is not the best –even among our (poor) model set –by exploring the following example. Suppose we have two equally probable categories (i.e., p(ω_1)=P(ω_2)=0.5). Further, we know that $p(x|\omega_1) \sim N(0,1)$ but *assume* that $p(x|\omega_2) \sim N(\mu,1)$. (That is, the parameter θ we seek by maximum likelihood techniques is the mean of the second distribution.) Image however that the *true* underlying distribution is $p(x|\omega_2) \sim N(1,10^6)$.
(a)what is the value of our maximum likelihood estimate μ, in our poor model, given a large amount of data.
(b) What is the decision boundary arising from this maximum likelihood estimate in the poor model.
(c) Ignore for the moment the maximum likelihood approach, and use the methods from chapter 7 to derive the Bayes optimal decision boundary given the *true* underlying distributions –$p(x|\omega_1) \sim N(0,1)$ and $p(x|\omega_2) \sim N(1,10^6)$. be careful to include all portions pf the decision boundary.
(d) Now consider again classifiers based on the (poor) model assumption of $p(x|\omega_2) \sim N(\mu,1)$. Using your result immediately above, find a *new* value of μ that will give lower error than the maximum likelihood classifier.
(e) Discuss these results, with particular attention to the role of knowledge of the underlying model.
Answer:

Problem 4

problem 10, Chapter 3 of the textbook (hint think about the bias and variance.) Suppose we employ a novel method for estimating the mean of a data set, $\mathcal{D} = x_1, x_2, ..., x_n$: we assign the mean to the value of the first point in the set, i.e., x_1.
(a) Show that tis method is unbiased.
(b) State why this method is nevertheless highly undesirable.

Anwser:
(a)Consider the novel method of estimating the mean of a set of points as taking its first value, which we denote $M = x_1$,
(a) If the expected value of a statistics is equal to the true value, then we call the statistics unbiased. for

this case, if we repeat the selection of the first point of a data set we have:

$$bias = E(M) - \mu = lim_{K\rightarrow\infty}\frac{1}{K}\sum_{k=1}^{K} M(k) - \mu = 0;$$

Where $M(k)$ is the first point in data set k drawn from the given distribution.

(b) While the unusual method for estimating the mean may indeed be biased, it will generally have large variance, and this is an undesirable property. Note that $E[(x_i) - \mu)^2] = \sigma^2$, and the RMS error, σ, is independent of n. This undesirable behavior is quite different from that of the measurement of:

$$\bar{x} = \frac{1}{n}\sum_{i=1} nx_i$$

where we see:

$$\begin{aligned} E[(\bar{x} - \mu)^2] &= E[(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)^2] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}[E[(x_i - \mu)^2] \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Thus the RMS error, $\sigma/\sqrt{n}$, approaches 0 as $1/\sqrt{n}$. Note that there are many superior method for estimating the mean, for instance the sample mean and some other re-sampling method such as "bootstrap" method.

Problem 5

Suppose that the prior distribution of θ and the parametric form (a uniform distribution) remain the same as in the example given in section 3.5 in the textbook, compute first the Bayesian estimation of θ and then the estimated class conditional $p(x|D)$ for $D = \{3, 9, 7\}$. You need to specify the Bayesian estimation and the class conditional fully (i.e., you need to specify the functions with all required constants). Then plot the class conditional form from 0 to 10.

Answer:

From the example of the textbook, we know that the prior distribution of θ is a uniform distribution, which listed in the following:

$$p(x|\theta) \sim U(0, \theta) = \begin{cases} 1/\theta & 0 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

Similarly with the example in the textbook, we will sue recursive Bayes methods to estimate θ and the underlying distribution. Before any data arrive, we have $(\theta|\mathcal{D}^0) = p(\theta) = U(0, 10)$ when our first data point

Problem 5 continued on next page. . .

x_1=3 arrives, we use the equation 54, from the textbook, to get an improved estimate:

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & 3 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

When the next data point $x_2 = 9$ arrives, we have :

$$p(\theta|\mathcal{D}^2) \propto p(x|\theta)p(\theta|\mathcal{D}^1) = \begin{cases} 1/\theta^2 & 9 \leq x \leq \theta \\ 0 & otherwise, \end{cases}$$

when the third data point, which is equal to 7, comes, we have:

$$p(\theta|\mathcal{D}^3) \propto p(x|\theta)p(\theta|\mathcal{D}^2) = \begin{cases} 1/\theta^3 & 9 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

The general form of the solution is:

$$p(\theta|\mathcal{D}^n) \propto 1/\theta^n \text{ for } \max_x[\mathcal{D}^n] \leq \theta \leq 10.$$

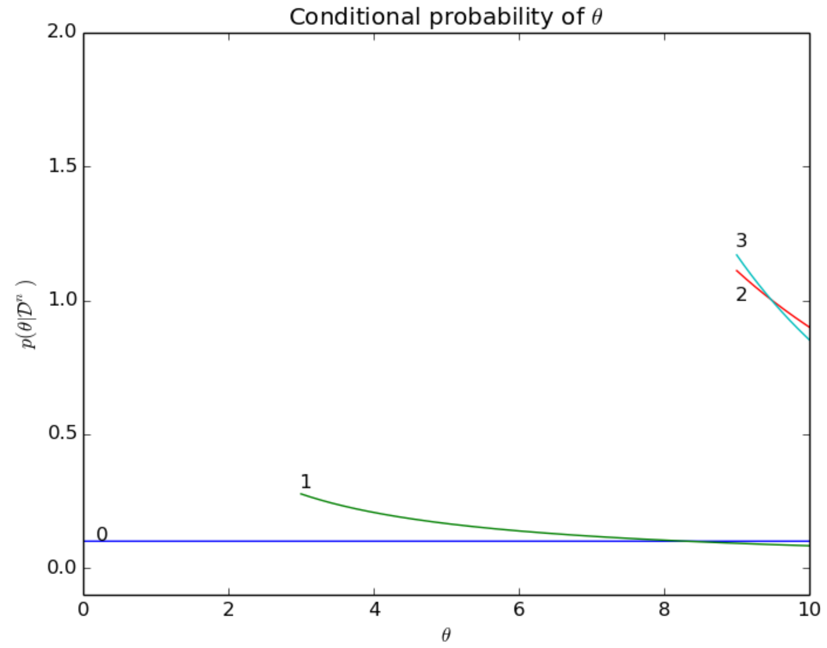
To make the above equation become the density function, we also calculate the constant term in all the $p(\theta|\mathcal{D}_i)$, which listed as follows:

$$p(\theta|\mathcal{D}^1) = \begin{cases} 1/(ln(10) - ln(3) * 1/\theta & 9 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

$$p(\theta|\mathcal{D}^2) = \begin{cases} 1/(1/9 - 1/10) * 1/\theta^2 & 9 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

$$p(\theta|\mathcal{D}^3) = \begin{cases} 1/(1/162 - 1/200) * 1/\theta^3 & 9 \leq x \leq 10 \\ 0 & otherwise, \end{cases}$$

We also plot the result for the $p(\theta|\mathcal{D}^i)$ under 4 cases as follows:



next part we try to solve the conditional class probability, $p(x|\mathcal{D}^3)$, given the equation 50 on the text book, we need to solve the integration of the following equation:

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$

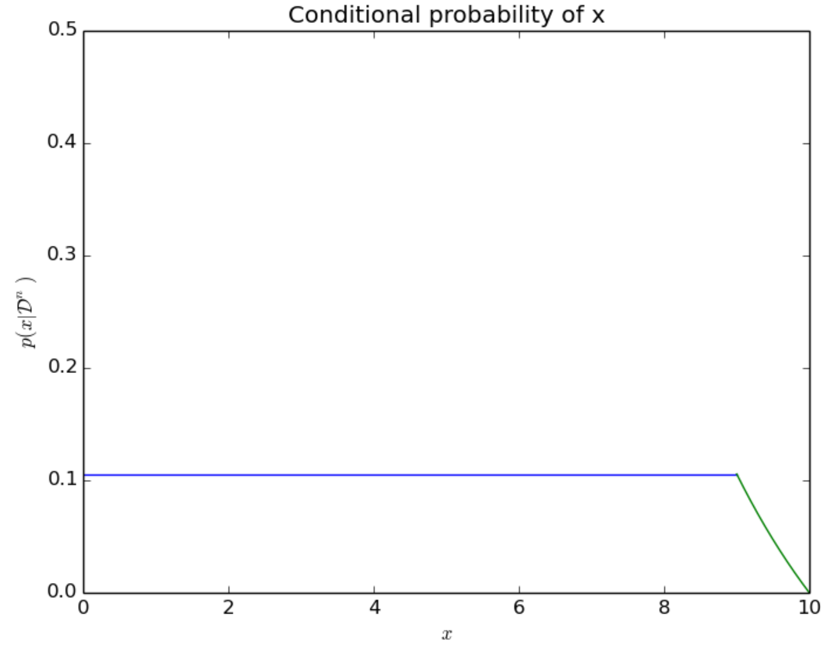
note that θ should be bigger than x and the domain of θ is 9 to 10 in our case, so in the domain from 0 to 9 of x, the integral region of θ is 9 to 10, when x bigger than 9, the integral region of θ becomes x to 10. This means that when x is from 0 to 9, it follows an uniform distribution, and when x is bigger than 9, it follows some polynomial density function.
Now we calculate the integration under two situation:
When x is less than 9:

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta = \int_9^{10} 1/(1/162 - 1/200) * 1/\theta^4 d\theta = 0.10565302144249514$$

when x is bigger than 9:

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta = \int_x^{10} 1/(1/162 - 1/200) * 1/\theta^4 d\theta = 1/(1/162 - 1/200) * 1/3 * (1/x^3 - 1/1000)$$

The following chart shows the result of the conditional class probability in the domain for x from 0 to 10.



Problem 6

Problem 11 chapter 3 of the textbook; you only need to show the univariate case.
One measure of the difference between two distribution in the same space is the *Kullback-l=Leibler divergence* of Kullback-Leibler" distance":

$$D_{KL}(p_1(x), p_2(x)) = \int p_1(x) ln \frac{p_1(x)}{p_2(x)} dx$$

(This "distance," does not obey the requisite symmetry and triangle inequalities for a metric.) Suppose we seek to approximate an arbitrary distribution $p_2(x)$ by a normal $p_1(x) \sim N(\mu, \Sigma)$. Show that the values that lead to the smallest Kullback-leibler divergence are the obvious ones:

$$\begin{aligned} \mu &= \mathcal{E}_2[x] \\ \Sigma &= \mathcal{E}_2[(x-\mu)(x-\mu)^t] \end{aligned}$$

where the expectation taken is over the density $p_(x)$.