Homework: aka the Maximum Likelihood and Bayesian Parameters

Dan Beatty

March 7, 2007

1 Introduction

Both of these problems are exercises on maximum-likelihood estimation.

2 Problem 2 from [1, 140-141]

Let x have a uniform density

$$p(x|\theta)\ U(0,\theta) \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

- Suppose that n samples $\mathcal{D} = (x_1, ..., x_n)$ are drawn independently according to $p(x|\theta)$. Show that the maximum-likelihood estimate for θ is $\max[\mathcal{D}]$ - that is, the value of the maximum element in $\mathcal{D}$.
- Suppose that $n = 5$ points are drawn from the distribution and the maximum value of which happens to be $\arg\max_k x_k = 0.6$. Plot the likelihood $p(\mathcal{D}, \theta)$ in the range $0 \leq \theta \leq 1$. Explain in words why you do not need to know the values of the other four points.

In this case,

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} \frac{1}{\theta} = \frac{n}{\theta} \tag{2}$$

$$l(\theta) = \ln p(\mathcal{D}|\theta) = \sum_{k=1}^{n} (\ln 1 - \ln \theta) \tag{3}$$

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln P(\vec{x}_k|\theta) \tag{4}$$

$$= \sum_{k=1}^{n} \nabla_\theta \ln 1 - \nabla_\theta \ln \theta = \sum_{k=1}^{n} \frac{1}{\theta} \tag{5}$$

In this case, the only way for $\max_{l(\theta)}$ to be satisfied is for $\theta = \infty$. This is actually the minimum.

On the other, the largest value available is the largest x_k of $\mathcal{D}$. Why? The value for $\frac{n}{\theta}$, assuming that θ is always positive is maximum as $\theta \to 0$. The reverse of the condition causes $p(D|\theta) = 0$ the instance that $\theta < x_k$. Thus its maximum is at x_k. This is the intuitive and trivial answer.

3 Problem 4 from [1, 141]

> Let $\vec{x}$ be a d-dimensional binary (0 or 1) vector with a multivariate Bernoulli distribution
> $$P(\vec{x}|\vec{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \tag{6}$$
> where $\vec{\theta} = (\theta_1, ..., \theta_d)^T$ is an unknown parameter vector, θ_i being the probability that $x_i = 1$. Show that the maximum-likelihood estimate for $\vec{\theta}$ is
> $$\hat{\vec{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \vec{x}_k \tag{7}$$

The estimation is simply the sample mean.

The book gives the following equations:

Likelihood of θ with respect to the set of samples.

$$p(\mathcal{D}|\vec{\theta}) = \prod_{k=1}^{n} p(\vec{x}_k|\theta) \tag{8}$$

Log likely hood function

$$l(\theta) = \ln p(D|\theta) = \sum_{k=1}^{n} \ln p(\vec{x}_k|\vec{\theta}) \tag{9}$$

$$\hat{\theta} = \max_{\theta} l(\theta) \tag{10}$$

Take the log likelihood function and maximize θ

$$\ln l(\theta) = \sum_{k=1}^{n} \ln p(\vec{x_k}|\vec{\theta}) \tag{11}$$

$$\text{substitution} \tag{12}$$

$$\sum_{k=1}^{n} \ln(\prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}) \tag{13}$$

$$\text{multiplication property of ln} \tag{14}$$

$$= \sum_{k=1}^{n} \ln(\prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}) \tag{15}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{d} (\ln(\theta_i^{x_i}(1-\theta_i)^{1-x_i})) \tag{16}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{d} (\ln(\theta_i^{x_i}) + \ln((1-\theta_i)^{1-x_i})) \tag{17}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{d} (x_i \ln(\theta_i) + (1-x_i)\ln(1-\theta_i)) \tag{18}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{d} (x_i \ln(\theta_i) + \ln(1-\theta_i) - x_i \ln(1-\theta_i)) \tag{19}$$

$$\text{Apply the gradient operator} \tag{20}$$

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \sum_{i=1}^{d} (\nabla_{\theta_i} x_i \ln(\theta_i) + \nabla_{\theta_i} \ln(1-\theta_i) - \nabla_{\theta_i} x_i \ln(1-\theta_i)) \tag{21}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{d} (\frac{x_i}{\theta_i} - \frac{1}{1-\theta_i} + \frac{x_i}{(1-\theta_i)}) \tag{22}$$

$$\text{Set to zero, and see the factors of } \theta \tag{23}$$

$$0 = \sum_{k=1}^{n} \sum_{i=1}^{d} (\frac{x_i}{\theta_i} - \frac{1}{1-\theta_i} + \frac{x_i}{(1-\theta_i)}) \tag{24}$$

$$\text{combine like terms} \tag{25}$$

$$0 = \sum_{k=1}^{n} \sum_{i=1}^{d} \frac{x_i - \theta_i}{\theta_i - \theta_i^2} \tag{26}$$

$$\text{To goto zero, use numerator for any i = d, and any d} \tag{27}$$

$$0 = \sum_{k=1}^{n} x_k - \theta \tag{28}$$

$$\text{for this to be true } \theta = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{29}$$

3

Another approach, the expected value for σ.

$$\mu = E[x] = \int \vec{x} p(\vec{x}|\vec{\theta}) d\vec{x} \tag{30}$$

$$\theta = E[(\vec{x} - \vec{\theta})(\vec{x} - \vec{\theta})] \tag{31}$$

$$= \int (\vec{x} - \vec{\theta})(\vec{x} - \vec{\theta}) p(\vec{x}|\vec{\theta}) d\vec{x} \tag{32}$$

$$= \int (\vec{x} - \vec{\theta})(\vec{x} - \vec{\theta})^T \prod_{i=1} d(\theta_i^{x_i}(1 - \theta_i)^{1-x_i}) d\vec{x} \tag{33}$$

$$\text{Backtrack steps to 31} \tag{34}$$

3.1 Facts about the Binomial (Bernoulli) Distribution

Definition of a binomial distribution

> If p is the probability that an event will happen in any single trial (called the probability of a success) and $q = 1 - p$ is the probability that it will fail to happen in any single trial, then the probability that the event will happen exactly X times in N trials is given by
>
> $$p(X) = \begin{pmatrix} N \\ X \end{pmatrix} p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \tag{35}$$
>
> where $X = 0, 1, 2, ..., N$ and $N! = N(N-1)(N-2)...1$ and $0! = 1$ by definition. [2, 155]

The mean of such a distribution is defined: $\mu = Np$ and the variance is defined $\sigma^2 = Npq$.

4 Problem 7

If the distribution has another distribution model.

> 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. $P(\omega_1) = P(\omega_2) = 0.5$). Furthermore, we know that $p(x|\omega_1)\ N(0,1)$ but assume that $p(x|\omega_2)\ N(\mu,1)$. (That is, the parameter θ we seek by maximum-likelihood techniques is the mean of the second distribution.) Imagine, however, that the true underlying distribution is $p(x|\omega_2)\ N(1,10^6)$.
>
> 1. What is the value of our maximum-likelihood estimate $\hat{\mu}$ in our poor model, given a large amount of data?
> 2. What is the decision boundary arising from this maximum-likelihood estimate in the poor model?
> 3. Ignore for the moment the maximum-likelihood approach, and use the methods from Chapter 2 to derive the Bayes optimal decision boundary given the true underlying distributions: $p(x|\omega_1)\ N(0,1)$ and $p(x|\omega_2)\ N(1,10^6)$. Be careful to include all portions of the decision boundary.
> 4. Now consider again classifiers based on the (poor) model assumption of $p(x|\omega_2)\ N(\mu|1)$. Using your result immediately above, find a new value for μ that will give lower error than the maximum-likelihood classifier.
> 5. Discuss these results, with particular attention to the role of knowledge of the underlying model.

$$p(\omega_1) = p(\omega_2) = 0.5 \tag{36}$$

$$p(x|\omega_1) \sim N(0,1) \tag{37}$$

$$p(x|\omega_2) \sim N(\mu,1) \tag{38}$$

$$p(x|\omega_2) \sim N(1,10^6) \tag{39}$$

MLE for $\hat{\mu}$

$$p(x|\omega_1, \mu_1) \sim N(0,1) \tag{40}$$

$$p(x|\hat{\omega_2}, \mu_2) \sim N(1, 10^6) \tag{41}$$

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) \tag{42}$$

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2] \tag{43}$$

$$p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2] \tag{44}$$

$$\ln(p(x|\omega_1)) = \ln(\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2]) \tag{45}$$

$$\ln(p(x|\hat{\omega_2})) = \ln(\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2]) \tag{46}$$

$$\ln(p(x|\omega_1)) = \ln(\frac{1}{\sqrt{2\pi}} \exp[-\frac{x^2}{2}]) \tag{47}$$

$$\ln(p(x|\omega_1)) = \ln(\frac{1}{\sqrt{2\pi}}) + \ln(\exp[-\frac{x^2}{2}]) \tag{48}$$

$$\ln(p(x|\omega_1)) = \ln 1 - \ln\sqrt{2\pi} - \frac{x^2}{2} \tag{49}$$

$$\ln(p(x|\omega_1)) = 0 - \frac{1}{2}\ln 2\pi - \frac{x^2}{2} \tag{50}$$

$$\ln(p(x|\omega_1)) = -\frac{1}{2}\ln 2\pi - \frac{x^2}{2} \tag{51}$$

$$\ln(p(x|\omega_2, \mu_2)) = \ln(\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu_2}{1})^2]) \tag{52}$$

$$\ln(p(x|\omega_2, \mu_2)) = \ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}(x-\mu_2)^2 \tag{53}$$

$$\ln(p(x|\omega_2, \mu_2)) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}(x-\mu_2)^2 \tag{54}$$

$$\frac{d}{d\mu_2}\ln(p(x|\omega_2, \mu_2)) = -(2)\cdot(-\frac{1}{2})(x-\mu_2) \tag{55}$$

$$\frac{d}{d\mu_2}\ln(p(x|\omega_2, \mu_2)) = (x-\mu_2) \tag{56}$$

$$\sum_{k=1}^{n}\frac{d}{d\mu_2}\ln(p(x|\omega_2, \mu_2)) = \sum_{k=1}^{n}(x_k-\mu_2) = 0 \therefore \mu_2 = \sum_{k=1}^{n} x_k \tag{57}$$

4.1 Silly question

Was there a decision boundary description made in this chapter that was different than what we saw in chapter 2? In chapter two, we saw a a concept called the discriminating function, denoted $g_i(x)$ and said that a sample x satisfied a particular $g_i(x)$ in conditions specified in terms of a partial-continuous function. In the two category case, we saw a special case where signs of the difference were enough to discriminate.

5 Problem 8

Consider an extreme case of general issue discussed in Problem 7, one in which it is possible that the maximum-likelihood solution leads to a worst possible classifier, that is, one with an error that approaches 100% (in probability). Suppose our data in fact comes from two one-dimensional distributions of the forms:

$$p(x|\omega_1) = [(1-k)\delta(x-1) + k\delta(x+X)] \tag{58}$$

$$p(x|\omega_2) = [(1-k)\delta(x+1) + k\delta(x-X)] \tag{59}$$

where X is positive, $0 \leq k < 0.5$ represents the portion of the total probability mass concentrated at the point $\pm X$ and $\delta(\cdot)$ is the Dirac delta function. Suppose our poor models are of the form $p(x|\omega_1, \mu_1)\ N(\mu_1, \sigma_1^2)$ and $p(x|\omega_2, \mu_2)\ N(\mu_2, \sigma_2^2)$ and we form a maximum likelihood classifier.

1. Consider the symmetries in the problem and show that in the infinite data case the decision boundary will always be at $x = 0$, regardless k and X.
2. Recall that the maximum-likelihood estimate of either mean, $\hat{\mu}_i$is the mean of its distribution. For a fixed k, find the value of X such that the maximum likelihood estimates of the means "switch" that is $\hat{\mu}_1 \geq \hat{\mu}_2$.
3. Plot the true distributions and the Gaussian estimates for the particular case $k = 0.2$ and $X = 5$. What is the classification error in this case?
4. Find a dependence $X(k)$ which will guarantee that the estimated mean $\hat{\mu}_1$ of $p(x|\omega_1)$ is less than zero (By symmetry, this will also ensure $\hat{\mu}_2$.)
5. Given your $X(k)$ just derived, state the classification error in terms of k.
6. Suppose we constrained our model space such that $\sigma_1^2 = \sigma_2^2 = 1$ (or indeed any other constant). Would that change the above results?
7. Discuss how if our model is wrong (here, does not include the delta functions), the error can approach 100% (in probability). Does this surprising answer arise because we have found some local minimum in parameter space?

As stated the assumption of MLE is in play, namely $p(x|\omega_1)\ N(\mu_1, \sigma_1^2)$ and $p(x|\omega_2^2)\ N(\mu_2, \sigma_2)$. Using the definitions of μ and σ^2 we can derive the approximate sample mean and sample variance.

$$\mu_1 = E[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{60}$$

$$= \int_{-\infty}^{\infty} x[(1-k)\delta(x-1) + k\delta(x+X)]dx \tag{61}$$

$$= \int_{-\infty}^{\infty} [(x-kx)\delta(x-1) + kx\delta(x+X)]dx \tag{62}$$

$$= \int_{-\infty}^{\infty} (x-kx)\delta(x-1)dx + \int_{-\infty}^{\infty} kx\delta(x+X)dx \tag{63}$$

$$= (x-kx) + kx \tag{64}$$

$$\sigma_1^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \tag{65}$$

6 Problem 15

References

[1] R. O. Duda, P. E. Hart, and D. E. Stork. *Pattern Classification.* Wiley and Sons, 2nd edition, 2000.

[2] L. J. S. Murray R. Spiegel. *Schaum's Outlines: Statistics.* McGraw Hill Companies, Inc, New York, 1999, 1988, 1961.