6.867: Exercises (Week 1)

Sept 8, 2016

1. (Bishop 1.11) We find ourselves with a data set consisting of the measured weights of a bunch of fish caught during an afternoon of fishing. We decide to model the distribution of these weights using a Gaussian distribution.

   Why might this not be a great modeling choice?

   Our goal is to select parameters μ, σ^2 of the Gaussian distribution in order to maximize the likelihood of our data, $\mathcal{D} = \{x^{(1)}, \ldots, x^{(n)}\}$. The parameters that maximize the log likelihood of the data, will also maximize the likelihood (due to its monotonicity) and the form is easier to deal with.

   Recall that the pdf of a Gaussian distribution is given by

   $$p_X(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \ .$$

   If we assume that the process whereby we caught the fish made their weights independent and identically distributed, then

   $$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_i p_X(x^{(i)} \mid \mu, \sigma^2) \ .$$

   The log likelihood function is then

   $$\log p(\mathcal{D} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi) \ .$$

   By setting its derivatives with respect to μ and σ^2 equal to zero and solving[1], verify that the maximum likelihood estimates of μ and σ are given by

   $$\mu_{\mathbf{ml}} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$$

[1] In the exercises of this class, we often solve the system $\nabla_\theta L(\theta) = 0$ for θ where θ is the parameter to be estimated (here, (μ, σ^2)) and L is the loss function to be minimized (here, $-\log p$). From calculus class, we know that this is a necessary condition of θ being a local extremum of $L : U \to \mathbb{R}$ (where U is an open subset of $\mathbb{R}^n$). If the loss function L is convex, this is also a sufficient condition of θ being a global minimum of L. In exercises, we often consider a convex loss function L where our approach of solving $\nabla_\theta L(\theta) = 0$ is justified.

$$\sigma^2_{\mathbf{ml}} = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)} - \mu_{\mathbf{ml}})^2$$

This solution may be different than the estimator you have previously seen for σ^2. See the discussion at the bottom of Bishop page 27 for an explanation.

2. As it happens, we caught 6 mega-guppies (a tasty type of fish), with these weights:

$$\mathcal{D}_0 = \{0.9, 1, 1.1, 1.2, 3, 3.1\} \ .$$

We looked in the USDA handbook which told us that the variance of the weight of North American mega-guppies is $\sigma^2 = 0.5^2 = 0.25$.

Find the maximum likelihood value of $\mu_{\mathbf{ml}}$ for $\mathcal{D}_0$. What is the data likelihood $p(\mathcal{D}_0|\mu_{\mathbf{ml}})$ and the data log likelihood?

3. Now, what if we ignore the USDA value of σ^2 and decide to estimate it ourselves? Find the maximum likelihood estimates $\mu_{\mathbf{ml}}$ and $\sigma^2_{\mathbf{ml}}$ of μ and σ^2 for our data set $\mathcal{D}_0$. What is the data likelihood $p(\mathcal{D}_0|\mu_{\mathbf{ml}}, \sigma^2_{\mathbf{ml}})$ and the data log likelihood?

What are the advantages and disadvantages of this model versus the one from the previous problem?

4. A supervillain has our hero trapped in an invisible one-dimensional force-field (hero can only move in one dimension) and we know that it has finite extent. Using a drone flying overhead, we make several measurements of the hero's position.

We wish to estimate the boundaries of the force-field given samples of the hero's position.

If we knew that our data are drawn uniformly from a finite interval, $[a, b]$, then we might want to find $a_{\mathbf{ml}}, b_{\mathbf{ml}}$ to maximize the likelihood of $\mathcal{D}$.

For our data set $\mathcal{D} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$, what are the maximum likelihood parameter estimates $a_{\mathbf{ml}}$ and $b_{\mathbf{ml}}$? What is the data likelihood $p(\mathcal{D}|a_{\mathbf{ml}}, b_{\mathbf{ml}})$?

Is this model of the hero data a good one? Why or why not?

5. Consider a simple prediction problem where the variable y is 0 or 1, and we know the probability $\Pr(y = 1)$ exactly. We use 0-1 loss (where g is the predicted ("g" for guessed) value and a is the actual value):

$$L(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{otherwise} \end{cases}$$

This type of loss makes sense when y takes on values in a discrete set. We would like to minimize the risk (expectation of the loss). What value g should you predict? Show the proof.

6. Now, imagine that you are working in the emergency room, predicting whether a patient presenting with chest pain is having a heart attack (a prediction of "H") or indigestion (a prediction of "I"). In this case, the two mistakes are not equally bad. We have the following

asymmetric loss function (where g is the predicted ("g" for guessed) value and a is the actual value):

$$L(g, a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{if } g = \text{"H" and } a = \text{"I"} \\ 10 & \text{if } g = \text{"I" and } a = \text{"H"} \end{cases}$$

Assuming you know $\Pr(Y = \text{"H"})$ and $\Pr(Y = \text{"I"})$ and that $\Pr(Y = \text{"H"}) + \Pr(Y = \text{"I"}) = 1$, in what cases should you predict "H"?

7. Pigeons[2], when put in a situation where $\Pr(y = 1) = p$ and $\Pr(y = 0) = 1 - p$, will select option 1 with probability p and option 0 with probability $1 - p$. What is the expected 0-1 loss for the pigeons' decision rule? What is the optimal decision rule and its expected loss?

   Actually, people[3] do this too!

8. (This is harder than the others)

   Consider the problem of predicting a real-valued random variable $y \in \mathbb{R}$. Assume we know the pdf $p_Y(a)$. Suppose that now we use the loss function $L(g, a)$ (where g is the predicted value and a is the actual value)

   $$L(g, a) = |g - a| \quad .$$

   What value should you predict to minimize the expected value of the loss?

[2] "Probability-Matching in the Pigeon", Donald H. Bullock and M. E. Bitterman, *The American Journal of Psychology* , Vol. 75, No. 4 (Dec., 1962), pp. 634-639

[3] "Banking on a Bad Bet: Probability Matching in Risky Choice is Linked to Expectation Generation," *Psychological Science*, Vol. 22, No. 6 (2011).