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?

   Solution:

   Maybe they come from different species, so we could expect the distribution to be multi-modal. Also, the Gaussian has infinite tails, so it will assign positive probability to fish with negative weight.

   Our goal is to select parameters μ, σ^2 of the Gaussian distribution in order to maximize the likelihood of our data, $\mathcal{D} = \{x^{(1)}, \dots, x^{(n)}\}$. 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

[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.

maximum likelihood estimates of μ and σ are given by

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

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

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.

Solution: Taking the partial derivatives of the log likelihood with respect to μ and σ^2 results in

$$\frac{\partial \log \Pr(x \mid \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x^{(n)} - \mu)$$

$$\frac{\partial \log \Pr(x \mid \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x^{(n)} - \mu)^2 - \frac{N}{2\sigma^2}$$

Setting the partial derivative with respect to μ to 0 gives

$$0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x^{(n)} - \mu) = \frac{1}{\sigma^2}\sum_{n=1}^{N} x^{(n)} - \frac{1}{\sigma^2}\sum_{n=1}^{N}\mu = \frac{1}{\sigma^2}\sum_{n=1}^{N} x^{(n)} - \frac{1}{\sigma^2}N\mu$$

so

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$$

For the Gaussian, we are fortunate that our estimate for the mean is independent of the variance.

Setting the partial derivative with respect to σ^2 (note – not σ) to 0 gives

$$0 = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x^{(n)} - \mu)^2 - \frac{N}{2\sigma^2}$$

$$\frac{N}{2\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x^{(n)} - \mu)^2$$

so

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)} - \mu)^2$$

We can relate this quantity back to by replacing μ with $\mu_{\mathbf{ml}}$ found above, since as we noticed, the $\mu_{\mathbf{ml}}$ does not depend on the variance.

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?

Solution:

$$\mu_{\mathbf{ml}} = 1.71666$$

$$p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}) = 5.387577e - 06$$

$$\log(p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}})) = -12.131415$$

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?

Solution:

Our model fits the data better. But it might be that it would have been better to use the other variance because it was based on a larger sample. But it might have been better to use our estimate because the local population of fish has a different distribution. It all depends on what you're trying to estimate and what assumptions you can make about the data and about the prior information.

Solution:

$$\mu_{\mathbf{ml}} = 1.716666$$

$$\sigma^2_{\mathbf{ml}} = 0.898055$$

$$p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}, \sigma_{\mathbf{ml}}) = 0.000277$$

$$\log(p(\mathcal{D}_0 \mid \mu_{\mathbf{ml}}, \sigma_{\mathbf{ml}})) = -8.191061$$

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?

Solution:

It might not be good if we have reason to think that for the hero, some parts of the force field are more interesting or comfortable than others. Also, if we are sampling positions finely in time, they will be very highly correlated with one another (not iid).

Solution:

The likelihood of the data is:

$$\prod_{i=1}^{n} \begin{cases} (b_{\mathbf{ml}} - a_{\mathbf{ml}})^{-1} & \text{if } a_{\mathbf{ml}} \leqslant x^i \leqslant b_{\mathbf{ml}} \\ 0 & \text{otherwise} \end{cases}$$

We can see that if $a_{\mathbf{ml}} > x^i$ or $b_{\mathbf{ml}} < x^i$, for any x^i, then the likelihood of the whole data set must be 0. So, we should pick $b_{\mathbf{ml}}$ to be as small as possible subject to the constraint that $b_{\mathbf{ml}} \geqslant x^i$, which means $b_{\mathbf{ml}} = \max_i x^i$. Similarly, $a_{\mathbf{ml}} = \min_i x^i$.

For $\mathcal{D}_0 = \{0.9, 1, 1.1, 1.2, 3, 3.1\}$ (the data from the previous question):

$$a_{\mathbf{ml}} = 0.9 \quad b_{\mathbf{ml}} = 3.1$$

$$p(\mathcal{D}_0 \mid a_{\mathbf{ml}}, b_{\mathbf{ml}}) = 0.008820$$

$$\log(p(\mathcal{D}_0 \mid a_{\mathbf{ml}}, b_{\mathbf{ml}})) = -4.730744$$

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.

Solution: The risk is the expected value of the loss. The notation $1\{y = g\}$ is an indicator which is 1 when the value in the brackets is True else 0.

$$\begin{aligned}
E_y[L(g,y)] &= \sum_y L(g,y)\,\Pr(Y=y) \\
&= \sum_y (1-1\{y=g\})\,\Pr(Y=y) \\
&= \sum_y \Pr(Y=y) - \sum_y 1\{y=g\}\Pr(Y=y) \\
&= 1-\Pr(Y=g)
\end{aligned}$$

This quantity is minimized when g is chosen to be the *mode* (most likely value) of the distribution.

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 $\Pr(Y = ''H'')$ and $\Pr(Y = ''I'')$ and that $\Pr(Y = ''H'') + \Pr(Y = ''I'') = 1$, in what cases should you predict "H"?

Solution: We begin by first constructing a table representing all the possible outcomes of this problem:

	$a = ''I''$	$a = ''H''$
$g = ''I''$	0	10
$g = ''I''$	1	0

Each row represents the possible outcomes from picking a particular guess and the columns represent the true diagnosis. We can then write down the expected loss when we pick each guess.

$$\begin{aligned}
L_{g=I} &= L(g=\text{“I”}, a=\text{“I”})p(a=\text{“I”}) + L(g=\text{“I”}, a=\text{“H”})p(a=\text{“H”}) \\
&= 10p(a=\text{“H”}) \\
L_{g=H} &= L(g=\text{“H”}, a=\text{“I”})p(a=\text{“I”}) + L(g=\text{“H”}, a=\text{“H”})p(a=\text{“H”}) \\
&= p(a=\text{“I”}) \\
&= 1-p(a=\text{“H”})
\end{aligned}$$

We should guess “I” whenever $L_{g=I} < L_{g=H}$ which implies,

$$\begin{aligned} L_{g=I} &< L_{g=H} \\ 10p(a = \text{“H”}) &< 1 - p(a = \text{“H”}) \\ p(a = \text{“H”}) &< 1/11 \end{aligned}$$

And therefore we should guess “I” whenever $p(a = \text{“H”}) < 1/11$ and “H” otherwise.

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!

Solution: The loss is 0 when the pigeon's random choice g agrees with the independent draw from the underlying distribution y. So, the loss is:

$$1 - (\Pr(g = 0)\Pr(y = 0) + \Pr(g = 1)\Pr(y = 1))$$

Recalling that $\Pr(g = 1) = \Pr(y = 1) = p$, we have that loss is

$$2p(1 - p)$$

Note that we saw that the optimal decision rule is to pick the mode and that the loss of that rule is:

$$1 - \max(p, (1 - p))$$

Note that for $p = 0.5$ the losses are the same. But, for other values, say $p = 0.6$, the pigeons' loss is 0.48 and the optimal loss is 0.4. Proving that pigeons are not so good at decision theory.

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)

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

   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).

Solution: The risk is the expected value of the loss.

$$\begin{aligned}
E_Y[L(g,y)] &= \int_{-\infty}^{\infty} p_Y(y)L(g,y)\ dy \\
&= \int_{-\infty}^{\infty} p_Y(y)|g-y|\ dy \\
&= \int_{-\infty}^{g} p_Y(y)(g-y)\ dy + \int_{g}^{\infty} p_Y(y)(y-g)\ dy \\
&= g\left(\int_{-\infty}^{g} p_Y(y)\ dy - \int_{g}^{\infty} p_Y(y)\ dy\right) + \int_{g}^{\infty} y p_Y(y)\ dy - \int_{-\infty}^{g} y p_Y(y)\ dy
\end{aligned}$$

We want to find the value of g that minimizes risk, so we set the derivative of the risk with respect to g to zero. Using the product rule for derivatives and the Fundamental Theorem of Calculus, we get:

$$\begin{aligned}
\frac{dE_Y[L(g,y)]}{dg} &= 2gp_Y(g) + \left(\int_{-\infty}^{g} p_Y(y)\ dy - \int_{g}^{\infty} p_Y(y)\ dy\right) - 2gp_Y(g) \\
&= \int_{-\infty}^{g} p_Y(y)\ dy - \int_{g}^{\infty} p_Y(y)\ dy = 0 \\
&= \Pr(y \leqslant g) - \Pr(y \geqslant g) = 0
\end{aligned}$$

A value of g satisfying this condition is known as the *median* of the distribution $\Pr(a)$; there's as much mass to the left of it as there is to the right. Note that this is not an algorithm for computing the optimal guess.