Foundations of Machine Learning

Brett Bernstein

August 22, 2018

Conditional Probability Models: Concept Check

Conditional Probability Models

MLE Learning Objectives

- Define the likelihood of an estimate of a probability distribution for some data $\mathcal{D}$.
- Define a parameteric model, and some common parameteric families.
- Define the MLE for some parameter θ of a probability model.
- Be able to find the MLE using first order conditions on the log-likelihood.

Conditional Probability Models

- Describe the basic structure of a linear probabilistic model, in terms of (i) a parameter θ of the probablistic model, (ii) a linear score function, (iii) a transfer function (kin to a "response function" or "inverse link" function, though we've relaxed requirements on the parameter theta).
- Explain how we can use MLE to choose w, the weight vector in our linear function (in (ii) above).
- Give common transfer functions for (i) bernoulli, (ii) poisson, (iii) gaussian, and (iv) categorical distributions. Explain why these common transfer functions make sense (in terms of their codomains).
- Explain the equivalence of EMR and MLE for negative log-likelihood loss.

MLE/Conditional Probability Model Concept Check Question

1. In each of the following, assume $X_1, \ldots, X_n$ are an i.i.d. sample from the given distribution.

(a) Compute the MLE for p assuming each $X_i \sim \text{Geom}(p)$ with PMF $f_X(k) = (1 - p)^{k-1}p$ for $k \in \mathbb{Z}_{\geq 1}$.

(b) Compute the MLE for λ assuming each $X_i \sim \text{Exp}(\lambda)$ with PDF $f_X(x) = \lambda e^{-\lambda x}$.

Solution.

(a) The likelihood L is given by

$$L(p; x_1, \ldots, x_n) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p$$

giving a log-likelihood

$$\log L(p; x_1, \ldots, x_n) = n \log p + \left(\sum_{i=1}^{n} x_i - 1 \right) \log(1-p).$$

Differentiating gives

$$\frac{d}{dp} \log L(p; x_1, \ldots, x_n) = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - 1}{1-p}.$$

Solving for a critical point we get

$$\frac{d}{dp} \log L(p; x_1, \ldots, x_n) = 0 \iff \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{p} \iff p = \frac{n}{\sum_{i=1}^{n} x_i}.$$

By the first or second derivative tests, this is the maximum. Thus the answer is

$$\hat{p}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

(b) The likelihood L is given by

$$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

giving a log-likelihood

$$\log L(\lambda; x_1, \ldots, x_n) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Differentiating gives

$$\frac{d}{dp} \log L(p; x_1, \ldots, x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Solving for a critical point we get

$$\frac{d}{dp}\log L(p; x_1, \dots, x_n) = 0 \iff \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

By the first or second derivative tests, this is a maximum. Thus the answer is

$$\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

2. We want to fit a regression model where $Y|X = x \sim \text{Unif}([0, e^{w^T x}])$ for some $w \in \mathbb{R}^d$. Given i.i.d. data points $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, give a convex optimization problem that finds the MLE for w.

   Solution. The likelihood L is given by

   $$L(w; x_1, y_1, \dots, x_n, y_n) = \prod_{i=1}^{n} \frac{\mathbf{1}(y_i \le e^{w^T x_i})}{e^{w^T x_i}}.$$

   Taking logs we get

   $$-\sum_{i=1}^{n} w^T x_i = -w^T\left(\sum_{i=1}^{n} x_i\right)$$

   if $y_i \le \exp(w^T x_i)$ for all i, or $-\infty$ otherwise. Thus we obtain the linear program

   $$\begin{aligned} &\text{minimize} \quad && w^T\left(\sum_{i=1}^{n} x_i\right) \\ &\text{subject to} \quad && \log(y_i) \le w^T x_i \quad \text{for } i = 1, \dots, n. \end{aligned}$$

3. Explain why softmax is related to computing the maximum of a list of values.

   Solution. Let $x_1, \dots, x_n \in \mathbb{R}$. Let $\text{ArgMax}(x_1, \dots, x_n)$ denote a 1-hot encoding of the argmax function:

   $$\text{ArgMax}(x_1, \dots, x_n) = \left(\mathbf{1}(\arg\max_i x_i = 1), \dots, \mathbf{1}(\arg\max_i x_i = n)\right).$$

   Recall that softmax has the following definition:

   $$\text{softmax}_\lambda(x_1, \dots, x_n) = \frac{1}{\sum_{i=1}^{n} e^{\lambda x_i}}\left(e^{\lambda x_1}, \dots, e^{\lambda x_n}\right),$$

   where $\lambda > 0$ is a fixed parameter. We claim that softmax is a differentiable approximation to ArgMax. Consider what happens when we let $x_j \to \infty$ while keeping the other values fixed. Then

   $$\frac{e^{\lambda x_j}}{\sum_{i=1}^{n} e^{\lambda x_i}} \to 1$$

and

$$\frac{e^{\lambda x_k}}{\sum_{i=1}^{n} e^{\lambda x_i}} \to 0$$

for all $k \neq j$. For example, suppose $x_1 = 1$, $x_2 = -3$, $x_3 = 5$. Then

$$\text{softmax}_1(1, -3, 5) = (0.0180, 0.0003, 0.9817)$$

while

$$\text{ArgMax}(1, -3, 5) = (0, 0, 1).$$

4. Suppose x has a Poisson distribution with unknown mean θ:

$$p(x|\theta) = \frac{\theta^x}{x!}\exp(-\theta), \qquad x = 0, 1, \cdots$$

Let the prior for θ be a gamma distribution:

$$p(\theta|\alpha, \beta) = \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)}\exp(-\beta\theta), \qquad \theta > 0$$

where Γ is the gamma function. Show that, given an observation x, the posterior $p(\theta|x, \alpha, \beta)$ is a gamma distribution with updated parameters $(\alpha', \beta') = (\alpha + x, \beta + 1)$. What does this tell you about the Poisson and gamma distributions?

Solution. From Bayes' theorem[1], we have:

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto (\theta^x \exp(-\theta)) \left(\theta^{\alpha-1}\exp(-\beta\theta)\right) \\ &= \theta^{x+\alpha-1}\exp\left(-(\beta+1)\theta\right) \\ &\propto \mathcal{G}(\alpha + x, \beta + 1) \end{aligned}$$

This shows that the gamma is the conjugate prior to the Poisson. Also, note here we exploit a common trick: we manipulate the numerator, ignoring constants independent of θ. If we can recognize the functional form as belonging to a distribution family we know, we can simply identify the parameters and trust that the distribution normalizes!

[1] Actually from Roman Garnett, from whom this problem was taken.