Machine Learning – Brett Bernstein

Week 10 Lab: Concept Check Exercises

Conditional Probability Models

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

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

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 $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, give a convex optimization problem that finds the MLE for w.

   Solution. The likelihood L is given by

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

   Taking logs we get

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

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

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

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

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

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

Recall that softmax has the following definition:

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

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

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

while

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