DS-GA 3001.009 Applied Statistics: Homework #4 Solutions

Due on Thursday, November 2, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. In this problem we prove another formulation of the KL divergence, called the Donsker-Varahdan variational formula:

$$D_{\mathrm{KL}}(P\|Q) = \max_f \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f].$$

Here P, Q are two probability distributions on a common discrete set $\mathcal{X}$, the function $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function on $\mathcal{X}$, $\mathbb{E}_P[f]$ is a shorthand for $\mathbb{E}_{X\sim P}[f(X)]$, and log is the natural logarithm (with base e).

(a) Given f, define another probability distribution Q_f on $\mathcal{X}$ as $Q_f(x) = cQ(x)e^{f(x)}$. Show that the normalization constant c is given by $c = (\mathbb{E}_Q[e^f])^{-1}$.

(b) Use the definition of the KL divergence, show that

$$D_{\mathrm{KL}}(P\|Q) - D_{\mathrm{KL}}(P\|Q_f) = \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f].$$

(c) Conclude that

$$D_{\mathrm{KL}}(P\|Q) \geq \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]$$

for every f, with equality if and only if $f(x) = \log \frac{P(x)}{Q(x)} + c'$ for some constant c'.

Solution:

(a) Since Q_f is a probability distribution, we have

$$1 = \sum_{x\in\mathcal{X}} Q_f(x) = c \cdot \sum_{x\in\mathcal{X}} Q(x)e^{f(x)} = c \cdot \mathbb{E}_Q[e^f] \Longrightarrow c = \frac{1}{\mathbb{E}_Q[e^f]}.$$

(b) By definition of the KL divergence, we have

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) - D_{\mathrm{KL}}(P\|Q_f) &= \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} - \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q_f(x)} \\ &= \sum_{x\in\mathcal{X}} P(x) \log \frac{Q_f(x)}{Q(x)} \\ &= \sum_{x\in\mathcal{X}} P(x) \log(ce^{f(x)}) \\ &= \sum_{x\in\mathcal{X}} P(x)f(x) + \log c \\ &= \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f], \end{aligned}$$

where the last identity uses (a).

(c) Since $D_{\mathrm{KL}}(P\|Q_f) \geq 0$, the target inequality holds. The equality holds if and only if $D_{\mathrm{KL}}(P\|Q_f) \geq 0$, or $P = Q_f$, i.e. $P(x) = cQ(x)e^{f(x)}$ for some constant $c > 0$. Consequently, in this case we have

$$f(x) = \log \frac{P(x)}{Q(x)} - \log c = \log \frac{P(x)}{Q(x)} + c',$$

with $c' = -\log c$.

2. In this problem, we apply the EM algorithm to a dataset consisting of both complete and missing data. Specifically, let $(x_1, y_1), \cdots, (x_{n+m}, y_{n+m})$ be i.i.d. drawn from some $p_\theta(x, y) = \exp(\langle \theta, T(x, y)\rangle - A(\theta))h(x, y)$ in an exponential family, but assume that we only observe $(x_1, y_1), \cdots, (x_n, y_n)$ and $y_{n+1}, y_{n+2}, \cdots, y_{n+m}$.

(a) Write out the incomplete log likelihood for the observations (up to additive constants).

(b) Describe the EM algorithm for the MLE computation. You should give the details of both E and M steps; you need not give proofs.

Solution:

(a) The incomplete log likelihood for the observations is

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log p_\theta(x_i, y_i) + \sum_{j=n+1}^{n+m} \log p_\theta(y_j) \\ &= \sum_{i=1}^{n} (\langle \theta, T(x_i, y_i)\rangle - A(\theta)) + \sum_{j=n+1}^{n+m} (A_{y_j}(\theta) - A(\theta)) + C,\end{aligned}$$

where

$$A_y(\theta) = \log \left[\int \exp(\langle \theta, T(x, y)\rangle) h(x, y) dx \right].$$

(b) Similar to the EM algorithm in class, to move from $\theta^{(t)}$ to $\theta^{(t+1)}$:

- E-step: compute the vector $(\mu_1^{(t+1)}, \cdots, \mu_{n+m}^{(t+1)})$ with

$$\mu_i^{(t+1)} = \begin{cases} T(x_i, y_i) & \text{if } 1 \leq i \leq n, \\ \mathbb{E}_{X \sim p_{\theta^{(t)}}(\cdot|y_i)}[T(X, y_i)] & \text{if } n+1 \leq i \leq n+m. \end{cases}$$

  In other words, for complete data we use the true value of T, and for missing data we compute the expected value of T based on the conditional distribution learned so far.
- M-step: compute $\theta^{(t+1)}$ from the estimating equation

$$\nabla A(\theta^{(t+1)}) = \frac{1}{n+m} \sum_{i=1}^{n+m} \mu_i^{(t+1)}.$$

3. Coding: we will implement the EM algorithm in the spatial test dataset. This dataset contains 26 pairs (x_i, y_i), but 13 of the y_i values are missing. Here we model the joint distribution of (x, y) by a bivariate Gaussian distribution

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

which is an exponential family with 5 parameters $(\mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ (note that $\Sigma_{12} = \Sigma_{21}$). We aim to estimate the mean and covariance parameters, and then fit the missing values in the dataset.

The detailed EM iteration is slightly involved to derive here, so we have implemented most of the steps. Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/y393htww`. Although not required, you are encouraged to understand why the current codes implement the EM algorithm correctly.

Be sure to submit a pdf with your codes, outputs, and colab link.

Solution: see `https://tinyurl.com/yc6ahxxh`.