DS-GA 3001.001 Applied Statistics: Homework #3

Due on Thursday, October 24, 2024

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

1. Compute the convex conjugate $f^\star(t)$ for the following functions:

   (a) $f(x) = \frac{x^2}{2}$;

   (b) $f(x) = e^x$.

   In your computation, note that $f^\star(t)$ is allowed to take the value $+\infty$.

2. Let $\mathcal{N}(\mu, \sigma^2)$ be the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Find the expression of

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \sigma_1^2) \| \mathcal{N}(\mu_2, \sigma_2^2)).$$

3. 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). You can use $A_y(\theta)$ in the lecture note but need to define it explicitly.

   (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.

4. Coding I: we will implement the EM algorithm for learning Gaussian mixtures. Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/477e9hfw`. Be sure to submit a pdf with your codes, outputs, and colab link.

5. Coding II: 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.