Math Foundations of ML, Fall 2022

Homework #6

Due Monday November 14, at 5:00pm ET

As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

1. Suppose that two random variables (X, Y) have joint pdf $f_{X,Y}(x, y)$. Find an expression for the pdf $f_Z(z)$ where $Z = X + Y$. You can start by realizing that

$$F_Z(u|X = \beta) = \mathrm{P}\left(Z \le u|X = \beta\right) = \mathrm{P}\left(Y \le u - \beta|X = \beta\right).$$

You can combine the expressions above by integrating over β, and see that the resulting expression corresponds to an integral of $f_{X,Y}(x, y)$ over a half plane. From this, you can get the pdf for Z by applying the Fundamental Theorem of Calculus. How does your expression simplify if X and Y are independent? (Convolution!)

2. Let $X_1, X_2, \ldots$ be independent uniform random variables,

$$X_n \sim \text{Uniform}(-1/2, 1/2), \quad \text{meaning} \quad f_X(x) = \begin{cases} 1, & -1/2 \le x \le 1/2 \\ 0, & \text{otherwise.} \end{cases}$$

   (a) What is the density function for $Y = X_1 + X_2 + X_3$? (If you compute this correctly, you will meet an old friend.)

   (b) The *moment generating function* of a random variable is

$$\varphi_X(t) = \mathrm{E}\left[e^{tX}\right].$$

   It is a fact that if $\varphi_X(t) = \varphi_W(t)$ for all t, then X and W have the same distribution. It is a fact that if $G \sim \text{Normal}(0, \sigma^2)$, then $\varphi_G(t) = e^{\sigma^2 t^2/2}$. Let

$$Y_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n.$$

   Find an expression for $\varphi_{Y_N}(t)$. Plot $\varphi_{Y_N}(t)$ and $\varphi_G(t)$ for $\sigma^2 = \text{var}(Y) = \text{var}(X_n) = 1/12$ on the same set of axes for $N = 1, 2, 5, 10$ and $0 \le t \le 5$. What might you conclude about Y_N as $N \to \infty$? (***Bonus question**: argue rigorously that $\varphi_{Y_N}(t) \to \varphi_G(t)$ for all t.*)

   (c) It is a fact that if $\phi(z)$ is a monotonically increasing function, then for any random variable Z,

$$\mathrm{P}\left(Z > u\right) = \mathrm{P}\left(\phi(Z) > \phi(u)\right).$$

   Use $\phi(z) = e^{tz}$ and the Markov inequality to derive a bound on $\mathrm{P}\left(Z_N > u\right)$, where

$$Z_N = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

   For the special case of $t = 4u/N$, compare this bound, as a function of u, to that obtained using the Chebsyshev inequality.

3. Let $Z_1, \ldots, Z_N$ be a sequence of independent Gaussian random variables with mean 0 and variance 1. You observe the random vector X in $\mathbb{R}^N$ that is generated through the autoregressive process

$$X_k = \begin{cases} Z_1, & k = 1 \\ aX_{k-1} + Z_k, & k > 1. \end{cases}$$

Given $X = \boldsymbol{x}$, find the MLE for $a \in \mathbb{R}$. (Hint: Conditional independence.) (Further hint: The conditional independence structure makes this a Markov process, meaning that we can factor the distribution for $X \in \mathbb{R}^N$ as

$$f_X(\boldsymbol{x}) = f_{X_1}(x_1) f_{X_2}(x_2|x_1) f_{X_3}(x_3|x_2) \cdots f_{X_N}(x_N|x_{N-1}).$$

)

4. Let X be a Gaussian random vector taking values in $\mathbb{R}^N$, let E be a Gaussian random vector taking values in $\mathbb{R}^M$, and let $\boldsymbol{A}$ be a $M \times N$ matrix. We have

$$X \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{R}_x), \quad E \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{R}_e), \quad X, E \text{ independent.}$$

We will make observation of the random vector

$$Y = \boldsymbol{A}X + E.$$

(a) From the lecture notes, it is clear that Y is a Gaussian random vector in $\mathbb{R}^M$ and that $\mathrm{E}[Y] = \boldsymbol{0}$. Find the covariance matrix for the Gaussian random vector $\begin{bmatrix} X \\ Y \end{bmatrix}$ that takes values in $\mathbb{R}^{N+M}$.

(b) Suppose we observe $Y = \boldsymbol{y}$. What is the minimum mean-square error estimate of X given $Y = \boldsymbol{y}$?

(c) Suppose $\boldsymbol{R}_x = \sigma_x^2 \mathbf{I}$ and $\boldsymbol{R}_e = \sigma_e^2 \mathbf{I}$. In this case, your MMSE estimator should look familiar, and you should see immediately that $\hat{x}_{MMSE}$ is in the row space of $\boldsymbol{A}$. What are the $\hat{\alpha}_n$ is the expression below?

$$\hat{\boldsymbol{x}}_{MMSE} = \sum_{n=1}^{N} \alpha_n \boldsymbol{v}_n, \quad \text{where the } \boldsymbol{v}_n \text{ are the right singular vectors of } \boldsymbol{A}.$$

(d) Take $\boldsymbol{R}_x$ and $\boldsymbol{R}_e$ as in part (c), and assume that $\boldsymbol{A}$ has full column rank. What is MSE $\mathrm{E}[\|\hat{\boldsymbol{x}}_{MMSE} - X\|_2^2]$ of the MMSE estimate $\hat{\boldsymbol{x}}_{MMSE}$?

5. Let $\boldsymbol{A}$ be an $M \times N$ matrix with full column rank. Let E be a Gaussian random vector in $\mathbb{R}^M$ with mean $\boldsymbol{0}$ and covariance $\boldsymbol{R}_e$. Suppose we observe

$$Y = \boldsymbol{A}\boldsymbol{\theta}_0 + E,$$

where $\boldsymbol{\theta} \in \mathbb{R}^N$ is unknown.

(a) What is the distribution of Y and how does it depend on $\boldsymbol{\theta}_0$?

(b) Find a closed form expression for the maximum likelihood estimate of $\boldsymbol{\theta}_0$. (In this case, we are working from a single sample of a random vector.)

(c) What is the distribution of the MLE estimator $\hat{\boldsymbol{\Theta}}$? Is $\hat{\boldsymbol{\Theta}}$ unbiased?

(d) What is the MSE of the MLE, $\mathrm{E}[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0\|_2^2]$?

(e) Compute the Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta}_0)$ and verify that the MLE meets the Cramer-Rao lower bound.

(f) Defend the following statement: The MLE is the best unbiased estimator of $\boldsymbol{\theta}_0$.

6. A Cauchy random variable with "location parameter" ν has a density function

$$f_X(x;\nu) = \frac{1}{\pi(1+(x-\nu)^2)}, \quad x \in \mathbb{R}. \tag{1}$$

Despite its simple definition, this is a strange animal. First of all, its mean is not defined, as the integral $\int x/(1+x^2)\,dx$ is not absolutely convergent. It is also easy to see that the variance is infinite. But as you can see (especially if you sketch it), the density is symmetric around ν, and ν is certainly the median.

Let $X_1, X_2, \ldots, X_N$ be iid Cauchy random variables distributed as in (1). From observed data $X_1 = x_1, \ldots, X_N = x_N$, we will compare three estimators: the sample mean

$$\hat{\nu}_{mn} = \frac{1}{N}\sum_{n=1}^{N} x_n,$$

the sample median

$$\hat{\nu}_{md} = \begin{cases} x_{((N+1)/2)}, & \text{N odd,} \\ \frac{x_{(N/2)}+x_{(N/2+1)}}{2}, & \text{N even,} \end{cases}$$

where $x_{(i)}$ is the ith largest value in $\{x_1, \ldots, x_N\}$, and the MLE

$$\hat{\nu}_{mle} = \arg\max_{\nu} L(\nu; x_1, \ldots, x_N) = \arg\max_{\nu} \sum_{n=1}^{N} \ell(\nu; x_n)$$

where $\ell(\nu; x_n) = \log f_X(x_n; \nu)$.

(a) One particular draw of data for $N = 50$ is variable `x` in the file `hw06p6a.mat`. Plot the log likelihood function, and report the MLE for ν. Your MLE will of course be approximate, but make sure yours is accurate to within 10^{-2} to the true MLE. I will give you a hint here and tell you that the true value of ν is somewhere in the interval $[0, 5]$.

(b) The file `hw06p6b.mat` contains a matrix `X`. This is an $N \times Q$ matrix, where $N = 50$ and $Q = 1000$; each entry is an independent Cauchy random variable with $\nu_0 = 3$. Treating each column of `X` as a single draw of the data for $N = 50$, compute the sample mean, sample median, and MLE for each column. From these, report the empirical mean squared error (by averaging $(\hat{\nu} - \nu_0)^2$ over all Q trials) for each of the three estimators.

(c) Find an integral expression for the expected log likelihood function $e(\nu) = \mathrm{E}[\ell(\nu; X)]$ when X has Cauchy density $f_X(x; \nu_0)$ as in (1). Your expression should have the form

$$e(\nu) = \int_{-\infty}^{\infty} (\text{something that depends on } x, \nu, \nu_0) \; dx.$$

Compute $e(\nu)$ for $\nu_0 = 3$ for 250 equally spaced values of ν between 0 and 5. You can do this using numerical integration (the `integral` function in MATLAB or `scipy.integrate.quad` in Python). Make a plot of $e(\nu) = \mathrm{E}[\ell(\nu; X)]$.

(d) Plot, overlayed on the same axes, the (renormalized) log likelihood functions $\frac{1}{N}\sum_{n=1}^{N} \ell(\nu; x_n)$ as a function of $\nu \in [0, 5]$ for each of the first 10 columns of `X` from part (b). On top of this, plot $e(\nu) = \mathrm{E}[\ell(\nu; X)]$ from part (c) as a dotted line.