(Fall 2022) 515502 Probability **Early Bird: 2022/12/25, 9pm; Normal: 2022/12/26, 9pm**

Homework 4: Concentration Inequalities, LLN, and Multivariate Normal

Submission Guidelines: Your deliverables shall consist of 2 separate files – (i) A PDF file: Please compile all your write-ups and your report into one .pdf file (photos/scanned copies are acceptable; please make sure that the electronic files are of good quality and reader-friendly); (ii) A zip file: Please compress all your source code into one .zip file. Please submit your deliverables via E3.

Problem 1 (Techniques of Chernoff Bound) (10+15=25 points)

Let $X_1, \cdots, X_N$ be non-negative independent random variables with continuous distributions (but $X_1, \cdots, X_N$ are not necessarily identically distributed). Assume that the PDFs of X_i's are uniformly bounded by 1 (that is, for every X_i, we have $f_{X_i}(x) \leq 1$, for all $x \in \mathbb{R}$).

(a) Show that for every i, $E[\exp(-tX_i)] \leq \frac{1}{t}$, for all $t > 0$.

(b) By using (a), show that for any $\varepsilon > 0$, we have

$$P\Big(\sum_{i=1}^{N} X_i \leq \varepsilon N\Big) \leq (e\varepsilon)^N.$$

(Hint: For any $t > 0$, $P(\sum_{i=1}^{N} X_i \leq \varepsilon N) = P(e^{t\sum_{i=1}^{N} X_i} \leq e^{t\varepsilon N}) = P\big(e^{-t\sum_{i=1}^{N} X_i} \geq e^{-t\varepsilon N}\big))$

Problem 2 (Strong Law of Large Numbers) (10+5=15 points)

Consider two sequences of random variables $X_1, X_2, \cdots$ and $Y_1, Y_2, \cdots$ defined on the same sample space. Suppose that X_n converges to a and Y_n converges to b, almost surely.

(a) Show that $X_n \cdot Y_n$ converges to $a \cdot b$, almost surely. (Hint: Consider two events A, B defined as $A = \{\omega : X_n(\omega)$ does not converge to $a\}$ and $B = \{\omega : Y_n(\omega)$ does not converge to $b\}$)

(b) Based on the result in (a), would $X_n^2 \cdot Y_n^3$ also converge almost surely? If so, what value would $X_n^2 \cdot Y_n^3$ converge to?

Problem 3 (Convergence in Probability) (10+10=20 points)

A sequence of random variables $X_1, X_2, \cdots$ is said to converge to a number c **in the mean square**, if

$$\lim_{n\to\infty} E[(X_n - c)^2] = 0.$$

(a) Show that convergence in the mean square implies convergence in probability. (Hint: For every $\varepsilon > 0$, consider $P(|X_n - c| \geq \varepsilon)$ and use Markov's inequality)

(b) Please construct an example that shows that "convergence in probability" does NOT imply "convergence in the mean square." (Hint: You may consider a random variable X which is 0 with probability $1 - \frac{1}{n}$ and is a large value with probability $\frac{1}{n}$)

Problem 4 (Bivariate and Multivariate Normal for Regression) (20+30=50 points)

We have learned "bivariate normal" in Lectures 19-22. The idea of bivariate normal can be readily extended to "multivariate normal". One interesting application of multivariate normal (MVN) random variables is to solve regression tasks. In this problem, you will implement a simple MVN-based predictor that predicts the outputs of the testing queries based on the training data. Specifically, let $D_{train} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ be the training dataset and let $D_{test} = \{x_{N+1}, \cdots, x_{N+M}\}$ be the testing queries. The goal is to predict $\{y_{N+1}, \cdots, y_{N+M}\}$ that correspond to $\{x_{N+1}, \cdots, x_{N+M}\}$. The following Figure 1 shows an example of MVN-based regression.

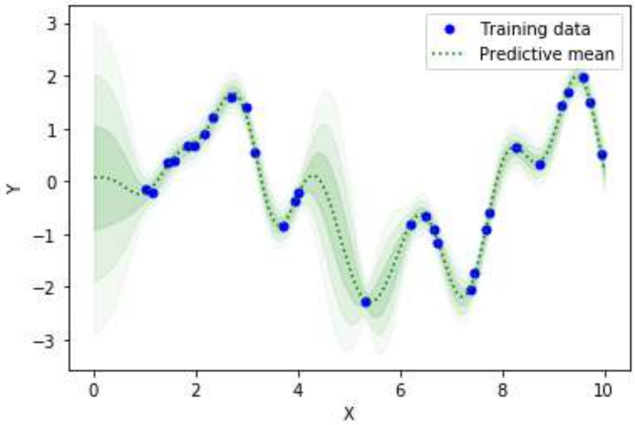

Figure 1: An example of MVN-based regression, where the shaded region shows the standard deviation of the predictive distribution.

(a) As a prep work, please show the following property of bivariate normal that we discussed in class: Let Z_1, Z_2 be a pair of bivariate normal random variable with mean μ_1, μ_2, variance σ_1^2, σ_2^2, and correlation coefficient ρ. Show that conditioned on that $Z_1 = z_1$, the conditional distribution of Z_2 is normal with mean $\mu_2 + \frac{\rho\sigma_2(z_1-\mu_1)}{\sigma_1}$ and variance $(1-\rho^2)\sigma_2^2$.

(b) The property in (a) can be extended to the multivariate normal case. Suppose that for every $k \in \{N+1, \cdots, N+M\}$, $\{Y_1, \cdots, Y_N, Y_k\}$ is multivariate normal with mean vector $\mu = [\mu_1, \cdots, \mu_N, \mu_k]^\intercal$ and a $(N+1) \times (N+1)$ covariance matrix Σ, where the covariance between Y_i and Y_j (denoted by $\Sigma_{i,j}$) has the following form:

$$\Sigma_{i,j} = \sigma_f^2 \exp\big(-\frac{(x_i - x_j)^2}{2\ell^2}\big) + \sigma^2\delta_{i,j}, \forall i, j,$$

where σ_f is a scale factor, ℓ is called the lengthscale, σ^2 is some positive constant (usually called the noise parameter), and $\delta_{i,j}$ is the delta function (i.e. $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$). Given that $Y_1 = y_1, \cdots, Y_N = y_N$, it can be shown that the conditional distribution of Y_k is **normal with mean** $K(x_k, x_{1:N})[K(x_{1:N}, x_{1:N}) + \sigma^2 I]^{-1} y_{1:N}$ **and variance** $K(x_k, x_k) - K(x_k, x_{1:N})[K(x_{1:N}, x_{1:N}) + \sigma^2 I]^{-1} K(x_{1:N}, x_k)$, where

- $K(x_k, x_k) = \Sigma_{k,k}$ is a scalar.
- $K(x_k, x_{1:N}) = [\Sigma_{k,1}, \cdots, \Sigma_{k,N}]$ is a $1 \times N$ vector.
- $K(x_{1:N}, x_{1:N})$ is an $N \times N$ matrix with the (i,j)-th entry equal to $\Sigma_{i,j}$.
- I is an identity matrix of size $N \times N$.
- $K(x_{1:N}, x_k)$ is the transpose of $K(x_k, x_{1:N})$.
- $y_{1:N} = [y_1, \cdots, y_N]^\intercal$ is an $N \times 1$ vector.

Based on the above conditional distribution, please write a program (e.g. in Python or MATLAB) to find the predictive distributions of the outputs of the test query points $\{x_{N+1}, \cdots, x_{N+M}\}$. What is the prediction result of the testing dataset under $\sigma_f = 1, \sigma = 0.1, \ell = 0.5$? What is the prediction result of the testing dataset if ℓ is set to be 0.05 instead? How about $\ell = 3.0$? Please briefly summarize your observation in a technical report (no more than 2 pages for this part).