Homework Assignment # 1

Due: Wednesday, September 14, 2016, 11:59 p.m.
Total marks: 100

Question 1. [10 MARKS]

Let $\Omega_X = \{a, b, c\}$ and $p_X(a) = 0.1, p_X(b) = 0.2$, and $p_X(c) = 0.7$. Let

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

(a) [3 MARKS] What is $E[f(x)]$?

(b) [3 MARKS] What is $E[1/p_X(x)]$?

(c) [4 MARKS] For an arbitrary pmf $p_X(x)$, what is $E[1/p_X(x)]$?

Question 2. [15 MARKS]

Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be independent multivariate Gaussian random variables, with $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{d \times d}$ for dimension $d \in \mathbb{N}$. Define $\mathbf{X} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_m\mathbf{X}_m$ as a convex combination, $a_i \geq 0$ and $\sum_{i=1}^m a_i = 1$.

(a) [5 MARKS] Write the expected value $E[\mathbf{X}]$ in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of $E[\mathbf{X}]$?

(b) [10 MARKS] Write the covariance $\text{Cov}[\mathbf{X}]$ in terms of the givens $a_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$. Show all you steps. What is the dimension of $\text{Cov}[\mathbf{X}]$? Briefly explain how the result for $\text{Cov}[\mathbf{X}]$ would be different if the variables X_1 and X_2 are not independent and have covariance $\text{Cov}[\mathbf{X}_1, \mathbf{X}_2] = \boldsymbol{\Lambda}$ for $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$.

Question 3. [10 MARKS]

This question involves some simple simulations, to better visualize random variables and get some intuition for sampling, which is a central theme in machine learning. Use the attached code called `simulate.py`. This code is a simple script for sampling and plotting with python; play with some of the parameters to see what it is doing. Calling `simulate.py` runs with default parameters; `simulate.py 1 100` simulates 100 samples from a 1d Gaussian.

(a) [5 MARKS] Run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 1.0$. Next run the code for 10, 100 and 1000 samples with dim=1 and $\sigma = 10.0$. What do you notice about the sample mean?

(b) [5 MARKS] The current covariance for dim=3 is

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

What does that mean about the multivariate Gaussian (i.e., about X, Y and Z)?

Change the covariance to

$$\Sigma = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

What happens?

Question 4. [30 MARKS]

Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$. That is, the prior density is

$$f(\lambda) = \theta e^{-\theta\lambda}$$

where $\lambda \in (0, \infty)$.

(a) [5 MARKS] Before observing any data (any reported accidents), what is the most likely value for λ?

(b) [5 MARKS] Now imagine there are 79 accidents over 9 days. Determine the maximum likelihood estimate of λ.

(c) [5 MARKS] Again imagine there are 79 accidents over 9 days. Determine the maximum a posteriori (MAP) estimate of λ.

(d) [5 MARKS] Imagine you now want to predict the number of accidents for tomorrow. How can you use the maximum likelihood estimate computed above? What about the MAP estimate? What would they predict?

(e) [5 MARKS] For the MAP estimate, what is the purpose of the prior once we observe this data?

(f) [5 MARKS] Look at the plots of some exponential distributions to better understand the prior chosen on λ. Imagine that now new safety measures have been put in place and you believe that the number of accidents per day should sharply decrease. How might you change θ to better reflect this new belief about the number of accidents?

Question 5. [20 MARKS]

Imagine that you would like to predict if your favorite table will be free at your favorite restaurant. The only additional piece of information you can collect, however, is if it is sunny or not sunny. You collect paired samples from visit of the form (is sunny, is table free), where it is either sunny (1) or not sunny (0) and the table is either free (1) or not free(0).

(a) [10 MARKS] How can this be formulated as a maximum likelihood problem?

(b) [5 MARKS] Assume you have collected data for the last 10 days and computed the maximum likelihood solution to the problem formulated in (a). If it is sunny today, how would you predict if your table will be free?

(c) [5 MARKS] Imagine now that you could further gather information about if it is morning, afternoon, or evening. How does this change the maximum likelihood problem?

Question 6. [15 MARKS]

Suppose you have three coins. Coin A has a probability of heads of 0.75, Coin B has a probability of heads of 0.5, and Coin C has a probability of heads of 0.25.

(a) [5 MARKS] Suppose you flip all three coins at once, and let X be the number of heads you see (which will be between 0 and 3). What is the expected value of X, $E[X]$?

(b) [10 MARKS] Suppose instead you put all three coins in your pocket, select one at random, and then flip that coin 5 times. You notice that 3 of the 5 flips result in heads while the other 2 are tails. What is the probability that you chose Coin C?

Homework policies:

Your assignment will be submitted as a single pdf document and a zip file with code, on canvas. The questions must be typed; for example, in Latex, Microsoft Word, Lyx, etc. or must be written legibly and scanned. Images may be scanned and inserted into the document if it is too complicated to draw them properly. All code (if applicable) should be turned in when you submit your assignment. Use Matlab, Python, R, Java or C.

Policy for late submission assignments: Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rule:

on time: your score 1

1 day late: your score 0.9

2 days late: your score 0.7

3 days late: your score 0.5

4 days late: your score 0.3

5 days late: your score 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Good luck!