COS302/SML305- Princeton University-Spring 2022
Assignment # 9

Due: April 11, 2022 at 11:59 pm

Upload at: https://www.gradescope.com/courses/355863/assignments

---

Remember to append your Colab PDF as explained in the first homework, with all outputs visible.
When you print to PDF it may be helpful to scale at 95% or so to get everything on the page.

Problem 1 (24 pts)

Let X and Y be random variables with a joint distribution P. Determine if the following statements are *true* or *false*. Justify your answers with examples or proofs.

(A) $P(X = x|Y) \leq P(X = x)$.

(B) $P(X = x|Y)$ is a random variable.

(C) $\mathbb{E}[X|Y = y]$ is a random variable.

(D) If X, Y are independent, then $\mathbb{E}[XY] = 0$.

(E) $\mathbb{E}[XY] \leq \mathbb{E}[X]\mathbb{E}[Y]$.

(F) For any function f,

$$\mathbb{E}[(Y - f(X))^2] \geq \mathrm{Var}(Y).$$

Problem 2 (14pts)

Let X and Y be random variables with a joint probability density function $p(x, y)$. Show that

$$\mathbb{E}_X[x] = \mathbb{E}_Y[\mathbb{E}_X[x \mid Y = y]],$$

where the notation $\mathbb{E}_X[x \mid Y = y]$ denotes the expectation of X under the conditional distribution $P(X \mid Y = y)$.

Problem 3 (30pts)

In this problem, you will use the entire 50k-digit MNIST data set. To remind you, the data are 28×28 greyscale images of the digits 0 through 9. Download the `mnist_full.pkl.gz` file and load the file into a `Colab` notebook using code such as the following.

```
import pickle as pkl
import numpy as np
import gzip
with gzip.open('mnist_full.pkl.gz', 'rb') as fh:
  mnist = pkl.load(fh)
```

This will result in a dictionary `mnist` with keys and values that should be self-explanatory. Use this data to solve the problems below. *Note:* It is possible that when solving the problems below, you will encounter issues when the covariance matrix is not positive definite. To solve this issue, you can add a small value to the diagonal of the covariance matrix; i.e., set $\tilde{\Sigma} = \Sigma + \alpha I$, for $\alpha \approx 10^{-6}$.

(A) Compute the empirical mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the training images.

(B) Reshape and display the mean as an image using `imshow`.

(C) Generate 5 samples from the multivariate Gaussian with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from (A). Do this only using `numpy.random.randn` and linear algebra operations. Reshape and display these samples using `imshow`.

(D) Now iterate over each of the possible labels from 0 to 9. Compute the mean and covariance of the training data that have that label. Here's a sketch of some code for obtaining the correct images.

```
for label in range(10):
  indices = train_labels == label
  images = train_images[indices,:]
```

Display the mean of each as an image, and generate 5 samples from the Gaussian distribution with the label-specific mean and covariance. Make a plot of these samples.

Problem 4 (30pts)

Consider the following joint distribution over random variables X and Y.

Y		x_1	x_2	x_3	x_4	x_5
	y_1	0.01	0.02	0.03	0.1	0.1
	y_2	0.05	0.1	0.05	0.07	0.2
	y_3	0.1	0.05	0.03	0.05	0.04

X

Compute the following quantities in a `Colab` notebook. Remember to compute these quantitites in **bits** which means using base 2 for logarithms. To get started, here the probability mass function (PMF) in Python.

```
PXY = np.array([[0.01, 0.02, 0.03, 0.1, 0.1],
                [0.05, 0.1, 0.05, 0.07, 0.2],
                [0.1, 0.05, 0.03, 0.05, 0.04]])
```

(A) What is the entropy $H(X)$?

(B) What is the entropy $H(Y)$?

(C) What is the conditional entropy $H(X \mid Y)$?

(D) What is the conditional entropy $H(Y \mid X)$?

(E) What is the joint entropy $H(X, Y)$?

(F) What is the mutual information $I(X\,;\,Y)$? Compute the mutual information once using the PMF and then once using relationships between the quantities you used in (A)-(E) to verify your answer.

Problem 5 (2pts)
Approximately how many hours did this assignment take you to complete?

My notebook URL: `https://colab.research.google.com/XXXXXXXXXXXXXXXXXXXXXXX`