Math Review Homework.
Probability Problem Set 1.

1. σ-algebras

   (a) The Fisher Iris Date Set is a famous data set in Pattern Recognition and Machine Learning. It can be found in many places, including:

   `https://archive.ics.uci.edu/ml/datasets/iris`

   The data set is a collection of 4 measurements on 3 flower species. The flower species are *Iris setosa, Iris virginica* and *Iris versicolor*. The measurements are Petal Length, Petal Width, Sepal Length, and Sepal Width. You may assume that all the measurements are in units of meters and lie in the open interval $(0, 1)$. Let $\mathcal{X}$ denote the set of all possible measurement vectors, $\mathbf{x} \in (0, 1)^4$. Describe the components of an appropriate probability space $(\mathcal{X}, \mathcal{S}, P)$.

2. Describe an experiment that treats the mean of a random variable as another random variable and uses multiple samples to approximate the distribution of the mean.

3. Self-Driving Cars

   (a) Suppose a self-driving car is programmed to receive information via a cellular connection from one source about estimated Travel times and from another source about road-construction Conditions; let's refer to these sources as E_{tr} and E_C. Assume that sometimes E_C are unavailable. The program has default settings to resolve ties or numbers that are close to ties. For example, consider the case of only using only source E_{tr} to make a choice about which fork in the road to take if the Estimated Travel Times, $\mathcal{E}_T$ are similar. The fastest route is desired, but because of uncertainty in $\mathcal{E}_T$, the fastest route may be unknown. A simple rule is to always take the rightmost fork. For example, let s denote a value between 0 and 1 that represents a percentage. If the program must choose between a left or right fork in the road and

   $$\mathcal{E}_T = \{T_L, T_R\} \text{ and } |T_L - T_R| < s\frac{(T_L + T_R)}{2},$$

   then take the right fork. If there are 3 forks with

   $$\mathcal{E}_T = \{T_L, T_M, T_R\}$$

   and

   $$\max\{|T_L - T_M|, |T_L - T_R|, |T_M - T_R|\} < s\frac{(T_L + T_M + T_R)}{3},$$

   then take the right fork, etc.

   A programmer must make a decision about what logic to use if additional information is received from source E_c. Consider the case of 3 forks in the road. If the program has already chosen to take the right fork using the rule above, should the programmer include logic that changes from the right fork to the middle fork if

E_c indicates that the traffic on the left fork is moving very slowly because of road construction? Why or why not?

(b) Assume the programmer has access to a very large database that stores pairs of the form $(TrueTrTime, EstTrTime)$ where

$$\begin{aligned} TrueTrTime &= \text{ Time recorded by a driver after traversing the route.} \\ EstTrTime &= \text{ Estimated time from source } E_{tr} \text{ before traversing the route.} \end{aligned}$$

How could you use this database to set the value of s?

4. Compute the moments: μ_1, μ_2, γ_1 of the multinomial distribution $\{(1, 0.1), (2, 0.3), (3, 0.6)\}$.

5. Assume $f(x)$ is the probability density function for a univariate Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Calculate $\mathbf{E}[x]$.

6. Assume that the diagonalization of a 2×2 covariance matrix can be written as $\mathbf{\Sigma} = \mathbf{V}^t \Lambda \mathbf{V}$ where

$$\Lambda = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$

Describe the level curves of $f(\mathbf{x}) = (\mathbf{x} - \mu_\mathbf{x})^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_\mathbf{x})$ in terms of the eigenvectors and eigenvalues of $\mathbf{\Sigma}$

7. If $\mathbf{x}$ is a random vector with a multi-variate Gaussian distribution having mean vector $\mu_\mathbf{x}$ and covariance matrix $\mathbf{\Sigma}$, then derive the mean vector $\mu_\mathbf{y}$ and covariance matrix $\mathbf{\Sigma_y}$ of the Principal Component Transform of $\mathbf{x}$, $\mathbf{y} = \mathbf{V}^t (\mathbf{x} - \mu_\mathbf{x})$.

8. Sum of Uniform and Gaussian Random Variable (Not the sum of the pdfs!)

(a) Suppose U and V are uniform random variables on $\mathbb{R}$ with pdfs given by $f_U(u) = \frac{1}{2}$ for $u \in [0, 2]$ and $f_V(v) = \frac{1}{4}$ for $v \in [-1, 3]$. What is the pdf of the random variable $U + V$?

(b) Suppose X and Y are univariate Gaussian random variables, X is distributed according to $\mathcal{G}_X(\mu_x, \sigma_x)$ ($X \sim \mathcal{G}_X$) and Y is distributed according to $\mathcal{G}_Y(\mu_y, \sigma_y)$ ($Y \sim \mathcal{G}_Y$) where $\mu_x = 0$, $\sigma_x = 1$, $\mu_y = 1$, and $\sigma_y = 2$. What is the pdf of $X + Y$?

9. Prove that if $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_M(\mathbf{x})$ are pdfs and $c_1, c_2, \ldots, c_M \in [0, 1]$ with $\sum_{m=1}^M c_m = 1$, then the mixture $f(\mathbf{x}) = \sum_{m=1}^M c_m f_m(\mathbf{x})$ is also a pdf.