Final examination

DSE 210

Name: Peyman Hesami

Be clear and concise. Write your answers in the space provided. Use the backs of pages for scratchwork.

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	

TOTAL POINTS: 80

1. **(10 points)** You are dealt two cards at random from a standard deck. What is the probability that:

   (a) The first card is an ace?

   There are 4 aces in each standard deck

   $\Rightarrow Pr(\text{first card} = \text{ace}) = \frac{4}{52}$

   (b) The first and second cards are both aces?

   $$Pr(\text{first \& second cards} = \text{ace}) = Pr(\text{first card} = \text{ace})\underbrace{Pr(\text{second card ace} \mid \text{first card ace})}_{=\frac{3}{51}}$$
   $$= \frac{4}{52} \times \frac{3}{51}$$

   (c) The second card is an ace? if we have no knowledge of first card then $\Rightarrow$

   $Pr(\text{second card ace}) = \frac{4}{52}$

   Solution 2: A: first card ace B: second card ace

   $$Pr(B) = Pr(B|A)Pr(A) + Pr(B|A')Pr(A') = \frac{3}{51} \times \frac{4}{52} + \frac{48}{52} \times \frac{4}{51} = \frac{4}{52}$$

   (d) The first card is an ace, given that it is a heart?

   $Pr(\text{first card} = \text{ace} \mid \text{first card} = \text{heart}) = \frac{1}{13}$

   (e) The second card is an ace, given that the first card is an ace?

   $Pr(\text{second card} = \text{ace} \mid \text{first card} = \text{ace}) = \frac{3}{51}$

2. **(3 points)** Ten cards are chosen at random from a standard deck. Which of the following pairs of events A, B are independent? Circle them.

   - A: first card is a ten, B: tenth card is a nine
   - A: first card is a ten, B: second card is a heart
   - A: second card is a heart, B: fifth card is a club

3. **(10 points)** Short answer questions.

(a) The letters G, H, I, R, T are randomly permuted. What is the probability that the result is the word R, I, G, H, T?

number of permutations $5 \times 4 \times 3 \times 2 \times 1 = 5!$

$\Rightarrow Pr(\text{result} = RIGHT) = \frac{1}{5!} = \frac{1}{120}$

(b) Three fair dice are rolled. What is the probability that they all have the same value?

we roll the first dice $\Rightarrow Pr(\text{second dice} = \text{first dice}) = \frac{1}{6}$

& $Pr(\text{third dice} = \text{second dice}) = \frac{1}{6}$

$\Rightarrow Pr(\text{all dices have same value}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

(c) Each time you go to the gym, you have a 20% chance of running into your worst enemy. What is the expected number of trips to the gym before you meet this person?

$p = Pr(\text{meeting worst enemy}) = \frac{1}{5}$

$\Rightarrow E\{\#\text{trips before meeting worst enemy}\} = \frac{1}{p} = 5$

(d) A certain population consists of 40% men and 60% women. Of the men, 20% are left-handed, and of the women, 10% are left-handed. A person is picked at random from this population and is found to be left-handed. What is the probability that this person is female?

$Pr(\text{femal}) = 0.6$, $Pr(\text{male}) = 0.4$, $Pr(\text{left}|\text{man}) = 0.2$ $Pr(\text{left}|\text{woman}) = 0.1$

$Pr(\text{left}) = 0.2 \times 0.4 + 0.1 \times 0.6 = 0.14$

$Pr(\text{female}|\text{left}) = \frac{Pr(\text{female})}{Pr(\text{left})} Pr(\text{left}|\text{female}) = \frac{0.6}{0.14} 0.1 = \frac{3}{7}$

(e) A man has a bottle containing ten identical-looking pills. Two of them contain medicine while the other 8 are placebos. Upon taking a pill, the man feels either **good** or **not good**, with the following probabilities:

$$\Pr(\text{feel } \textbf{good} \mid \text{medicine}) = \frac{3}{4}$$

$$\Pr(\text{feel } \textbf{good} \mid \text{placebo}) = \frac{1}{2}$$

Today, the man picks a pill at random and finds that he feels good. What is the probability that the pill contained medicine?

$Pr(\text{medicine}) = 0.2$ $\quad Pr(\text{placebo}) = 0.8$

$Pr(\text{feel good}) = \frac{3}{4} \times 0.2 + 1/2 \times 0.8 = 0.55$

$Pr(\text{medicine}|\text{feel good}) = \frac{Pr(\text{medicine})}{Pr(\text{feel good})} Pr(\text{feel good}|\text{medicine})$

$= \frac{0.2}{0.55} \times \frac{3}{4} = 0.2727$

4. **(8 points)** A die has six sides that come up with different probabilities.

$$\Pr(1) = \Pr(2) = \Pr(3) = \frac{1}{12}, \quad \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{4}.$$

(a) You roll the die; let X denote the outcome. What is $\mathbb{E}(X)$?

$E\{X\} = 1/12(1+2+3) + \frac{1}{4}(4+5+6) = 4.25$

(b) What is $\text{var}(X)$?

$E\{X^2\} = 1/12(1+2^2+3^2) + \frac{1}{4}(4^2+5^2+6^2) = 20.416$

$\text{var}(X) = E\{X^2\} - E\{X\}^2 = 20.416 - 18.062 = 2.3535$

(c) Now you roll this die a hundred times, and let Z be the sum of all the rolls. What is $\mathbb{E}(Z)$?

$Z = X_1 + \cdots + X_{100}$

$E\{Z\} = E\{X_1\} + \cdots + E\{X_{100}\} = 100 \times 4.25 = 425$

(d) What is $\text{var}(Z)$? X_i's are independent

$\text{Var}(Z) = 100 \times 2.3535 = 235.35$

5. **(3 points)** A pair of random variables X_1 and X_2 have the following properties:

- They both take values in $\{-1, 1\}$
- X_1 has mean 0 while X_2 has mean 0.5
- The correlation between X_1 and X_2 is 0.25

$\mu = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & \frac{\sqrt{3}}{8} \\ \frac{\sqrt{3}}{8} & \frac{3}{4} \end{pmatrix}$

Suppose we fit a (bivariate) Gaussian to (X_1, X_2). Give the mean and covariance matrix of this Gaussian.

$\Pr(X_1 = 1) = P_1$
$\Pr(X_1 = -1) = P_2$ $\Rightarrow E\{X_1\} = P_1 - P_2 = 0$, $P_1 + P_2 = 1$ $\Rightarrow P_1 = P_2 = 1/2 \Rightarrow \text{Var}(X_1) = 1/2 \times 1 + 1/2 \times 1 = 1$

sum of prob should be 1

$\Pr(X_2 = 1) = q_1$
$\Pr(X_2 = -1) = q_2$ $\Rightarrow E\{X_2\} = q_1 - q_2 = 1/2$, $q_1 + q_2 = 1$ $\Rightarrow q_1 = 3/4,\ q_2 = 1/4 \Rightarrow \text{Var}(X_2) = \frac{3}{4}\left(\frac{1}{2}\right)^2 + \frac{1}{4}\left(\frac{3}{2}\right)^2 = \frac{3}{4}$

$\frac{1}{4} = \text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\text{Cov}(X_1, X_2)}{1 \times \frac{\sqrt{3}}{2}} \Rightarrow \text{Cov}(X_1, X_2) = \frac{\sqrt{3}}{8}$

$u_3 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix} \Rightarrow Cov(X)u_3 = \lambda_3 u_3 \Rightarrow \begin{pmatrix}5 & -3 & 0\\-3 & 5 & 0\\0 & 0 & 4\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\\0\end{pmatrix} = \begin{pmatrix}\frac{8}{\sqrt{2}}\\-\frac{8}{\sqrt{2}}\\0\end{pmatrix} = \begin{pmatrix}\frac{\lambda_3}{\sqrt{2}}\\-\frac{\lambda_3}{\sqrt{2}}\\0\end{pmatrix} \Rightarrow \lambda_3 = 8$

6. **(10 points)** A certain random variable $X \in \mathbb{R}^3$ has mean and covariance as follows:

$$\mathbb{E}X = \begin{pmatrix}1\\2\\3\end{pmatrix}, \quad \text{cov}(X) = \begin{pmatrix}5 & -3 & 0\\-3 & 5 & 0\\0 & 0 & 4\end{pmatrix}$$

→ Cov matrix is positive semi-definite ⇒ all of its eigenvectors are orthonormal

(a) The eigenvectors of cov(X) can be found in the following list:

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\1\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

Circle them. only these three satisfy $Cov(X)u = \lambda u$

continuing

(b) Find the eigenvalues corresponding to each of the eigenvectors in part (a). Make it clear which eigenvalue belongs to which eigenvector.

$u_1 = \begin{pmatrix}0\\0\\1\end{pmatrix} \Rightarrow Cov(X)u_1 = \lambda_1 u_1 \Rightarrow \begin{pmatrix}5 & -3 & 0\\-3 & 5 & 0\\0 & 0 & 4\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}0\\0\\4\end{pmatrix} = \lambda_1\begin{pmatrix}0\\0\\1\end{pmatrix} \Rightarrow \lambda_1 = 4$

$u_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix} \Rightarrow Cov(X)u_2 = \lambda_2 u_2 \Rightarrow \begin{pmatrix}+5 & -3 & 0\\-3 & 5 & 0\\0 & 0 & 4\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\\0\end{pmatrix} = \begin{pmatrix}\frac{2}{\sqrt{2}}\\\frac{2}{\sqrt{2}}\\0\end{pmatrix} = \frac{\lambda_2}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix} \Rightarrow \lambda_2 = 2$

(c) Suppose we used principal component analysis (PCA) to project points X into *two* dimensions. Which directions would it project onto?

we choose the direction with largest eigen values (largest variance) ⇒ λ_1 & λ_3

⇒ two dimentions $= (u_3 \ u_1) = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\\-\frac{1}{\sqrt{2}} & 0\\0 & 1\end{pmatrix} = V$

(d) Continuing from part (c), what would be the resulting two-dimensional projection of the point $x = (4, 0, 2)$?

$V = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\\-\frac{1}{\sqrt{2}} & 0\\0 & 1\end{pmatrix}$: transformation (dimensionality reduction) matrix

$\text{projection}(X) = X.V = V^T X = \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0\\0 & 0 & 1\end{pmatrix}\begin{pmatrix}4\\0\\2\end{pmatrix} = \begin{pmatrix}\frac{4}{\sqrt{2}}\\2\end{pmatrix} = \begin{pmatrix}2\sqrt{2}\\2\end{pmatrix}$

(e) Continuing from part (d), suppose that starting from the 2-d projection, we tried to reconstruct the original x. What would the three-dimensional reconstruction be, exactly?

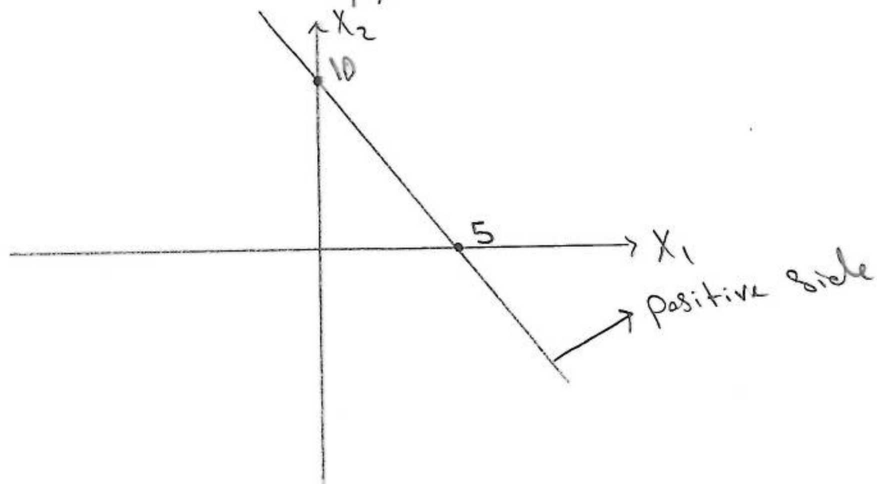
$\text{reconstructed}(X) = VV^T X = \begin{pmatrix}\frac{1}{\sqrt{2}} & 0\\-\frac{1}{\sqrt{2}} & 0\\0 & 1\end{pmatrix}\begin{pmatrix}2\sqrt{2}\\2\end{pmatrix} = \begin{pmatrix}2\\-2\\2\end{pmatrix}$

7. **(4 points)** Consider the linear classifier $w \cdot x \geq \theta$, where

$$w = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \theta = 10.$$

Sketch the decision boundary in $\mathbb{R}^2$. Make sure to indicate where the boundary intersects the two axes, and which side of the boundary is the positive side.

$w \cdot x = x^T \cdot w = (x_1, x_2)\begin{pmatrix} 2 \\ 1 \end{pmatrix} \geq 10 \Rightarrow 2x_1 + x_2 \geq 10$



8. **(4 points)** A survey is taken to determine what fraction of freshman computer science majors have prior programming experience. Call this unknown fraction p. Out of the nationwide pool of computer science freshmen, 100 are chosen at random. Of them, 40% had prior programming experience.

   (a) The natural estimate of p is 0.4. Give a 95% confidence interval for the estimate.

$\hat{p} = 0.4 \qquad X = \frac{X_1 + \cdots + X_{100}}{100} \Rightarrow X \sim N\left(p, \frac{p(1-p)}{100}\right)$

$p = \hat{p} = 0.4 \Rightarrow \sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{100}} = \sqrt{\frac{0.4 \times 0.6}{100}} = 0.0489$

95% confidence interval $\equiv$ $0.4 \pm 2(0.0489) = 0.4 \pm 0.09797$

   (b) Suppose we now want to estimate p more accurately, to within a 95% confidence interval of ± 0.01. What sample size should we use?

95% confidence interval of $\pm 0.01 \equiv 2\sigma = 0.01$

$\Rightarrow 2\sqrt{\frac{p(1-p)}{n}} = 0.01 \Rightarrow 2\sqrt{\frac{0.4 \times 0.6}{n}} = 0.01$

$\Rightarrow 200 \times 0.4898 = \sqrt{n}$

$\Rightarrow n \approx 9600$

9. **(2 points)** A school wants to determine the average number of hours that the students spend on homework; call this unknown number μ. 100 students are chosen at random, and each of them is asked to report the typical number of hours per week that he or she spends on homework. The reported numbers have a mean of 12.2 and a standard deviation of 5.4. Give a 95% confidence interval for μ.

$$M = \frac{X_1 + \cdots + X_{100}}{100}$$

$$E\{X_1, \ldots, X_{100}\} = 12.2 \qquad Std(X_1, \ldots, X_{100}) = 5.4 \Rightarrow \begin{cases} E\{M\} = \frac{100 \times 12.2}{100} = 12.2 \\ Var\{M\} = \frac{1}{100^2} \times 100 \times 5.4^2 = \frac{5.4^2}{100} \end{cases} \Rightarrow 12.2 \pm 2\frac{5.4}{10} = 12.2 \pm 1.08$$

10. **(6 points)** Genius Academy is a high school that claims to prepare its students exceptionally well for the SAT exam. A random sample is taken of 100 Genius Academy students, and their SAT scores turn out to have a mean of 1930 and a standard deviation of 150. A random sample is also taken of 100 students from the other local high school, and their scores have a lower mean, of 1860, with a standard deviation of 200.

We wish to determine whether the difference between these observed averages is significant.

(a) State the null hypothesis.

The mean of two distributions (SAT scores in genius academy and SAT scores in local highscools) are the same

(b) Compute a suitable z-statistic for this situation. assume null is true $\Rightarrow$

genius academy scores : $X_1 \sim N(\mu_1, \sigma_1^2)$

local highschool scores : $X_2 \sim N(\mu_2, \sigma_2^2)$

$$\begin{cases} \mu_1 = 1930 \\ \sigma_1 = \frac{150}{\sqrt{100}} = 15 \end{cases} \qquad \begin{cases} \mu_2 = 1860 \\ \sigma_2 = \frac{200}{\sqrt{100}} = 20 \end{cases}$$

$$\Rightarrow X_1 - X_2 \sim N(0, \sigma_1^2 + \sigma_2^2) = N(0, 625)$$

$$\Rightarrow z = \frac{1930 - 1860 - 0}{\sqrt{625}} = 2.8$$

(c) What is the p value, and what conclusion would you draw?

p-value = Gaussian($X = 1930 - 1860$, $\mu = 0$, $\sigma^2 = 625$) = 0.0026

the observed difference has probability about 0.26% under the null hypothesis and it's strong evidence against the null hypothesis

11. **(20 points)** For this last problem, you should turn in an iPython notebook.

Download the IRIS data set from:

`https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`

This is a data set of 150 points in $\mathbb{R}^4$, with three classes; refer to the website for more details of the features and classes.

(a) Use a PCA projection to 2-d to visualize the entire data set. You should plot different classes using different colors/shapes. Do the classes seem well-separated from each other?

(b) Now build a classifier for this data set, based on a generative model (you can choose whichever you like).

- Split the data set into training/test data as follows: use the first 35 points in each class for training, and use the remaining 15 points for testing.
- What error rate do you achieve?