Final examination

DSE 210

Name: Kevin Kannappan

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?

   $$P(X_1 = \text{Ace}) = \boxed{1/13}$$

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

   $$P(X_1 = \text{Ace} \cap X_2 = \text{Ace}) = \boxed{(1/13) \cdot (3/51)}$$

   (c) The second card is an ace?

   $$P(X_2 = \text{Ace}) = \boxed{(12/13) \cdot (1/13) + (1/13) \cdot (3/51)}$$

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

   $$P(\text{Ace} \mid \text{Heart}) = \boxed{1/13}, \text{ all hearts equally likely}$$

   ~~or independent~~

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

   $$Pr(\text{Ace} \mid 1st \text{ Ace}) = \boxed{3/51}$$

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?

$$\frac{1}{5!} = \boxed{\frac{1}{120}}$$

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

$$\frac{6}{6^3} = \boxed{\frac{1}{36}}$$

(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?

coin with bias $p = 0.2$, $\quad 0.2(n) = 1$

$n = 5$

$\boxed{5 \text{ trips}}$

(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?

$Pr(m) = 0.4$

$Pr(f) = 0.6$

$Pr(L|m) = 0.2$

$Pr(L|f) = 0.1$

$$Pr(f|L) = \frac{Pr(L|f) \cdot Pr(f)}{Pr(L)} = \frac{0.06}{0.14} \boxed{\approx 0.43}$$

$$Pr(L) = Pr(L|m) \cdot Pr(m) + Pr(L|f) \cdot Pr(f)$$
$$(0.2) \cdot (0.4) + (0.1) \cdot (0.6) = 0.14$$

(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:

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

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

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

$Pr(\text{medicine}) = 0.2$

$Pr(\text{placebo}) = 0.8$

$$Pr(\text{medicine} \mid \text{feel good}) = \frac{pr(\text{feel good} \mid \text{medicine}) \cdot Pr(\text{medicine})}{pr(\text{feel good})}$$

$$pr(\text{feel good}) = Pr(\text{feel good} \mid \text{medicine}) \cdot Pr(\text{medicine}) + Pr(\text{feel good} \mid \text{placebo}) \cdot Pr(\text{placebo})$$

$$= (3/4)(1/5) + (1/2)(4/5)$$

$$= 0.55$$

$$\frac{0.15}{0.55} = \boxed{0.27}$$

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 $\mathbb{E}(X)$?

$$E(x) = 1(1/12) + 2(1/12) + 3(1/12) + 4(1/4) + 5(1/4) + 6(1/4)$$
$$= 51/12 = \boxed{4.25}$$

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

$$Var(x) = E(x^2) - (E(x))^2$$
$$= 1^2(1/12) + 2^2(1/12) + 3^2(1/12) + 4^2(1/4) + 5^2(1/4) + 6^2(1/4)$$
$$= 245/12 - (51/12)^2 \boxed{\approx 2.35}$$

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

By expectation rules (and each roll is independent):

$$100 \cdot (4.25) = \boxed{425 = E(Z)}$$

(d) What is $\mathrm{var}(Z)$?

By variance rules (and each roll is independent):

$$100 \cdot (2.35) = \boxed{235 = Var(Z)}$$

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

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

$$E(x_1) = 1(p) + -1 \cdot (1-p) = 0 \Rightarrow 2p = 1; \ p_1 = 0.5$$
$$E(x_2) = 1(p) + -1 \cdot (1-p) = 0.5 \Rightarrow 2p = 1.5; \ p_2 = 0.75$$
$$E(x_1^2) = 1^2(0.5) + -1^2(0.5) = 1; \ Var(x_1) = 1 - 0 = 1; \text{ hence } \sigma_1 = 1$$
$$E(x_2^2) = 1^2(0.75) + -1^2(0.25) = 1; \ Var(x_2) = 1 - (0.5)^2 = 0.75; \text{ hence } \sigma_2 = \sqrt{0.75}$$

$$corr(x_1, x_2) = \frac{cov(x_1, x_2)}{std(x_1)\, std(x_2)} \Rightarrow 0.22 = \frac{cov(x_1, x_2)}{1 \cdot \sqrt{75}} \Rightarrow \begin{matrix} cov(x_1,x_2) = \\ 0.22 \end{matrix}$$

$$\boxed{\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}, \quad cov(x_1, x_2) = \begin{bmatrix} 1 & 0.22 \\ 0.22 & 0.75 \end{bmatrix}}$$

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

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

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

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \underset{\text{3rd}}{\boxed{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}}, \underset{\text{2nd}}{\boxed{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}}}, \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \underset{\text{1st}}{\boxed{\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}}}$$

Circle them.

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

$$\begin{bmatrix} \overset{\text{1st}}{8}, & \overset{\text{2nd}}{2}, & \overset{\text{3rd}}{4} \end{bmatrix}$$

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

Directions with largest variance:

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

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

$$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4/\sqrt{2} \\ 2 \end{bmatrix}$$

(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?

$$\begin{bmatrix} 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4\sqrt{2} \\ 2 \end{bmatrix} = \begin{bmatrix} 4/2 \\ -4/2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$$

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.

$2x + y \geq 10$

$y \geq -2x + 10$

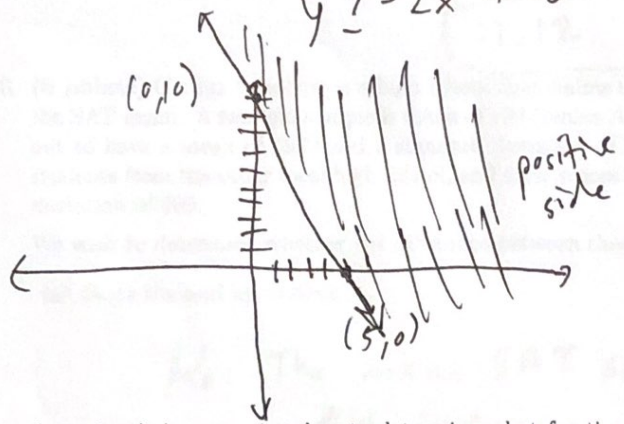
solve intercepts

$y = 0$; $0 \geq -2x + 10$

$-10 \geq -2x$

$x = 5$

$x = 0$; $y \geq 10$



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.

$\sigma = \sqrt{\frac{0.4 \cdot 0.6}{100}} \approx 0.049$

$0.4 - 2(0.049)$ to $0.4 + 2(0.049)$

$\boxed{-0.06 \text{ to } 0.14}$

(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?

P at its largest variance = 0.5, use that to generate a confidence interval at least 95% estimated within 0.01

$2 \cdot \sqrt{\frac{(0.5)^2}{n}} = 0.01$

$\sqrt{\frac{0.25}{n}} = \frac{0.01}{2}$;

if we take $p = 0.4$;

then $\sqrt{\frac{0.24}{n}} = \frac{.01}{2}$

$\boxed{\begin{array}{l} n = 9{,}600 \text{ for } p = 0.4 \\ n = 10{,}000 \text{ for } p = 0.5 \end{array}}$

~~n = 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.

use sample μ and σ:

$$\sigma = \frac{5.4}{\sqrt{100}} = \sigma = 0.54$$

$$12.2 - 2(0.54) \text{ to } 12.2 + 2(0.54)$$

$$\boxed{11.12 \text{ to } 13.28}$$

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.

H_0: The mean SAT scores of the shcols are the sare

H_1: The mean SAT scores of the schools are diffrenent

(b) Compute a suitable z-statistic for this situation.

$$X_G = 1930 \quad , X_H = 1860$$

$$\sigma_G = \frac{150}{\sqrt{100}} = 15 \; ; \; \sigma_H = \frac{200}{\sqrt{100}} = 20 \; ; \; \sigma = \sqrt{15^2 + 20^2}$$

$$\sigma = 25$$

$$Z = \frac{1930 - 1860}{25} \approx \boxed{2.8}$$

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

Using p-vabe calculator, the $\boxed{\text{p-value is } .00256}$

I would reject the null hypothesis that the mean SAT scores of the schools are the sare and conclude that the differences between these observed averages is significant.

Final Exam NB

March 18, 2019

1 Final Exam Notebook

Problem 6 using Python packages:

```
In [42]: import pandas as pd
         import matplotlib.pyplot as plt
         %matplotlib inline
         import numpy as np
         import random
         from sklearn.decomposition import PCA
         from sklearn.naive_bayes import GaussianNB
```

```
In [5]: m = np.matrix([[5,-3,0], [-3, 5,0],[0,0,4]])
        print(m)
```

```
[[ 5 -3  0]
 [-3  5  0]
 [ 0  0  4]]
```

```
In [6]: lamda, evectors = np.linalg.eig(m)
        lamda = np.float64(lamda)
        evectors = np.float64(evectors)
```

```
In [7]: print(lamda)
        print(evectors)
```

```
[ 8.  2.  4.]
[[ 0.70710678  0.70710678  0.        ]
 [-0.70710678  0.70710678  0.        ]
 [ 0.          0.          1.        ]]
```

Problem 11:

```
In [29]: # Part a)
         iris_df = pd.read_csv('./Iris.csv')
         iris_df.set_index('Id',inplace=True)
         print(iris_df.shape)
         iris_df.head()
```

```
(150, 5)
```

```
Out[29]:     SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
         Id
         1             5.1           3.5            1.4           0.2  Iris-setosa
         2             4.9           3.0            1.4           0.2  Iris-setosa
         3             4.7           3.2            1.3           0.2  Iris-setosa
         4             4.6           3.1            1.5           0.2  Iris-setosa
         5             5.0           3.6            1.4           0.2  Iris-setosa
```

```
In [30]: # Target values
         target_names = iris_df.Species.unique()
         target_values = np.array(iris_df["Species"])
         iris_df.drop(['Species'], axis=1,inplace=True)
         print(target_names)
```

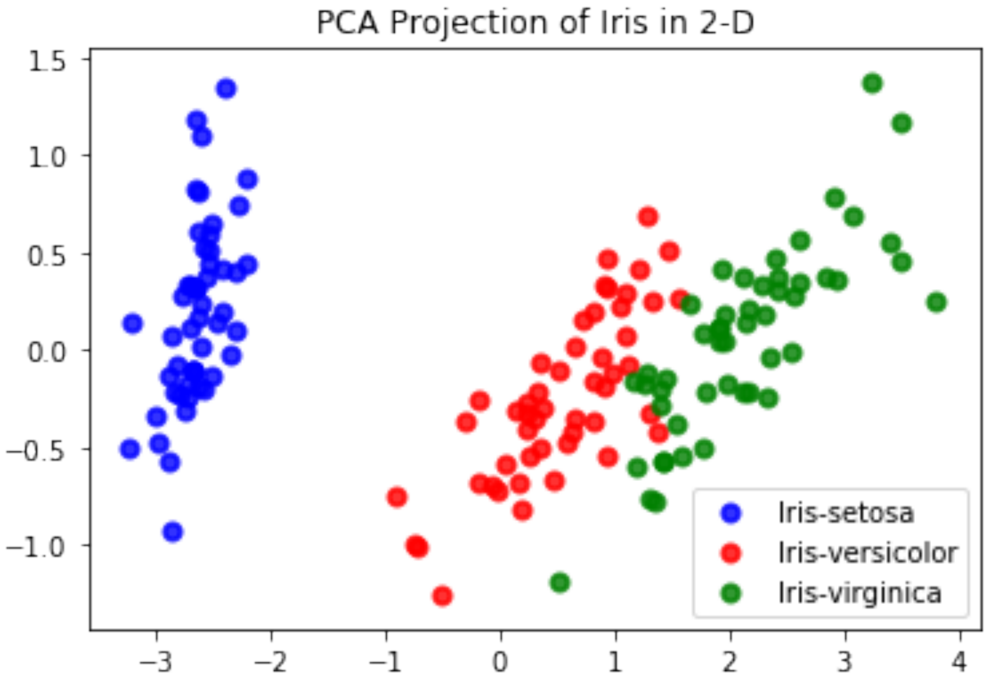
```
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
```

```
In [31]: # Fit 2 components to PCA
         pca = PCA(n_components=2)
         iris_2d = pca.fit(iris_df).transform(iris_df)
         print('explained variance ratio (first two components): %s'
               % str(pca.explained_variance_ratio_))
```

```
explained variance ratio (first two components): [ 0.92461621  0.05301557]
```

```
In [33]: plt.figure()
         colors = ['blue', 'red', 'green']

         for color, i, target_name in zip(colors, target_names, target_names):
             plt.scatter(iris_2d[target_values == i, 0], iris_2d[target_values == i, 1], color=c
                         label=target_name)
         plt.legend(loc='best', shadow=False, scatterpoints=1)
         plt.title('PCA Projection of Iris in 2-D');
```



These classes appear to be very well separated from each other. While virginica and versicolor appear closer together, the transformation does an adequate job separating them.

```
In [35]: # Part b)
         # Load data-set again to make it easier:
         iris_df2 = pd.read_csv('./Iris.csv')
         df_train = iris_df2[iris_df2['Species'] == target_names[0]][0:35]
         for t in target_names[1:]:
             df_train = pd.concat([df_train, iris_df2[iris_df2['Species'] == t][0:35]])

         df_test = iris_df2[iris_df2['Species'] == target_names[0]][35:]
         for t in target_names[1:]:
             df_test = pd.concat([df_test, iris_df2[iris_df2['Species'] == t][35:]])
```

```
In [40]: print(df_train.shape)
         print(df_test.shape)
```

```
(105, 6)
(45, 6)
```

```
In [41]: df_train_y = df_train["Species"]
         df_train.drop(['Species'], axis=1,inplace=True)
         df_test_y = df_test["Species"]
         df_test.drop(["Species"],axis=1,inplace=True)
```

```
In [43]: # Use Gaussian Naive Bayes classifier
         classifier = GaussianNB()
         classifier.fit(df_train, df_train_y)

Out[43]: GaussianNB(priors=None)

In [46]: pred = classifier.predict(df_test)

         miss = (df_test_y != pred).sum()
         accuracy = 1.0 - miss / df_test_y.shape[0]
         print("Accuracy: ", accuracy*100,"%")

Accuracy:  100.0 %
```

It would appear no smoothing constant is necessary, or the input of prior probabilities, as the base classifier did exceptionally well on the held-out test set. As indicated, I achieved a 0% error rate.