retail IRA marketing dataset

cse 103 HW 6 #2

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?

$P = (4/52) = 0.077$

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

$P = (4/52)(3/51) = 0.0045$

(c) The second card is an ace?

$P = (1-4/52)(4/52) + (4/52)(3/51) = .076$

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

$P(\text{ace/heart}) = \frac{P(A \wedge H)}{P(H)} = \frac{(1/52)}{(13/52)} = 1/13$

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

$P(\text{2nd Ace} / \text{1st Ace}) = \frac{P(\text{1st Ace} \wedge \text{2nd Ace})}{P(\text{1st Ace})} = \frac{(4/52)(3/51)}{4/52} =$

$3/51$

cse 103 #9

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 — dependent
- A: first card is a ten, B: second card is a heart — independent
- A: second card is a heart, B: fifth card is a club — dependent

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

DSE 210 ws 1 #5

(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?

5 characters $\Rightarrow \frac{1}{5!} \Rightarrow$ 0.008333

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

$6(1/6)(1/6)(1/6) =$ .0278

(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?

$\frac{1}{p} =$ 5 times

cse 103 #2

(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?

$$P(F/L) = \frac{P(L/F)P(F)}{P(L)} = \frac{P(L/F)P(F)}{P(L/M)P(M) + P(L/F)P(F)} = \frac{(.10)(.6)}{(.2)(.4)+(.1)(.6)} \Rightarrow$$

$$\frac{.06}{.14} = .43$$

DSE 210 ws 5 a)

(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?

$P(\text{medicine}) = 0.2 \qquad P(\text{placebo}) = 0.8$

$$P(\text{medicine} \mid \text{feel good}) = \frac{P(\text{feel good} \mid \text{medicine})\,P(\text{medicine})}{P(\text{feel good})} \Rightarrow$$

$$P(\text{feel good}) = P(\text{medicine})P(\text{feel good} \mid \text{medicine}) + P(\text{placebo})P(\text{feels good} \mid \text{placebo})$$

$$= 0.2(0.75) + (0.8)(.5) = 0.55$$

$$P(\text{medicine} \mid \text{feel good}) = \frac{(0.75)(0.2)}{.55} = 0.27$$

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 $\mathbb{E}(X)$?

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

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

$Var(X) = E(X^2) - \mu^2 = \left(\frac{1^2+2^2+3^2}{12} + \frac{4^2+5^2+6^2}{4}\right) - 4.25^2 =$

$\boxed{2.35}$

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

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

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

$Var\ 100 = \boxed{235}$

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

$means = \begin{bmatrix} 0 \\ .5 \end{bmatrix} \begin{bmatrix} var_1 & cov \\ cov & var_2 \end{bmatrix}$

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

$E(X_1) = P_1(1) + (1-P_1)(-1) = 0 \Rightarrow P_1 = 0.5$

$E(X_1^2) = P_1(1^2) + (1-P_1)(-1)^2$

$= (.5^2)(1^2) + (1-.5)(-1)^2$

$= 1$

$\sigma^2(X_1) = E(X_1^2) - [E(X_1)]^2$

$= 1 - 0 \Rightarrow 1$

$Corr(x_1, x_2) = \frac{cov(x_1,x_2)}{\sigma(x_1)\sigma(x_2)} \Rightarrow cov(x_1,x_2) = corr(x_1,x_2)\sigma(x_1)\sigma(x_2)$

$= .25\sqrt{1}\sqrt{.75}$

$= .22$

$E(X_2) = P_2(1) + (1-P_2)(-1) = .5$

$P_2 = 0.75$

$E(X_2^2) = P_2(1)^2 + (1-P_2)(-1)^2 = 1$

$\sigma^2(x_2) = E(X_2^2) - [E(x_2)]^2$

$= 1 - .5^2$

$= .75$

$\Sigma = \begin{bmatrix} 1 & .22 \\ .22 & .75 \end{bmatrix} \quad M = \begin{bmatrix} 0 \\ .5 \end{bmatrix}$

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 $\text{cov}(X)$ can be found in the following list:

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

3rd 2nd 1st

Circle them.

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

[8, 2, 4]

1st 2nd 3rd

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

need eigenvector with highest variance

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

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

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

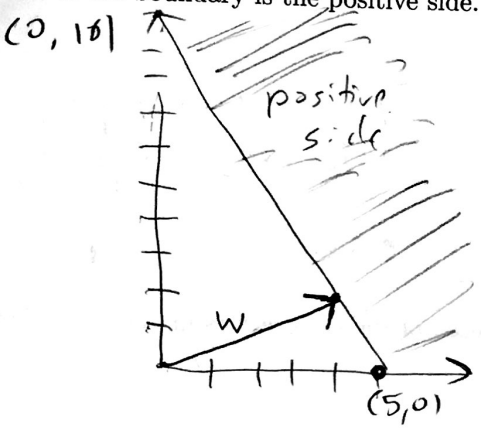
(e) Continuing from part (d), suppose that starting from the 2-d projection, we tried to reconstruct the original x. What would the three-dimensional reconstruction be, exactly?

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

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.



$2x+y \geq \theta$

$2x+y \geq 10$

$y \geq -2x+10$

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.

$\sigma = \sqrt{\frac{.4(1-.4)}{100}} = .0245$

$2\sigma = 0.098$

$.4 \pm 0.098$

(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?

$p = .4$

$2\sqrt{\frac{p(1-p)}{n}} = \pm .01 \Rightarrow$

$n = 9600$

main, yes, scale

check

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.

$n = 100$

$\text{mean} = \mu = 12.2$

$\sigma_{\text{sample}} = 5.4 \Rightarrow \sigma_{\bar{x}} = \frac{5.4}{\sqrt{100}} = .54$

12.2 ± 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.

Null: The means of the two schools have no significant statistical difference.

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

$\bar{X}_1 = 1930 \quad \bar{X}_2 = 1860$

$\sigma_2 = \frac{150}{\sqrt{100}} = 15 \qquad \sigma_2 = \frac{200}{\sqrt{100}} = 20$

$z = \frac{1930 - 1860}{\sqrt{15^2 + 20^2}} = 2.8$ Since $z > 2.8$, difference is significant.

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

p-value = .0026, the students are very different. which is strong evidence against the null.