CSE 151A - Homework 02

Due: Wednesday, April 15, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Wednesday at 11:59 p.m.

Essential Problem 1.

The table below shows the SAT scores for 10 randomly-selected applicants to UCSD, along with whether or not they were admitted.

Score	Admitted?
1240	Y
1310	Y
1470	Y
1500	Y
1200	N
1200	N
1220	N
1250	N
1290	N
1400	N

Your friend is applying to UCSD with an SAT score of 1300. Using Gaussians to estimate the conditional probabilities involved, use a Bayes classifier to predict whether your friend will be admitted or not. Show your work and all calculations involved.

Hint: You can use `scipy.stats.norm.pdf` to from the Python package `scipy` to evaluate the normal PDF (or another similar function in a different language). But make sure you know how the function works. In particular, does it require the standard deviation, or the variance?

Essential Problem 2.

The CDC is testing a vaccine for COVID-19 and has gathered a group of 44 people to experiment on. The people are classified by the state they are from, as shown in the table below.

State	#
Ohio	12
California	27
Texas	5

Suppose that 3 of the people from Ohio have COVID-19, 10 of the people from California have the virus, and one person from Texas has the virus. You randomly select one person from the study and learn that they are healthy – they do not have the virus. What is the probability that the person is from Texas?

Essential Problem 3.

In each part below, assume that you have gathered a data set consisting of the quantities described. Respond with a matrix containing the *sign* of each entry of the data's covariance matrix.

In each case there will be a preferred answer – but the correct answer isn't necessarily unique. If you feel unsure as to whether the sign of the covariance between two variables is positive or negative, make a guess and provide your reasoning. Otherwise, you do not need to show your work for this problem if you don't want to.

Example: Let X_1 be a person's height, and X_2 be their weight.

Solution: The signs of the entries of the covariance matrix are $\begin{pmatrix} + & + \\ + & + \end{pmatrix}$ because a person's weight tends to be larger the taller they are.

a) Let X_1 be a person's midterm score, let X_2 be their final exam score, and let X_3 be their GPA.

b) For a particular day, let let X_1 be the temperature, X_2 be the number of hours that the air conditioner ran, and X_3 be the number of winter coats sold on that day.

c) Let X_1 be the longest distance a person can run, let X_2 be their age, and let X_3 be a measure of the efficiency of their lungs (the larger X_3, the more efficient their lungs).

Essential Problem 4.

Let X be an $n \times d$ matrix, let A be a $d \times r$ matrix, and let B be an $r \times n$ matrix. Let $\vec{x}$ be a vector in $\mathbb{R}^n$ (that is, an $n \times 1$ column vector), let $\vec{y}$ be a vector in $\vec{R}^d$ (that is, a $d \times 1$ column vector). For each of the following, state whether the result is a scalar, a vector, or a matrix. If it is a vector or a matrix, state its shape (number of rows and columns).

For the purposes of this question, a matrix with one column is considered a column vector, and a matrix with one row is considered a row vector. If the result of an expression is 1×1, it is a scalar. You do not need to show your work.

a) $\vec{x} \cdot \vec{x}$

b) XA

c) $XX^\intercal$

d) $X^\intercal X$

e) $(XA)^\intercal \vec{x}$

f) $\vec{y}^\intercal \vec{y}(XX^\intercal)^{-1}$

g) $(\vec{x} \cdot \vec{x} + \vec{y} \cdot \vec{y}) + x^\intercal B^\intercal A^\intercal X^\intercal XAB\vec{x}$

h) $B^\intercal A^\intercal X^\intercal XAB$

Essential Problem 5.

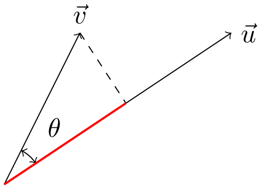
We will soon need to remember the key properties of the dot product. This question is meant to help you remember them.

a) Recall from your class on vector algebra that one way to define the dot product of two vectors, $\vec{u}$ and $\vec{v}$, is:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta,$$

where $\|\vec{u}\|$ is the length of the vector $\vec{u}$, $\|\vec{v}\|$ is the length of $\vec{v}$, and θ is the angle between the two vectors.

Two vectors $\vec{u}$ and $\vec{v}$ are shown below.



Argue that the length of the red segment is $(\vec{u} \cdot \vec{v})/\|\vec{u}\|$.

b) The function $f(\vec{x}) = 5x_1 + 2x_2 - 3x_3$ can be written as $f(\vec{x}) = \vec{w} \cdot \vec{x}$ for some vector $\vec{w}$. What is $\vec{w}$?

c) Let $\vec{x} = (x_1, \ldots, x_d)^T$ be a vector in $\mathbb{R}^d$. If $\vec{x}$ is a unit vector (that is, the length of $\vec{x}$ is 1) and $x_1 = 0.1$, what is the largest that any of the remaining entries $x_2, \ldots, x_d$ can possibly be?

d) What is the angle between $\vec{x} = (1, 2, 3)^T$ and $\vec{y} = (3, 2, 1)^T$?

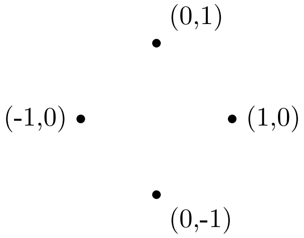
Plus Problem 1. (5 plus points)

In this problem, let X_1 and X_2 be random variables.

a) Show that if X_1 and X_2 are independent, then $\text{Cov}(X_1, X_2) = 0$

b) Independence implies zero covariance, but zero covariance does not imply independence in general. Here's an example demonstrating this.

Consider the four points below:



A point is chosen from these four, uniformly at random. Let X be its x-coordinate, and let Y be its y-coordinate. Show that X and Y are dependent, but that $\text{Cov}(X, Y) = 0$.

c) Zero covariance does not imply independence in general. However, in the special case that X and Y are *jointly* Gaussian random variables, $\text{Cov}(X, Y) = 0$ *does* imply that X and Y are independent. Recall that random variables X and Y are *jointly* Gaussian if the random vector $\vec{S} = (X, Y)^T$ has a density which is a two-dimensional Gaussian. The statement that X and Y have zero covariance is saying that the covariance matrix of the Gaussian describing their joint density is diagonal.

Prove that jointly Gaussian random variables with zero covariance are independent.

Plus Problem 2. (9 plus points)

Tumors are often diagnosed as malignant or benign through medical imaging. The file http://cse151a.com/data/cancer/train.csv contains data on 400 tumors collected as part of a breast cancer study at the University of Wisconsin. The data contains 30 measurements of each tumor, including the tumor's area, its perimeter, a measure of its texture, and so on. All features are continuous. The first column of the data reports whether the tumor was benign (`B`) or malignant (`M`). The file http://cse151a.com/data/cancer/test.csv contains a test set.

a) Create scatter plots of the `radius_mean` column versus the `texture_mean` column for benign and malignant tumors using the training data. Overlay your plots on the same graph.

b) Perform Linear Discriminant Analysis by estimating each class-conditional density with a Gaussian; the two Gaussians should share the same diagonal covariance matrix. Standardize each feature before performing your analysis. Report the error of your classifier on both the training set and the test set. Provide your code.

Hint 0: You can use libraries like `scipy` to evaluate the multivariate Normal pdf, but don't use code which performs LDA itself.

Hint 1: The top left entry of your shared covariance matrix should be roughly 0.46.

Hint 2: How do you get one covariance matrix for both classes? The lecture describes the standard approach.

Hint 3: The test set should be standardized too. When standardizing it, what makes the most sense: using the mean and variance from the training set, or from the test set? Oftentimes in practice we don't see the whole test set at once, but rather see one point at a time – you can make that limiting assumption here.

c) Perform Quadratic Discriminant Analysis by estimating each class-conditional density with a Gaussian; the two Gaussians should have different full covariance matrices. Standardize each feature before performing your analysis. Report the error of your classifier on both the training set and the test set. Provide your code.