DSC 40A - Final Exam

December 11, 2019

Name: SOLUTIONS

PID:

Version: A

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone enrolled in the course who has not yet taken the exam (this includes posting questions about the exam on Piazza!)

Signature:

Name of student to your **left**:

Name of student to your **right**:

Exam version of student to your **left**:

Exam version of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.) Your version should be different than either of the students to your left/right.

Instructions:

- Write your solutions to the following problems in the boxes provided.
- Scratch paper is provided at the end of the exam.
- No calculators are permitted, but a cheat sheet is.
- Write your name or PID at the top of each sheet in the space provided.

(Please do not open your exam until instructed to do so.)

Problem 1.

Determine the single best answer for each question. You are not penalized for guessing.

a) Let $L(h)$ be a loss function and let c be a constant. Define a new loss function $L'(h) = L(h) + c$. Then if h^* minimizes L, it also minimizes L'.

● True ○ False

b) Let A, B, and C be events. If A and B are independent, they must also be conditionally independent given C.

○ True ● False

c) Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be real numbers. Then

$$\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} x_i y_i.$$

○ True ● False

d) Suppose that when a straight line is fit to a data set using least squares regression the sum of squared errors is S_1. A new point is added to the data set and a new straight line is fit; let S_2 be the resulting sum of squared errors. It is possible for $S_2 < S_1$.

○ True ● False

Solution: Suppose the new SSE is smaller. The contribution to the new SSE of the new point is non-negative, so the contribution of the previous $n-1$ points to the SSE is smaller than the previous SSE. This implies that the SSE of the original line wasn't as small as possible, which contradicts the fact that we are using least squares regression.

e) Let $x_1, \ldots, x_n$ be a data set of real numbers. Then $L(h) = \sum_{i=1}^{n}(x_i - h)^2$ is minimized by the:

○ mode ○ midpoint ● mean ○ median

f) Let $x_1, \ldots, x_n$ be a data set of real numbers. Then

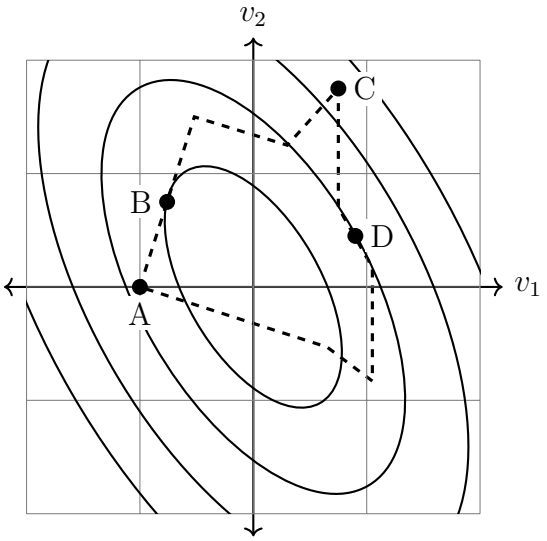
$$L(h) = \max\{|x_1 - h|, |x_2 - h|, \ldots, |x_n - h|\}$$

is minimized by the:

○ mode ● midpoint ○ mean ○ median

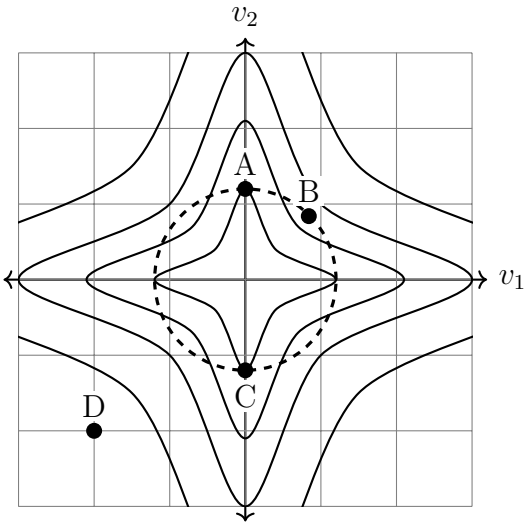
PID or Name: _____

g) Let $f(v_1, v_2)$ be a function whose value increases away from the origin. Suppose that the contour lines of f are shown below as solid lines, and suppose that the dashed lines represent a constraint. Which of the points shown **maximizes** the function f subject to the constraint?



◯ Point A ◯ Point B ● Point C ◯ Point D

h) Let $f(v_1, v_2)$ again be a function whose value increases away from the origin. Suppose that the contour lines of f are shown below as solid lines, and suppose that the dashed lines represent a constraint. Which of the points shown **maximizes** the function f subject to the constraint?



◯ Point A ● Point B ◯ Point C ◯ Point D

Problem 2.

Using least squares regression, fit a **quadratic** function of the form $y \approx c_0 + c_1 x^2$ to the data below:

x	y
-2	5
1	2
2	3
3	6

For your reference, recall that the least squares solutions for the slope b_1 and intercept b_0 of a linear fit to the data are:

$$b_1 = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right)$$

PID or Name: _____

$c_0 =$ $\frac{50}{33}$ $\quad$ $c_1 =$ $\frac{16}{33}$

Show your work here:

Solution: The function $c_0 + c_1 x^2$ is linear in the parameters. We can use the familiar equations for the estimators of slope and intercept if we replace each x_i with x_i^2.

That is, let $z_i = x_i^2$. Then:

$$c_1 = \frac{\sum_{i=1}^{n} z_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} z_i\right)}{\sum_{i=1}^{n} z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} z_i\right)^2}$$

$$c_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - c_1 \sum_{i=1}^{n} z_i\right)$$

x	y	z	zy	z^2
-2	5	4	20	16
1	2	1	2	1
2	3	4	12	16
3	6	9	54	81
sum:	16	18	88	114

So:

$$c_1 = \frac{88 - \frac{1}{4}\cdot 16 \cdot 18}{114 - \frac{1}{4}\cdot 18^2} = \frac{88 - 4\cdot 18}{114 - \frac{1}{4}\cdot 324} = \frac{88-72}{114-81} = \frac{16}{33}$$

And

$$c_0 = \frac{1}{4}\left(16 - \frac{16}{33}\cdot 18\right) = 4 - \frac{72}{33} = \frac{132-72}{33} = \frac{60}{33} = \frac{20}{11}$$

Problem 3.

Recall that an $n \times n$ matrix A is said to be *orthogonal* if $A^\intercal A = I$, where I is the $n \times n$ identity matrix. Show that if $\vec{u}$ is an eigenvector of A, and A is an orthogonal matrix, then $\vec{u}$ is also an eigenvector of $A^\intercal$.

Solution: Since $\vec{u}$ is an eigenvector of A, we have $A\vec{u} = \lambda\vec{u}$. Multiplying both sides by $A^\intercal$ gives:

$$A^\intercal A\vec{u} = A^\intercal\lambda\vec{u} = \lambda A^\intercal\vec{u}$$

Since A is orthogonal, the left hand side is simply $I\vec{u} = \vec{u}$. Hence:

$$\vec{u} = \lambda A^\intercal\vec{u} \implies A^\intercal\vec{u} = \frac{1}{\lambda}\vec{u}$$

Hence $\vec{u}$ is an eigenvector of $A^\intercal$ with eigenvalue λ.

Problem 4.

Let $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ be vectors in $\mathbb{R}^2$. Suppose that W is the matrix:

$$W = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}.$$

Let $L : \mathbb{R}^2 \to \mathbb{R}$ be the loss function defined by:

$$L(\vec{h}) = \sum_{i=1}^{n} \left\| W\left(\vec{x}^{(i)} - \vec{h}\right) \right\|^2.$$

Find the vector $\vec{h}^* = (h_1^*, h_2^*)^\intercal$ which minimizes L.

Solution: We will first write $W(\vec{x}^{(i)} - \vec{h})$ in coordinates:

$$\begin{aligned} W(\vec{x}^{(i)} - \vec{h}) &= \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix} \left((x_1^{(i)}, x_2^{(i)})^\intercal - (h_1, h_2)^\intercal \right) \\ &= \begin{pmatrix} w_1(x_1^{(i)} - h_1) \\ w_2(x_2^{(i)} - h_2) \end{pmatrix} \end{aligned}$$

So:

$$\begin{aligned} \|W(\vec{x}^{(i)} - \vec{h})\|^2 &= \left\| \begin{pmatrix} w_1(x_1^{(i)} - h_1) \\ w_2(x_2^{(i)} - h_2) \end{pmatrix} \right\|^2 \\ &= w_1^2(x_1^{(i)} - h_1)^2 + w_2^2(x_2^{(i)} - h_2)^2 \end{aligned}$$

And:

$$L(\vec{h}) = \sum_{i=1}^{n} \|W(\vec{x}^{(i)} - \vec{h})\|^2 = \sum_{i=1}^{n} \left[w_1^2(x_1^{(i)} - h_1)^2 + w_2^2(x_2^{(i)} - h_2)^2 \right]$$

Differentiating with respect to h_1, we find:

$$\frac{dL}{dh_1} = \sum_{i=1}^{n} w_1^2 \cdot 2(x_1^{(i)} - h_1) \cdot -1 = -2 \sum_{i=1}^{n} w_1^2(x_1^{(i)} - h_1)$$

Setting to zero and solving for h_1, we find:

$$h_1 = \frac{1}{n} \sum_{i=1}^{n} x_1^{(i)}$$

Repeating the above for h_2 yields $h_2 = \frac{1}{n} \sum_{i=1}^{n} x_2^{(i)}$.

Problem 5.

The holiday season is the busiest time of year for air travel. Flying anywhere at this time of year can be chaotic: tickets are oversold, flights are delayed, and luggage is lost. We can better understand each of these events using the tools of probability and combinatorics.

In what follows, assume that a particular airliner has 246 seats.

a) How many ways are there for the airline to assign seats if the flight is full (that is, 246 passengers must be assigned to 246 seats)?

Solution: $246!$

b) Suppose the flight is not full; it has only 150 passengers. How many ways are there for the airline to assign 150 passengers to 246 seats?

Solution: $\frac{246!}{(246-150)!} = \frac{246!}{96!}$

c) The airline oversold the flight, and 300 people showed up – but there are only 246 seats. How many ways are there for the airline to choose the people will board the plane?

Solution: $\binom{300}{246}$

d) Assume once again that 300 passengers arrive for a flight with only 246 seats. How many ways are there for the airline to assign seats? (Not everyone will get a seat).

Solution: $\binom{300}{246}246! = \frac{300!}{(300-246)!} = \frac{300!}{54!}$

e) You and your friends are 6 of the 300 people who arrived for the flight. Suppose that the airline chooses people at random to board the plane. What is the probability that all six of you are able to board?

Solution: We'll use as the outcome space all possible sets of boarded passengers. There are

$$\binom{300}{246}$$

such sets. How many of these sets contain you and your friends? There is one set for each way of choosing the 240 other people who board the plane from the other 294 people who arrived. That is, there are:

$$\binom{294}{240}$$

sets which include you and your friends. The probability is therefore:

$$\frac{\binom{294}{240}}{\binom{300}{246}}.$$

f) When doing their calculations, the airline assumes that an arbitrary person has a 98% chance of actually showing up for their flight. Assuming that this is true, what is the probability that exactly 300 people show up?

Solution: 0.98^{300}

g) Out of the 246 seats on the plane, 82 are window seats, 82 are middle seats, and 82 are aisle seats. How many ways are there of choosing who sits next to a window, who sits in the middle, and who sits next to the aisle if 246 passengers board the plane?

Solution:

$$\binom{246}{82}\binom{164}{82}\binom{82}{82}$$

Airline	% of flights on time	% of all flights
Northeast	7/8	1/6
Epsilon	3/4	1/2
Divided	4/5	1/3

The table above shows the fraction of flights which are on time for three airlines, along with the fraction of all flights that the airline is responsible for. For instance, Northeast Airlines operates 1/6 of all flights, and 7/8 of their flights are on time.

h) Suppose that the probability that a flight encounters a storm is 1/5, independent of which airline operates the flight. What is the probability that a randomly-selected flight is operated by Epsilon **or** it encounters a storm?

Solution: We will calculate

$$\begin{aligned} P(\text{Epsilon or Storm}) &= P(\text{Epsilon}) + P(\text{Storm}) - P(\text{Epsilon and Storm}) \\ &= 1/2 + 1/5 - P(\text{Epsilon and Storm}) \end{aligned}$$

Since the event that a storm occurs is independent of airline, we have:

$$\begin{aligned} &= 1/2 + 1/5 - P(\text{Epsilon}) \cdot P(\text{Storm}) \\ &= 1/2 + 1/5 - 1/2 \cdot 1/5 \\ &= \frac{1}{2} + \frac{1}{5} - \frac{1}{10} \\ &= \frac{1}{2} + \frac{1}{5} - \frac{1}{10} \\ &= \frac{6}{10} = \frac{3}{5} \end{aligned}$$

i) Let E_1 be the event that a flight arrives on time, and let E_2 be the event that the pilot forgot the keys to the airplane. Are these two events independent?

◯ Yes ● No

j) A randomly-selected flight arrived on time. What is the probability that the flight was operated by Northeast?

Solution: This is an opportunity to use Bayes' Theorem. We have:

$$P(\text{Northeast} \,|\, \text{on-time}) = \frac{P(\text{on-time} \,|\, \text{Northeast}) \cdot P(\text{Northeast})}{P(\text{on-time})}$$

We know all of these quantities except for $P(\text{on-time})$, but we can calculate it:

$$\begin{aligned} P(\text{on-time}) = & \\ & P(\text{on-time} \,|\, \text{Northeast}) \cdot P(\text{Northeast}) \\ & + P(\text{on-time} \,|\, \text{Epsilon}) \cdot P(\text{Epsilon}) \\ & + P(\text{on-time} \,|\, \text{Divided}) \cdot P(\text{Divided}) \\ & = \frac{7}{8} \cdot \frac{1}{6} + \frac{3}{4} \cdot \frac{1}{2} + \frac{4}{5} \cdot \frac{1}{3} \end{aligned}$$

So:

$$P(\text{Northeast} \,|\, \text{on-time}) = \frac{\frac{7}{8} \cdot \frac{1}{6}}{\frac{7}{8} \cdot \frac{1}{6} + \frac{3}{4} \cdot \frac{1}{2} + \frac{4}{5} \cdot \frac{1}{3}}$$

Problem 6.

In parts of the world other than San Diego, the weather changes from day to day. In these places, people try to guess tomorrow's weather using the current conditions.

Weather data for 20 random days in Columbus, Ohio are recorded below, along with the next day's weather (rainy, cloudy, or sunny).

Suppose that today's humidity is $> 50\%$, the temperature is hot, and the air pressure is low. Use naïve Bayes to predict whether tomorrow will be rainy, cloudy, or sunny. Show your work.

Next Day's Weather	Humidity	Temperature	Air Pressure
Rainy	$> 50\%$	Cool	Low
Rainy	$> 50\%$	Hot	Low
Rainy	$> 50\%$	Cool	Low
Rainy	25%-50%	Hot	High
Rainy	25%-50%	Hot	Low
Rainy	25%-50%	Cool	Low
Rainy	25%-50%	Cool	Low
Rainy	$< 25\%$	Cool	Low
Rainy	$< 25\%$	Hot	Low
Rainy	$< 25\%$	Hot	High
Cloudy	$> 50\%$	Cool	Low
Cloudy	$> 50\%$	Cool	Low
Cloudy	25%-50%	Hot	High
Cloudy	$< 25\%$	Cool	High
Cloudy	$< 25\%$	Cool	Low
Sunny	$> 50\%$	Cool	Low
Sunny	$> 50\%$	Hot	High
Sunny	$> 50\%$	Cool	High
Sunny	25%-50%	Hot	High
Sunny	$< 25\%$	Hot	High

PID or Name: _____

Prediction: tommorow will be ● rainy ○ cloudy ○ sunny

Show your work here:

Solution: Naïve Bayes calls for computing three things and seeing which is the largest:

$$\begin{aligned}
P(\text{Rainy} \mid >50\% \text{ and hot and low}) &\propto P(>50\% \mid \text{Rainy}) \cdot P(\text{hot} \mid \text{Rainy}) \cdot P(\text{low} \mid \text{Rainy}) \cdot P(\text{Rainy}) \\
&\approx \frac{3}{10} \cdot \frac{5}{10} \cdot \frac{8}{10} \cdot \frac{10}{20} \\
&= \frac{1200}{20000} \\
P(\text{Cloudy} \mid >50\% \text{ and hot and low}) &\propto P(>50\% \mid \text{Cloudy}) \cdot P(\text{hot} \mid \text{Cloudy}) \cdot P(\text{low} \mid \text{Cloudy}) \cdot P(\text{Cloudy}) \\
&\approx \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{20} \\
&= \frac{30}{2500} = \frac{240}{20000} \\
P(\text{Sunny} \mid >50\% \text{ and hot and low}) &\propto P(>50\% \mid \text{Sunny}) \cdot P(\text{hot} \mid \text{Sunny}) \cdot P(\text{low} \mid \text{Sunny}) \cdot P(\text{Sunny}) \\
&\approx \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{5}{20} \\
&= \frac{45}{2500} = \frac{360}{20000}
\end{aligned}$$

So it is most likely to be rainy tomorrow.

Problem 7. (Extra Credit)

Draw a picture that conveys your feelings about the fact that winter break is (almost) here.

Solution: :)

Before turning in your exam, please check that your name is on every page.
After turning in your exam, have a good break!