https://www.overleaf.com/1872993542yympnmrdjfxg

CSE 152 Mid-Term Examination

University of California San Diego

Oct 29, 2019

Name: _____

UCSD ID: _____@ucsd.edu

Question	**Points**	**Points Earned**
1	10	
2	10	
3	10	
4	10	
5	10	
6	10	
7	10	
8	10	
9	10	
Total	90	

Instructions:

1. This examination contains 12 pages, including this page.
2. You have **one (1) hour** to complete the examination. As a courtesy to your classmates, we ask that you not leave during the last fifteen minutes.
3. Write your answers in this booklet. We scan this into Gradescope, so **please try to avoid writing on the backs of pages**. If you must do so, please indicate **very** clearly on the front of the page that you have written on the back of the page.
4. You may use two (2) double-sided 8.5" $\times$ 11" pages with notes that you have prepared. You may not use any other resources, including lecture notes, books, other students or other engineers.
5. You may use a calculator. You may not share a calculator with anyone. If you didn't bring a calculator, you may use your phone, **but** you must put it on **flight mode** and clear all visible notifications **before** the examination starts, and you must not open any applications other than the calculator and a timer.

Question 1: Basic Linear Algebra

$$A = \begin{bmatrix} 1 & 0 & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & \lambda \end{bmatrix}$$

(a) Calculate the rank of A when $\lambda = 1$. Is A invertible? [2 pts]

Rank is 2. Not invertible as it is not full-rank.

(b) What's the null space of A when $\lambda = 1$? What's the dimension of its null space? [3 pts]

Solve Ax = 0. Find out that $x = [k, 0, -k]^\top, k \in \mathbf{R}$. Dimension is 1.

(c) Give the condition of λ when A is a full rank matrix. What is the dimension of A's null space for such λ's? [5 pts]

Calculate the determinate, which is $2\lambda - 2$. So when $\lambda \neq 1$, A will be full rank. And the dimension of its null space is be zero, as Ax=0 have no non-zero solution for x. (rank of null(A) = size(A)-rank(A))

Question 2: Least Squares Problem

(a) For an over-determined linear system $Ax = b$, where A is an $m \times n$ tall matrix ($m > n$) with linearly independent columns (full rank). When would this linear system have no solution?

Suppose $A = [a_1, a_2, \ldots, a_n]$. So $Ax = \Sigma_1^n a_i x_i$, which is a linear combination of A's columns. If b is not in the range of A's columns, which means that b cannot be represented as the linear combination of a_i, then there will be no such x for $Ax = b$.

(b) Write down the least squares problem formulation [2 pts], and derive the solution [2 pts].

The objective is min $\|Ax - b\|^2$. By expand the square term we will get the objective to be $x^\top A^\top Ax - x^\top A^\top b - b^\top AX + b^\top b$. Calculating the gradient to x, and set it to zero will give the solution, which is $x = (A^\top A)^{-1} A^\top b$

(c) Consider an under-determined linear system, where A is a fat matrix ($m < n$). How many solution(s) does $Ax = b$ have? Select from below (you can make multiple choices if more than one could happen). Briefly explain why. [2 pts]

No solution | **one solution** | **Infinite solutions**

Similar as (a), if b is not in the range of A's columns, the system will have no solution. Otherwise, as A's columns are not linearly independent (A is a fat matrix), the linear combination of $Ax = b$ can have infinite solutions.

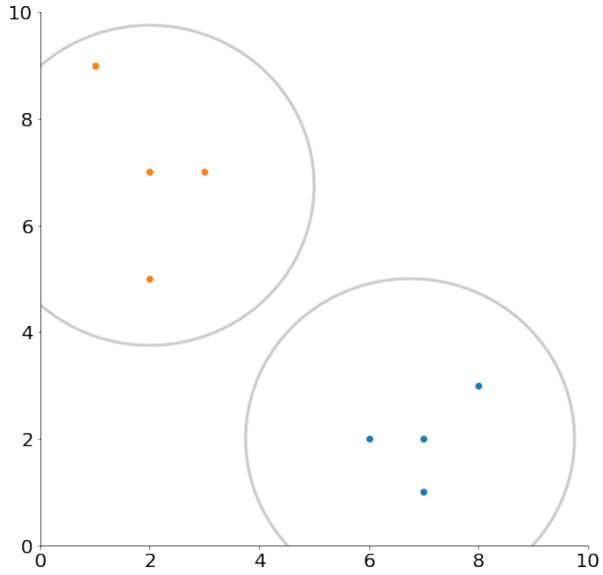
(d) In class, we taught that we may obtain a special solution for this under-determined system by $x = A^\dagger b$ using the pseudo inverse, where $A^\dagger = A^T(AA^T)^{-1}$. What is the property (geometric interpretation) of this special solution? [2 pts]

It's the minimum norm solution, which is $min\|x\|^2, \text{s.t.} Ax = b$. Geometric interpretation is, $Ax = b$ gives a hyper-plane, and this hyper-plane is closest to the zero point (origin).

Question 3: K-Means

Point	X Coord.	Y Coord.
1	6	2
2	7	1
3	8	3
4	7	2
5	3	7
6	1	9
7	2	7
8	2	5

(a) Plot the points on the graph and draw the circles around the two clusters of data. [2 pts]



b) The first iteration of your k-means algorithm initialized the cluster centers at $C_1 = (2,6)$ and $C_2 = (6,2)$. Calculate the distance between Point 2 and it's cluster center. [3 pts]

Point 2 is at $(7, 1)$

$$\sqrt{(7-6\cdot)^2 + (1-2)^2} = \sqrt{2}$$

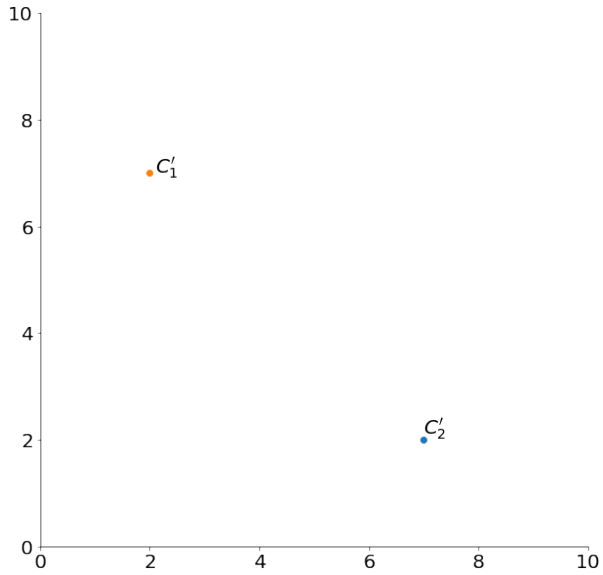
c) Using the work you have done in part a), draw the (rough) new locations for clusters C_1 and C_2 for the next iteration of the algorithm. [5 pts]

Points 1 through 4 belong to cluster 1, points 5 through 8 belong to cluster 2. To calculate the new cluster centers one has to calculate the mean location of the points in each of the clusters.

Call the new cluster centers C_1' and C_2':

$$C_1' = \left(\frac{3+1+2+2}{4}, \frac{7+9+7+5}{4}\right) = (2, 7)$$

$$C_2' = \left(\frac{6+7+8+7}{4}, \frac{2+1+3+2}{4}\right) = (7, 2)$$



Question 4: Bag of Visual Words

Creating and implementing a Bag of Words model involves the following steps:

Extract Features
↓
Learn Visual Vocab
↓
Quantize Features
↓
Represent Image by Frequency of Visual Words

a) Explain what occurs at each stage of the process. [8 pts]

Extract Features:

Information is extracted from patches of pixels in the image. This information can be the individual pixel values, gradients, edges, corners, or other feature information. That information is subsequently vectorized into a higher dimensional space.

Learn Visual Vocab:

The information is classified and assigned a label. In our class we use K-means clustering to create a visual vocabulary of words from the vectorized data. An index is assigned to each cluster and this index becomes a word in the visual dictionary.

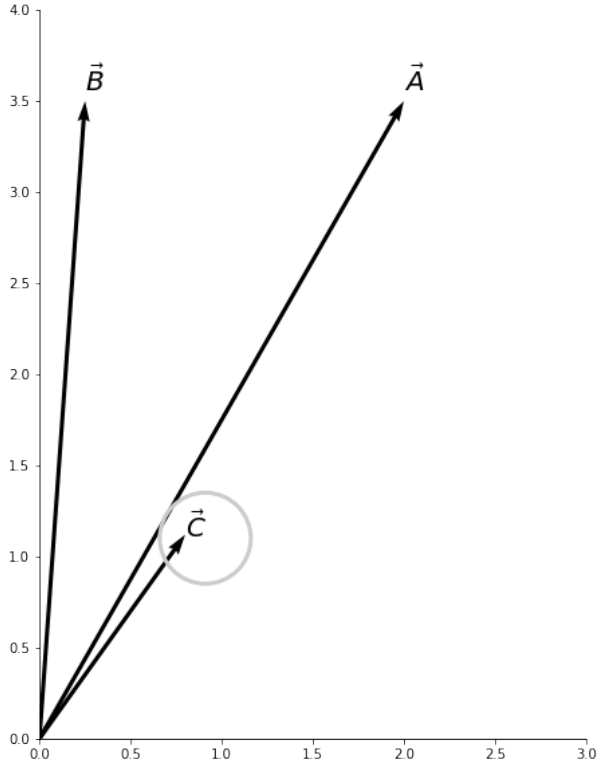
Quantize Features:

At this stage of the pipeline we are looking to classify an image. Therefore, for a given image we have to extract features for each patch as we did in step one, and then use the features to retrieve the closest cluster center. By assigning the cluster index to the patch, we quantize a patch.

Represent Image by Frequency of Visual Words:

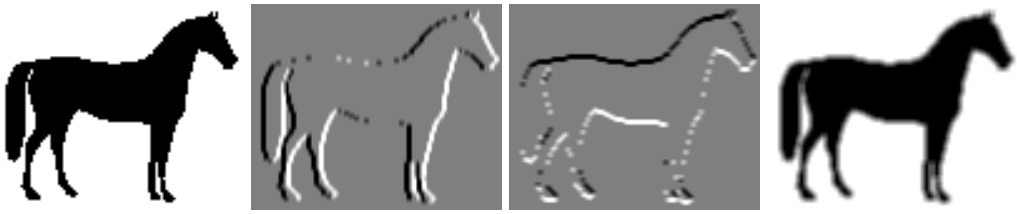
Once the visual words have been aggregated, a histogram is built to represent the frequency of visual words in an image. In this form, the image can be compared to other images via cosine similarity or can be fed into a machine learning model trained to classify images based on visual word distributions.

b) We utilize the cosine similarity metric to compare the visual words distributions between images. In the image below, indicate which vector has the higher cosine similarity to $\vec{A}$ by circling the vector letter. [2 pts]



Question 5: Filters

(a) An input image shown below is filtered by some image filters to produce A, B, and C. Briefly explain the effect of each filter. [6 pts]



input A B C

1. $\frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$ 2. $\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}$ 3. $\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix}$

Explain.

1. C, box filter, blur the input by averaging with the neighbors.
2. A, x gradient, find vertical edges
3. B, y gradient, find horizontal edges

(b) Given a gray-scale image I with shape (H, W). Let $\text{Filter}(K, I)$ be the operation that filters image I with kernel K. In this operation, the filtered result has shape (H, W) and the original image is padded with 0s if necessary. Prove

$$\text{Filter}\left(\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}, I\right) = \text{Filter}\left(\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}, \text{Filter}(\begin{bmatrix}1 & 1 & 1\end{bmatrix}, I)\right)$$

[4 pts]

Method 1: Since image filtering is associative,

$$\text{Filter}\left(\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}, \text{Filter}(\begin{bmatrix}1 & 1 & 1\end{bmatrix}, I)\right) = \text{Filter}\left(\text{Filter}\left(\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}, \begin{bmatrix}1 & 1 & 1\end{bmatrix}\right), I\right) = \text{Filter}\left(\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}, I\right)$$

Method 2 (proof outline, you can prove it with more concrete math):

LHS: each pixel in the image is the sum of itself and 8 neighbors.

RHS:

inner filter: each pixel is summed with its left neighbor and its right neighbor.

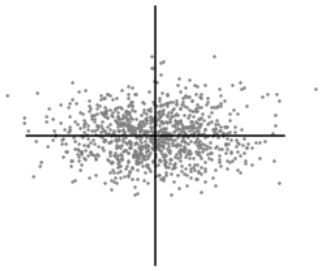
outer filter: each pixel is the sum of itself, its upper neighbor, and lower neighbor, which are summed with their left and right original neighbors. So the overall effect is also the sum of itself and 8 neighbors.

Question 6: Principal Component Analysis

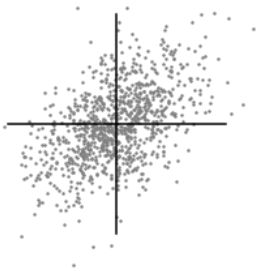
(a) Draw the direction of the first principal component in figure (a). [1 pt]

(b) Draw the direction of the second principal component in figure (b). [1 pt]

Note that this is the second component.



(a)



(b)

(c) Suppose that we have already obtained the SVD result for some centered data matrix X (each row is the vector of a data point), which satisfies $X = U\Sigma V^\top$. By this SVD decomposition, what are the principal components and eigenvalues of PCA? [4 pts]

$$X^\top X = V\Sigma^\top \Sigma V^\top$$

So eigenvalues are given by Σ^2, and principal components are given by the the right singular vectors of X, which are the columns of $V^\top$.

(d) Can you explain PCA from the optimization's perspective? What is the goal/objective of PCA [2 pts]? Is there any intuitive explanation? [2 pts]

The goal of PCA is to find a low dimensional space to preserve the structure of the data as much as possible. Suppose that $U = [u_1, u_2, ..., u_k]$ is the basis of the low dimensional space, and we can project and reconstruct the data by $XUU^\top$. So the objective of PCA will be $\min \|XUU^\top - X\|^2$, which will minimize the reconstruction error.

Question 7: Harris Corner

(a) Here is the second moment matrix for Harris corner detection.[6 pts]

$$\sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

When evaluated at 3 patches A, B, C of an image, the eigenvalues of the matrices are

$$A : \lambda_1 = 0.9, \lambda_2 = 0.01; B : \lambda_1 = 0.012, \lambda_2 = 0.014, C : \lambda_1 = 0.9, \lambda_2 = 0.92$$

If we know that among A, B, C, there is a corner, an edge, and a flat region. What are A, B, C respectively? Fill in the blanks below with A, B, C.

Flat region: B Corner: C Edge: A

(b) Fill in the the blank space below in the pseudo-code for Harris corner response. Given the second moment matrix M, the Harris corner response is computed by

$$\det(M) - 0.06(\text{tr}(M))^2$$

where

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc, \text{tr}\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = a + d$$

You can assume addition, subtraction, multiplication, division, and square operations are performed element-wise on multi-dimensional arrays. [4 pts]

---

function HARRISCORNER(I)
  $I_x \leftarrow x\_gradient(I)$
  $I_y \leftarrow y\_gradient(I)$

  $W \leftarrow \begin{bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{bmatrix}$ ▷ W is a 3x3 box filter

  $A \leftarrow \text{convolve}(W, I_x I_x)$

  $B \leftarrow \text{convolve}(W, IxIy)$

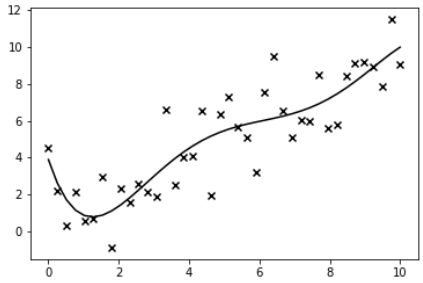
  $C \leftarrow \text{convolve}(W, IyIy)$

  $response \leftarrow AC - B^2 - 0.06(A + C)^2$

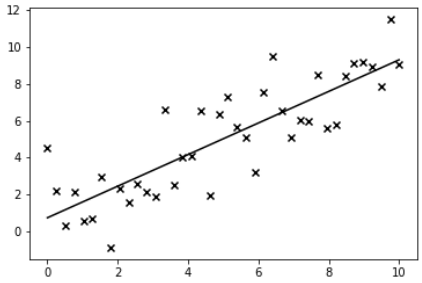
  return *response*
end function

Question 8: Bias and Variance

(a) The following images show two type of models for a regression task. [5 pts]



model A (K-Nearest Neighbor)



model B (Linear Regression)

Fill in the blanks (circle the correct answer) below.

Compared to Model B, Model A has _lower_ (lower/higher) bias and _higher_ (lower/higher) variance. Model A has _worse_ (better/worse) generalizability and is prone to _overfitting_ (overfitting/underfitting). Model A is _more_ (more/less) sensitive to changes in the training data.

(b) Explain what "hyperparameter" for a model is [1 pt].

Hyperparamters of a model are choices about the algorithm that we set rather than learn.

(c) Give an example of "hyperparameter" (e.g., a hyperparameter of Harris corner, KNN, K-Means, or Neural network) [1 pt].

Number of centers in K-means. Number of neighbors in KNN. etc.

(d) Describe the procedure to choose hyperparameters with cross-validation. (You need to list the steps in cross-validation.)[3 pts]

1. Split training data into training set and testing set.
2. Split training data into k folds.
3. For each fold, train on the other folds and validate on this fold. Record some evaluation metric.
4. Pick hyperparameters based on average of metrics validated on all folds from training-validation.
5. Train on the whole training set and test on the unseen testing set.

Question 9: Logistic Regression

Class	$S = f(x_i; w)$	Normalized Probability	$P(Y = k \mid X = x_i)$
Cat	3.4		
Dog	5.9		
Plane	-.5		

a) Explain the core function of the Softmax classifier in one to two sentences. [4 pts]

The Softmax function transforms raw classifier scores from a machine learning model into probabilities for each of the classes in a multi-class model. It allows us to interpret scores as probabilities

b) Given the model scores in the first column, what steps must you take in order to calculate class probability in the third column? (Explain process, do not do perform calculations) [4 pts]

The raw scores from the model are the unnormalized log probabilities, therefore to calculate the values in column 2, we must take the natural exponent raised to the value in the first column. The value in row one of column 2 is $e^{3.4}$. Column 3 is the normalized probability. Therefore it is calculated by normalizing the scores in the second column by the sum of all of the values in the second column. For example:

$$P(Y = cat \mid X = x_i) = \frac{e^{3.4}}{\sum e^{x_i}}$$

c) Which class has the highest probability? [2 pts]

The Dog