MLT - Week 1

- Popularity of AI and ML is ever increasing. Egs. of ML applications are - vision, speech, text, etc.
- Data-driven, generalize (do well on unseen data)
- Spam vs non span - Binary, rainfall classification - regression, recommending movies - ordinal, friend suggestions - link prediction, grouping pictures in phone - clustering, robot navigation - reinforcement learning, stock market prediction - online learning

Activity Question 1

1) What are some of the most important types of data that machine learning models deal with?

- [x] Text ✓
- [x] Speech ✓
- [x] Vision ✓

directly from the lecture

2) Map the dataset on the left to the domain on the right:

Dataset description	Domain
(1) Record of all IPL matches played so far	(a) Finance
(2) Credit profile of customers	(b) E-Commerce
(3) 3D structure of proteins	(c) Bioinformatics
(4) Click-stream data of Amazon consumers	(d) Sports analytics

sports
finance
related to biology
e-commerce store

We could define a machine learning problem associated with each dataset on the left.

- ( ) (1) - (d), (2) - (b), (3) - (c), (4) - (a)
- (•) (1) - (d), (2) - (a), (3) - (c), (4) - (b) ✓

3) E-commerce recommendations were discussed briefly in the lecture. When a customer visits a website to buy a product, which of the following are important features that a recommendation system should take into account while making its recommendation?

- [ ] Name of the customer ✗
- [x] Age of the customer → helps since preferences differ for kids, adults & so on.
- [x] Gender of the customer → M/F choices
- [x] Products that the customer has purchased in the past → it was his liking at a pt. of time
- [x] Profiles of customers who have a similar purchase behaviour → how other customers behave/choices & when they buy that prod.
- [ ] The customer's PAN number

4) Can you think of domains other than e-commerce where recommendations play an important role?

Films, Funds, etc.

5) You are given the data of marks scored by students in a class in a csv file. Your task is to find the class topper. Is this a machine learning problem? 1 p

- ( ) Yes
- (•) No

No, becoz there is no such ML thing going on, no analytics, etc.

6) In the previous problem, you have access to a dataset. Consider the following claim:

"Data is an important aspect of any machine learning problem. So, every problem that has a dataset associated with it can be classified as a machine learning problem."

What is wrong with this claim?

dataset is good, but what do we do, with that, ML is the sc. of study of the given data, and then find new predictions

- Supervised L - classification, regression, ranking, structure learning
- Unsupervised L - clustering, representational learning
- Sequential L - online learning, multi-armed bandits, reinforcement learning
- Develop an ability to pose anything as ML problem and understand algorithms
- Structure - Algebra, Uncertainity - Statisitics, Decisions - Optimization

Activity Questions - 2

1) What are some of the *broad* paradigms of machine learning discussed in the lecture?

- [x] Supervised learning
- [ ] Regression
- [x] Unsupervised learning
- [ ] Clustering
- [x] Sequential learning
- [ ] Online learning

there 3 broad paradigms are just parts of the major 3.

2) Which of the following scenarios can be modeled as a binary classification problem?

- [x] Given the image of a chest X-Ray, determine the presence or absence of a tumor.
- [x] Given a customer's financial background, determine if the customer can be given a loan or not.
- [ ] Given the history of rainfall in a given region, predict the amount of rainfall for a given duration
- [ ] Given a sound clip, determine the number of musical instruments that are present in it.

it has only 2 options

either 0 or 1 , {-1, 1}

3) For every mail that comes to your inbox, you have to design an algorithm that can assign exactly one of these four labels to it:

- family
- friends
- work
- spam

What type of machine learning problem does this correspond to?

- ( ) Regression
- ( ) Binary classification (exactly 2 categories)
- (•) Multi-class classification (more than 2 categories) ✓
- ( ) This is not a supervised learning problem

4 labels → multi-classification problem ⇒ more than 2

4)

Scenario-1: You are given a bucket of red and blue balls. Someone commands you to separate them into two separate buckets of uniform color.

Scenario-2: You are given a bucket of black balls that look identical, but weigh differently. There is no one around you to tell you what to do. However, you try to separate the balls into different buckets, such that balls in a given bucket have more or less the same weight.

if a machine were to be trained to do these actions, what paradigms of machine learning would they correspond to?

- [x] Scenario-1: supervised learning ✓
- [x] Scenario-2: unsupervised learning ✓
- [ ] Scenario-1: unsupervised learning
- [ ] Scenario-2: supervised learning

supervision

no supervision

original: # real numbers = dn

real numbers in compressed representation = $d + n$

inner product / dot product of x and w.

$c^* = \left(\frac{x_1 w_1 + x_2 w_2}{w_1^2 + w_2^2} \right)$ scalars

length$^2 \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ ←

NOTE: can always pick $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ s.t. length$\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}\right) = 1$

$\Rightarrow c^* = (x^T w) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

Activity Question - 3

1) Is the following statement true or false:

Compression is the act of throwing away a fraction of data-points from the dataset.

- ( ) True
- (•) False ✓

it's not throwing away, it's just that we want to show original data in a minor way, w/o loosing anything.

2) Consider the following dataset:

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ -3 \\ 4 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \\ 9 \\ -4 \end{bmatrix} \right\}$$

each x_i is the data pt., we 5 pts. here (n)

these are features in each data pt. ⇒ $d = 4$

If n represents the number of data-points and d represents the number of features, what are the values of n and d?

- ( ) $n = 4, d = 5$
- (•) $n = 5, d = 4$ ✓
- ( ) $n = d = 4$

3) Consider the following dataset?

$$\left\{\begin{bmatrix}3\\6\\-3\end{bmatrix}, \begin{bmatrix}5\\10\\-5\end{bmatrix}, \begin{bmatrix}-2\\-4\\2\end{bmatrix}, \begin{bmatrix}0\\0\\0\end{bmatrix}\right\}$$

these 2 are compressed representations, less no. of data pts. to show the same data ✓

$$\begin{bmatrix}-2\\-4\\2\end{bmatrix}\left((-1.5), (-2.5), (1), (0)\right)$$

Which of the following vectors could be chosen as a representative for the above dataset?

- ☐ $\begin{bmatrix}0\\0\\0\end{bmatrix}$
- ☐ $\begin{bmatrix}1\\1\\1\end{bmatrix}$

} → but they are not good representatives, their representation takes more than usual.

- ☑ $\begin{bmatrix}-2\\-4\\2\end{bmatrix}$ ✓
- ☑ $\begin{bmatrix}1\\2\\-1\end{bmatrix}$ ✓

$$\begin{bmatrix}1\\2\\-1\end{bmatrix}\left((3), (5), (-2), (0)\right)$$ ✓

4) Consider the following dataset in $\mathbb{R}^2$:

$$\left\{\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\4\end{bmatrix}, \begin{bmatrix}-3\\-6\end{bmatrix}, \begin{bmatrix}0\\0\end{bmatrix}\right\}$$ → 8

Let's say $\begin{bmatrix}1\\2\end{bmatrix}\left((1), (2), (-3), (0)\right)$ → 6 data pts.

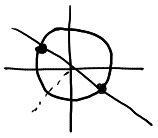
If we use a representative and the corresponding coefficients to come up with an alternative representation of the points, what would be the reconstruction error for this dataset?

0

since, we can get the whole dataset using a representation → 0 reconstruction error,

5) You have a dataset of 100 points in $\mathbb{R}^3$ where all of them lie on a line passing through the origin. You wish to choose a representative $\mathbf{w}$ for this dataset such that $||\mathbf{w}|| = 1$. How many representatives can be chosen?

- ○ 0
- ○ 1
- ◉ 2
- ○ infinite



$\|w\| = 1 \Rightarrow$ circle

there 2 such pts. for representation.

6) Let $\mathbf{w}$ be a vector that represents a line L passing through the origin. What is the projection of a point $\mathbf{x}$ onto this line? Note that $||\mathbf{w}||$ is not necessarily equal to 1.

- ○ $\mathbf{x}^T\mathbf{w}$
- ○ $(\mathbf{x}^T\mathbf{w})\mathbf{w}$
- ◉ $\left(\frac{\mathbf{x}^T\mathbf{w}}{||\mathbf{w}||^2}\right)\mathbf{w}$

$$p = \left(\frac{x^T w}{\|w\|^2}\right) w$$

if not

$\because \|w\| = 1 \Rightarrow p = (x^T w)\, w$

Goal: Find the line that has the least "reconstruction" error.

$$\min_{w:\ \|w\|^2=1} g(w) = \frac{1}{n}\sum_{i=1}^{n} -(x_i^T w)^2$$

Equivalently

$$\max_{w:\ \|w\|^2=1} \frac{1}{n}\sum_{i=1}^{n} (x_i^T w)^2$$

$$\max_{w:\ \|w\|^2=1} w^T C w$$

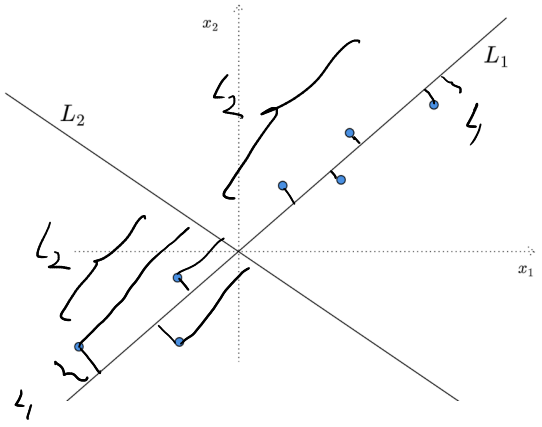
$$C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

↳ Co-variance matrix

Soln: w is the eigenvector corresponding to the maximum eigen value of C

Activity Question - 4

1) Consider the following dataset and two lines L_1 and L_2:



Which of the two lines is a good choice for a compressed representation of the dataset and why?

- ○ L_1, as it gives a better compression ratio
- ○ L_2, as it gives a better compression ratio
- ◉ L_1, as it gives a smaller reconstruction error
- ○ L_2, as it gives a smaller reconstruction error

→ L_2 gives a larger reconstruction error.

2) Which of the following expressions is the reconstruction error for a dataset of n points, with respect to a line passing through the origin represented by the vector $\mathbf{w}$. Note that $||\mathbf{w}|| = 1$. (MSQ)

- ☑ $\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{w})\mathbf{w}||^2$ ✓ → $\frac{1}{n}\sum_{i=1}^{n} ||x_i - (x_i^T w)w||^2$
- ☑ $\frac{1}{n}\sum_{i=1}^{n} [\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{w})\mathbf{w}]^T [\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{w})\mathbf{w}]$ → $\frac{1}{n}\sum_{i=1}^{n} [x_i - (x_i^T w)w]^T$
- ☑ $\frac{1}{n}\sum_{i=1}^{n} [\mathbf{x}_i^T\mathbf{x}_i - (\mathbf{x}_i^T\mathbf{w})^2]$ ✓ $[x_i - (x_i^T w)w]$
- ☐ $\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{w})^2$ → $\frac{1}{n}\sum_{i=1}^{n} [x_i^T x_i - (x_i^T w)^2]$
- ☐ $\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i||^2$

3) Is the following statement true or false?

The reconstruction error is always a non-negative scalar quantity.

- ◉ True ✓ true since it is the sum of squared quantities
- ○ False

4) Select all formulations of the optimization problem that are equivalent to the one given below:

$\min_{\mathbf{w},||\mathbf{w}||=1} \frac{1}{n}\sum_{i=1}^{n} -(\mathbf{x}_i^T\mathbf{w})^2$ = $\max \frac{1}{n}\sum_{i=1}^{n} (x_i^T w)^2$

- ☑ $\max_{\mathbf{w},||\mathbf{w}||=1} \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{w})^2$ or $\max\ w^T \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) w$
- ☑ $\max_{\mathbf{w},||\mathbf{w}||=1} \mathbf{w}^T \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right] \mathbf{w}$
- ☐ $\min_{\mathbf{w},||\mathbf{w}||=1} \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{w})^2$

5) Given a dataset of n points in $\mathbb{R}^d$, what is the dimension of the covariance matrix? *1 point*

- ○ $n \times n$
- ◉ $d \times d$
- ○ $n \times d$
- ○ $d \times n$

n pts in $\mathbb{R}^d$

$C \Rightarrow d \times d$

$C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

6) What is the optimal value of $\mathbf{w}$ for the optimization problem posed towards the end of the lecture? *1 point*

- ○ It is the eigenvector corresponding to the minimum eigenvalue of the covariance matrix.
- ○ It could be any one of the eigenvectors of the covariance matrix. That is, every eigenvector is a solution of the optimization problem.
- ◉ It is the eigenvector corresponding to the maximum eigenvalue of the covariance matrix.

$x \in \mathbb{R}^d$

↓ Find w

$(x^T w) \cdot w$

↓ Residue / error

$x - (x^T w) \cdot w$

Might not be error but has "information"

POSSIBLE ALGORITHM

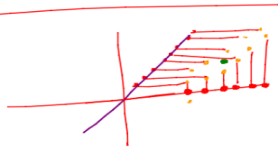
Input: $\{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$

→ $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ $x_i = x_i - \mu$ $\forall i$

→ Find "best" line $w_1 \in \mathbb{R}^d$

→ Replace $x_i \leftarrow x_i - (x_i^T w)\, w$

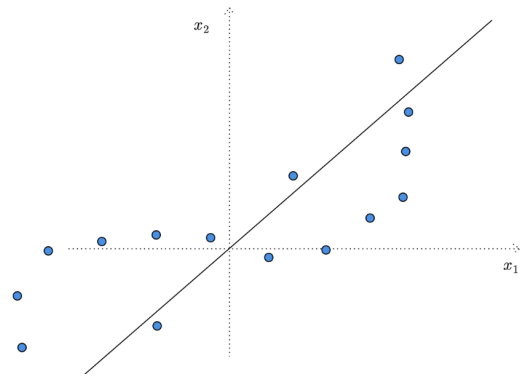
→ Repeat to obtain w_2



Issue: Data may not be centered

Activity Question - 5

1) Consider the following figure:



The line shown in the figure is obtained after one round of running the "possible algorithm". Does a second round seem necessary here?

- (●) Yes — Yes, becoz. in another round, over reconstruction error may reduce.
- ( ) No

2) Which of the following represents the expression for the residue of a point $\mathbf{x}$ after projecting it onto a line passing through the origin, represented by $\mathbf{w}$? Note that $||\mathbf{w}|| = 1$. *1 point*

- ( ) $(\mathbf{x}^T\mathbf{w})\mathbf{w}$
- ( ) $\mathbf{x} - \mathbf{w}$
- ( ) $\mathbf{x} - \mathbf{x}^T\mathbf{w}$
- (●) $\mathbf{x} - (\mathbf{x}^T\mathbf{w})\mathbf{w}$ ✓

Residue $= \left(x_i - \frac{(x^T w)w}{\|w\|^2} \right)$ ← but, $\|w\| = 1$ then,

3) We are given a dataset of n points in $\mathbb{R}^d$ and a line passing through the origin that is represented by $\mathbf{w}$.

S1: The residue corresponding to every point is perpendicular to $\mathbf{w}$. ✓

S2: All the residues are collinear, that is, they lie on a line (passing through the origin).

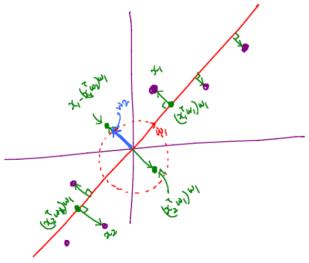
- (●) Only S1 is true — ↳ this is not necessary, they might also be non-linear
- ( ) Only S2 is true
- ( ) Both S1 and S2 are true
- ( ) Both S1 and S2 are false

4) Is the following dataset centered?

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

- (●) Yes ✓ — Yes, Calculate the mean of the data set $\bar{x}$ then $\sum_{i=1}^{n} (\bar{x} - x_i) = 0$, then the dataset is centered.
- ( ) No

Question: What can we say about w_1 & w_2?



$\Rightarrow$ $w_2^T w_1 = 0$

Observation:

→ All residues are orthogonal to w_1

→ Any line which minimizes sum of errors wrt residues must also be orthogonal to w_1 [ARGUE WHY]

$\forall i \quad x_i = (x_i^T w_1) w_1 + (x_i^T w_2) w_2 + \dots$

Eg. $\{w_1, w_2, w_3\} \in \mathbb{R}^{100}$

Rep $\{w_1, w_2, w_3\}$ ↳ Common for dataset

Co-efficients $x_i \to [x_i^T w_1 \quad x_i^T w_2 \quad x_i^T w_3] \in \mathbb{R}^3$ ↳ Data point specific

original: $100 \times n$, $100 \times 100 = 10000$, dxn

now: $3 \times 100 + 3n$, $3 \times 100 + 3 \times 100 = 600$, $dxk + kxn$

What have we gained?

- If data lives in a "low" dimensional linear sub-space, then residues become 0 much earlier than d rounds.

" Larger the value of $\frac{1}{n} \sum_{i=1}^{n} (x_i^T w)^2$, the better the fit"

ENTER LINEAR ALGEBRA

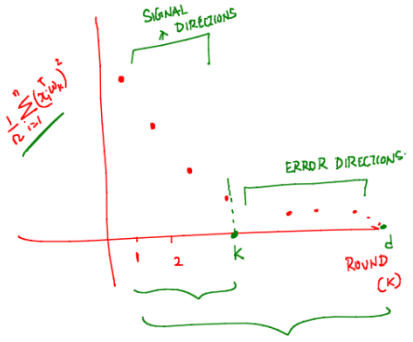
$\max_{w: \|w\|_2^2=1} w^T C w$ $\quad C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

C → Covariance matrix.

Soln: w_1 is eigenvector corresponding to the "largest" eigenvalue of C. [HILBERT'S min-max THEOREM]

In fact $\{w_1, \cdots, w_d\}$, the eigenvectors of C form an orthonormal basis.

w_k → Best line one can obtain in round k



$\lambda_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i^T w_1)^2$ ← term we used earlier

Average? $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^T w = 0^T w$ [for a centered dataset]

$= 0$

Variance $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w - \underbrace{\text{mean}}_{=0})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$

Rule of thumb for # dimensions

$\left(\sum_{i=1}^{k}\lambda_i(C) \Big/ \sum_{i=1}^{d}\lambda_i(C)\right) \geq 0.95$ → usually in practice

ERROR MINIMIZATION on CENTERED DATASET $\iff$ VARIANCE MAXIMIZATION

PCA Finds combination of features [linear] that are de-correlated. [loosely speaking independent of each other]

"EIGEN FACES"

Activity Question - 6

1) $\mathbf{w}_1$ is the line which minimizes the reconstruction error of a set of points and $\mathbf{w}_2$ is the line which minimizes the reconstructic error of the residues. What is the value of $\mathbf{w}_1^T\mathbf{w}_2$?

0 $\quad w_1^T w_2 = w_2^T w_1 = 0$

1 poi

2) $B = \{\mathbf{w}_1, \cdots, \mathbf{w}_d\}$ is a set of d orthonormal vectors in $\mathbb{R}^d$ that are obtained by running the algorithm for d rounds on some centered dataset. Select all correct statements. **1 poi**

- [x] The vectors in B are linearly independent ✓
- [x] The vectors in B span $\mathbb{R}^d$ } → independence = span + basis
- [x] B is a basis for $\mathbb{R}^d$ ✓
- [x] $\mathbf{w}_i^T\mathbf{w}_j = 0$ for $i \neq j$ ✓ } direct constraints
- [x] $\mathbf{w}_i^T\mathbf{w}_i = 1$ for $1 \leq i \leq d$ ✓

3) Consider a dataset of 1000 points, each of which belongs to $\mathbb{R}^5$. If you know nothing else about the dataset to begin with, what is the number of rounds of the algorithm after which the residues completely vanish? → $\mathbb{R}^d$

5 $\quad$ round $\leq d$ $\quad \therefore$ at max 5 rounds

1 point

4) Consider a dataset of 1000 points, each of which belongs to $\mathbb{R}^{10}$. Upon running the algorithm on this dataset, the residues completely vanish after 4 rounds. Which of the following statements is true? **1 point**

- [ ] The dataset is a subset of $\mathbb{R}^4$
- [ ] The dataset is a subset of a 4-dimensional linear subspace of $\mathbb{R}^{1000}$
- [x] The dataset is a subset of a 4-dimensional linear subspace of $\mathbb{R}^{10}$

$n = 1000$, $d = 10$, $k = 4$

k-dimensional of $\mathbb{R}^d$

Activity Question - 7

1) What is the best line one can obtain in the k^{th} round of the algorithm?

- [ ] It is the eigenvector corresponding to the k^{th} smallest eigenvalue of the covariance matrix.
- [x] It is the eigenvector corresponding to the k^{th} largest eigenvalue of the covariance matrix. → Hilbert min-max theorem

2) What is the value of the following expression if λ_k is an eigenvalue of the covariance matrix $\mathbf{C}$ for eigenvector $\mathbf{w}_k$?

$\mathbf{w}_k^T\mathbf{C}\mathbf{w}_k$ $\quad \lambda_k = w_k^T C w_k$

- [ ] 1
- [ ] 0
- [x] λ_k ✓
- [ ] λ_k^2

3) Given a dataset of n points, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, the k^{th} eigenvector and the corresponding eigenvalue of the covariance matrix are $\mathbf{w}_k$ and λ_k respectively. Consider the following statements:

S1

$\lambda_k = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w}_k)^2 = w_k^T C w_k$

S2

λ_k is non-negative → Sum of squared things

- [ ] Only S1 is true
- [ ] Only S2 is true
- [x] Both S1 and S2 are true ✓
- [ ] Both S1 and S2 are false

4) Given a centered dataset of n points, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, and a line $\mathbf{w}$ that passes through the origin, consider the following set of quantities:

$\{\mathbf{x}_1^T\mathbf{w}, \cdots, \mathbf{x}_n^T\mathbf{w}\}$

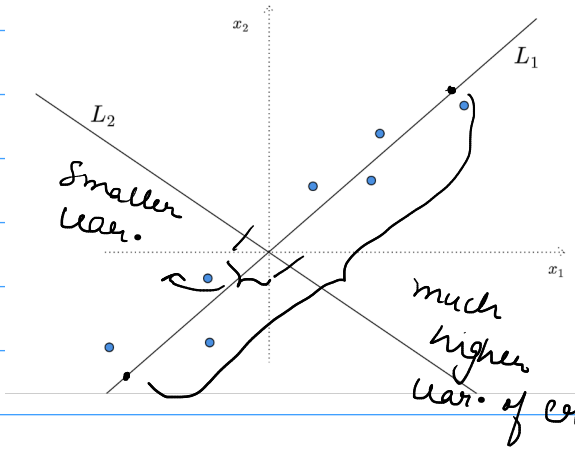
What is the average of the quantities in the above set?

0

Avg. = 0

Var. = $\frac{1}{n}\sum_{i=1}^{n}(x_i^T w)^2$

5) Consider the following dataset and two lines L_1 and L_2:



Which of the two lines is a good direction for a compressed representation of the dataset and why?

- ○ L_1, as it gives a better compression ratio
- ○ L_2 as it gives a better compression ratio
- ◉ L_1 as the variance of the dataset is higher along this direction
- ○ L_2 as the variance of the dataset is higher along this direction
- ○ L_1 as the variance of the dataset is lower along this direction
- ○ L_2 as the variance of the dataset is lower along this direction

Practice Assignment - 1

1) An image is a collection of pixels. A pixel is stored as a float value and typically occupies 4 bytes of memory. Consider a ***1 point*** dataset of 1000 images, where each image has dimensions 100×100. Approximately, how much memory does the entire dataset occupy?

- ○ 4 KB
- ○ 4 MB
- ◉ 40 MB ✓
- ○ 4 GB

1 img. = 100 x 100 = 10000 ⇒ 4 bytes – 1 img.

1000 ⇒ 4000 bytes x 10000 (pen)

→ 4,00,0 0,000 bytes

⇒ 40000 kB

⇒ 40 MB

2) Consider a dataset that has 100 points that belong to $\mathbb{R}^3$. All of them are found to lie on a line that passes through the origin. We use a unit vector along the line as a representative and the coefficients with respect to it to represent the individual data-points. Compute the percentage decrease in the size of the dataset if we move to this new representation. Assume that it takes one unit of space to store one feature. Enter your answer correct to two decimal places; it should be in the range $[0, 100]$.

65.66

Original ⇒ d x n = 300

later ⇒ d + k = 103

$\frac{300-103}{300} = \frac{197}{300} \times 100 = 65.67\%$

3) **Common Data for questions (3) and (4)**

Consider the following dataset that has four points, all of which lie on a line:

$S = \left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3\\4\end{bmatrix}, \begin{bmatrix}6\\8\end{bmatrix}, \begin{bmatrix}1/5\\4/15\end{bmatrix}\right\}$

Answer questions (3), (4) and (5) based on this data. The statement of questions-(3) is given below the horizontal line.

Among the vectors given below, choose a representative that has unit length.

- ○ $\begin{bmatrix}3\\4\end{bmatrix}$ $(i^2 + j^2)^{1/2} = 1$
- ○ $\begin{bmatrix}1/15\\4/15\end{bmatrix}$
- ◉ $\begin{bmatrix}3/5\\4/5\end{bmatrix}$ ✓ $\left(\frac{3}{5}\right)^2 + \left(\frac{4}{5}\right)^2 = \frac{9}{25} + \frac{16}{25} = \frac{25}{25}$ $= \sqrt{1} = 1$
- ○ $\begin{bmatrix}1\\1\end{bmatrix}$

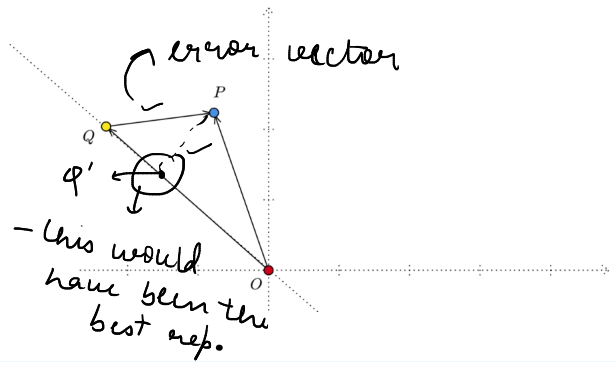
4) With respect to the representative in the previous question, compute the coefficients for these four points. The i^{th} element from the left in each option is the coefficient for the i^{th} element from the left in the set S.

- ◉ $\{0, \quad 5, \quad 10, \quad 1/3\}$
- ○ $\{0, \quad 1, \quad 2, \quad 1/3\}$
- ○ $\{0, \quad 5, \quad 10, \quad 3\}$
- ○ $\{0, \quad 1, \quad 10, \quad 1/3\}$

$\begin{bmatrix}3/5\\4/5\end{bmatrix}$ $(0, 5, 10, 1/3) = \begin{bmatrix}3/5 & 3 & 6 & 3/15\\4/5 & 4 & 8 & 4/15\end{bmatrix}$

5) **Common Data for questions (5) to (7)** *1 point*

Consider the following image. P is a point in 2D space. Q is a proxy for this point on a line passing through the origin. The image is drawn to scale.



error vector

Answer questions (5), (6) and (7) that follow. The statement of question-(5) is given below the horizontal line.

Which of the following is the error vector?

- $\overrightarrow{OP}$
- $\overrightarrow{OQ}$
- (●) $\overrightarrow{QP}$

6) Is Q the "best" representation of P on the line?

- Yes
- (●) No ✓

this would have been the best rep.

7) If Q' is the "best" representation of P on the line, then which of the following statements are true? Notation: $|\overrightarrow{AB}|$ is the length of the vector $\overrightarrow{AB}$. *1 point*

- ☐ $|\overrightarrow{QP}| < |\overrightarrow{Q'P}|$ $\quad |Q'P| < QP$ (clearly)
- ☑ $|\overrightarrow{Q'P}| < |\overrightarrow{QP}|$ ✓
- ☐ $|\overrightarrow{QQ'}| = 0$ ✗ since $Q \neq Q'$
- ☑ $|\overrightarrow{QQ'}| > 0$ ✓

8) Is the following statement true or false? *1 point*

The projection of $\mathbf{x}$ onto $\mathbf{w}$ was derived to be $(\mathbf{x}^T\mathbf{w})\mathbf{w}$, where $\mathbf{w}$ is a unit vector. Since this derivation was done for the special case of 2D vectors, this formula is not applicable in the general case of d-dimensional vectors.

- ( ) True
- (●) False

no, not this formula is applicable. $x^T y = \sum_i^d x_i y_i$

9) Consider a mean-centered dataset of n points where each point belongs to $\mathbb{R}^d$. $\mathbf{w_1}, \cdots, \mathbf{w_k}$ are the first k principal components obtained by running PCA on the dataset, where $k < d$. The following relationship is observed:

$$\mathbf{x}_i - \left[\sum_{j=1}^{k} (\mathbf{x}_i^T \mathbf{w}_j)\mathbf{w}_j\right] = 0, \quad 1 \leq i \leq n$$

Which of the following statement about the dataset is true?

- The dataset lies in a d-dimensional subsapce of $\mathbb{R}^n$
- (●) The dataset lies in a k-dimensional subspace of $\mathbb{R}^d$
- The dataset lies in a d-dimensional subspace of $\mathbb{R}^k$
- The dataset lies in a k-dimensional subspace of $\mathbb{R}^n$

remember above eq. in AQ

4-dim subspace in R10

10) In the context of PCA, given n data-points in $\mathbb{R}^d$ that are mean-centered, after estimating $\mathbf{w_1}$ in the first round, what is the mean of the residues?

- $\mathbf{w}_1$
- (●) $\mathbf{0}$

Avg. / Mean = 0

11) Consider two ways of representing n datapoints that belong to $\mathbb{R}^d$ in the form of a matrix:

Approach-1: A matrix $\mathbf{X_1}$ of dimension $n \times d$

Approach-2: A matrix $\mathbf{X_2}$ of dimension $d \times n$

Assume that the dataset is mean-centered. Select all correct expressions for the covariance matrix.

- ☑ $\frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1$ ✓ $n \times d \Rightarrow X_1^T X_1 = n \cdot C \Rightarrow C = \frac{1}{n} X_1^T X_1$
- ☐ $\frac{1}{n}\mathbf{X}_2^T\mathbf{X}_2$
- ☐ $\frac{1}{n}\mathbf{X}_1\mathbf{X}_1^T$ $\quad d \times n \Rightarrow X_2 X_2^T = n \cdot C \Rightarrow C = \frac{1}{n} X_2 X_2^T$
- ☑ $\frac{1}{n}\mathbf{X}_2\mathbf{X}_2^T$ ✓

12) Consider a mean-centered dataset obtained from the banking domain that has 100 data-points, each of which is described by 7 features. The dataset is represented as a 100×7 matrix, $\mathbf{X}$. You run PCA on this dataset and observe that the residues vanish completely after k iterations.

A little later, a domain expert makes the following observations. If $\mathbf{c}_i$ represents the i^{th} column of $\mathbf{X}$, then:

(1) The set of vectors $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ are linearly independent.
(2) The following relations are satisfied:
(a) $\mathbf{c}_5 = \mathbf{c}_1 + \mathbf{c}_3$
(b) $\mathbf{c}_6 = 2\mathbf{c}_3 - 3\mathbf{c}_4$
(c) $\mathbf{c}_7 = \mathbf{c}_2 + 3\mathbf{c}_4$

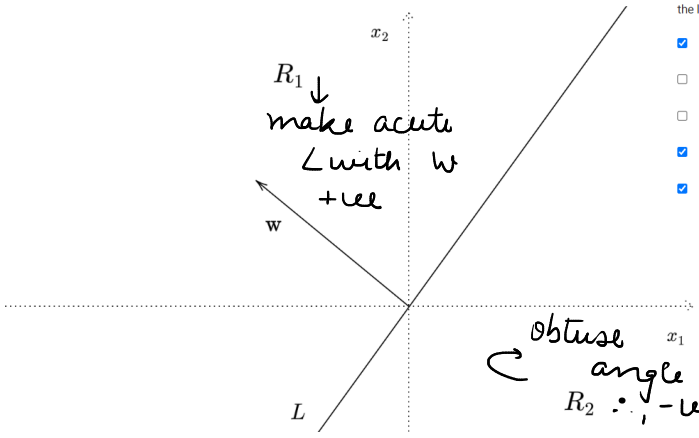
7 features
4 independent $c_1 \to c_4$
but c_5, c_6 & c_7 are dep. on 1st 4 features.

What is the value of k? Assume that the dataset is already mean-centered.

$\therefore, k = 4$

4

13) Consider the following image:



Here, $\mathbf{w}$ is a vector and L is a line perpendicular to $\mathbf{w}$ that passes through the origin. R_1 and R_2 are two regions on either side of the line L. If $\mathbf{x}$ is an arbitrary vector in the plane, select all correct statements.

- [x] $R_1 : \mathbf{w}^T\mathbf{x} > 0$
- [ ] $R_2 : \mathbf{w}^T\mathbf{x} > 0$
- [ ] $R_1 : \mathbf{w}^T\mathbf{x} < 0$
- [x] $R_2 : \mathbf{w}^T\mathbf{x} < 0$
- [x] $L : \mathbf{w}^T\mathbf{x} = 0$

Graded Assignment - 1

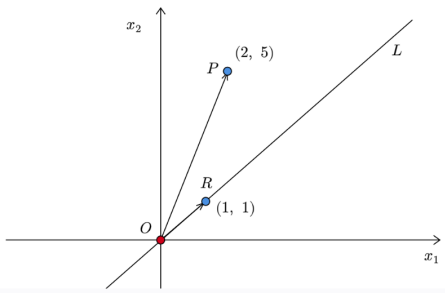
1) Consider a point $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and a line passing through the origin which is represented by the vector $\mathbf{w} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$. What can you say about the following quantities? (MSQ)

(1) the projection of $\mathbf{x}$ onto the line

(2) the residue

- [ ] The residue is equal to the zero vector.
- [x] The residue is equal to the vector $\mathbf{x}$.
- [x] The projection is the zero vector.
- [ ] The projection is equal to the vector $\mathbf{x}$.

Consider a point P and a line L that passes through the origin O. The point R lies on the line.



We use the following notation:

$\mathbf{w} = \overrightarrow{OR}$

$\mathbf{x} = \overrightarrow{OP}$

2) Consider the following statements:

Statement-1: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{w}$

Statement-2: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{w})\mathbf{x}$

Statement-3: The projection of $\mathbf{x}$ on the line L is given by $(\mathbf{x}^T\mathbf{x})\mathbf{w}$

Statement-4: The projection of $\mathbf{x}$ on the line L is given by $\mathbf{w}^T\mathbf{x}$

Which of the above statements is true?

- Statement-1
- Statement-2
- Statement-3
- Statement-4
- None of these statements are true.

3) Find the length of the projection of $\mathbf{x}$ on the line L. Enter your answer correct to two decimal place

4.94

4) Find the residue after projecting $\mathbf{x}$ on the line L.

- $\begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix}$
- $\begin{bmatrix} -1.5 \\ 1.5 \end{bmatrix}$
- $\begin{bmatrix} 2 \\ 5 \end{bmatrix}$
- $\begin{bmatrix} 1 \\ 4 \end{bmatrix}$

5) Find the reconstruction error for this point. Enter your answer correct to two decimal places.

4.50

6) Consider the following images of points in 2D space. The red line segments in one of the images represent the lengths of the residues after projecting the points on the line L. Which image is it?

Image-1

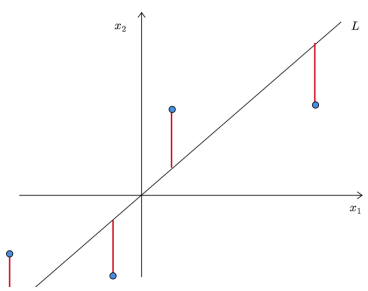
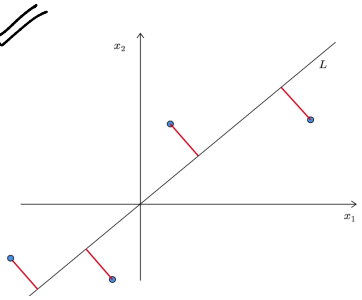


Image-2



7) Consider a dataset that has 1000 samples, where each sample belongs to $\mathbb{R}^{30}$. PCA is run on this dataset and the top 4 principal components are retained, the rest being discarded. If it takes one unit of memory to store a real number, find the percentage decrease in storage space of the dataset by moving to its compressed representation. Enter your answer correct to two decimal places; it should lie in the range $[0, 100]$.

86.26

$n = 1000$, $d = 30$, $k = 4$, % decrease = ?

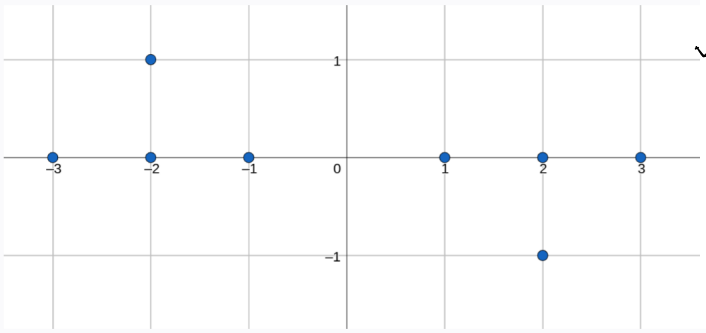
Original $= n \times d = 30{,}000$

Later $= n \times k + d \times k = 4000 + 120 = 4120$

$\% = \frac{30000 - 4120}{30000} \times 100 = 86.27\%$

Common Data for questions (8) to (9)

Consider a dataset that has 8 points all of which belong to $\mathbb{R}^2$:



$2 \times 8 \quad 8 \times 2 = 2 \times 2$

8) Find the covariance matrix of this dataset.

- ✓ $\begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$
- $\begin{bmatrix} 36 & -4 \\ -4 & 2 \end{bmatrix}$
- $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

$C = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$

$n = 8$

$d = 2$

$\begin{bmatrix} -3 & -2 & -1 & -2 & 1 & 2 & 3 & 2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}_{2 \times 8} \times \begin{bmatrix} -3 & 0 \\ -2 & 0 \\ -1 & 0 \\ -2 & 1 \\ 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 2 & 1 \end{bmatrix}_{8 \times 2}$

$C = \begin{bmatrix} 4.5 & -0.5 \\ -0.5 & 0.25 \end{bmatrix}$

9) If PCA is run on this dataset, find the variance of the dataset along the first principal component. The eigenvectors of the covariance matrix are given below:

$\begin{bmatrix} -0.993 \\ 0.115 \end{bmatrix}, \begin{bmatrix} -0.115 \\ -0.993 \end{bmatrix}$

Recall that the first principal component is the most important.

4.55

if (λ_k, W_k) is the k^{th} eigenpair, then

$\lambda_k = W_k C W_k^T$

$\sigma_j^2 = \lambda_j \Rightarrow W_1^T C W_1$

$\Rightarrow \begin{bmatrix} -0.115 & -0.993 \end{bmatrix} \begin{bmatrix} 4.5 & 0.5 \\ -0.5 & 0.25 \end{bmatrix} \begin{bmatrix} -0.115 \\ -0.993 \end{bmatrix}$

$= 4.55$

10) Consider a dataset of 100 points all of which lie in $\mathbb{R}^5$. The eigenvalues of the covariance matrix are given below:

$3.4, \quad 2.8, \quad 0.5, \quad 0.4, \quad 0.01$

If we run the PCA algorithm on this dataset and retain the top-k principal components, what is a good choice of k? Use the heuristic that was discussed in the lectures.

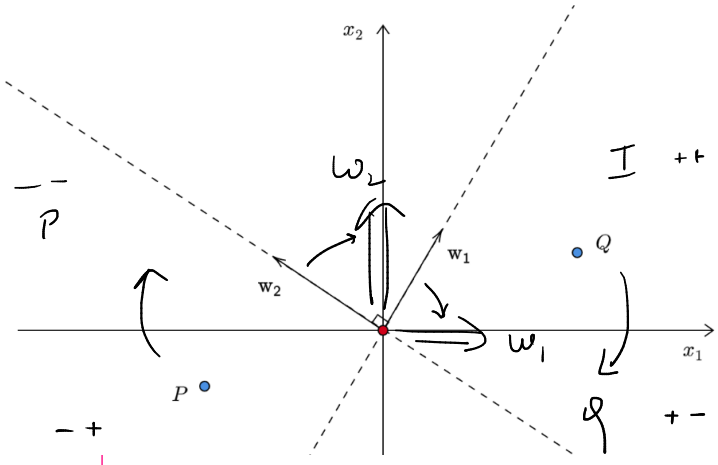
4 ✓

top k comp. should capture 95% ↓

$3.4 + 2.8 + 0.5 + 0.4 + 0.1 = 7.2$

47.2% / 86.1% / 93.5% / 98.6% ✓

11) PCA is run on a dataset that has 2 features. The resulting principal components are $\mathbf{w}_1$ and $\mathbf{w}_2$. We represent the points in 2D space in terms of this new coordinate system made up of the principal components. The first coordinate corresponds to $\mathbf{w}_1$ and the second to $\mathbf{w}_2$. In such a scenario, what would be the sign of the coordinates for the points P and Q? ***1 point***



- $P : (-ve, -ve)$
- ✓ $P : (-ve, +ve)$
- $Q : (+ve, +ve)$
- ✓ $Q : (+ve, -ve)$

MLT - Week 2

- With PCA, C => R d*d, so time complexity is O(d^3). Problem arises when d is large. Face recog.
- It's not necessary that the data lives in a low-dimensional linear subspace.

Activity Question -1

1) Which principal component captures the most variance in the dataset?

OR

Along which principal component do the projected data points have the highest variance?

- (•) First component
- ( ) The last component along which data points have non-zero projections.
- ( ) d^{th} component where d is the number of features.
- ( ) Can not be determined.

2) What will be the time complexity of running PCA using covariance matrix on a dataset containing n examples in d-dimensional space?

- ( ) $O(n^2)$
- ( ) $O(d^2)$
- ( ) $O(n^3)$
- (•) $O(d^3)$

$C \in R^{d \times d}$ → $O(d^3)$

3) Which of the following may be the issues with PCA? *1 p*

- [x] When the number of features becomes large, the time complexity of PCA goes high. $O(d^3)$ Issue 1
- [ ] When the number of examples becomes large, the time complexity of PCA goes high.
- [x] Features in the dataset have non-linear relationships. → Issue 2
- [ ] Features in the dataset have linear relationships. → not a low-d linear subspace

4) PCA algorithm is run on the Yale face dataset discussed in the lecture to get the principal components. Now, these principal components were used to represent a dog image. Which of the following number of principal components should be ideal for a better representation? *1 p*

- ( ) 2
- ( ) 10
- ( ) 20
- (•) 200

From lecture value can out as 200

$C = \frac{1}{n} XX^T$

Let w_k be the eigenvector corresponding to the k^{th} largest eigenvalue of C (λ_k ← e-value)

$C w_k = \lambda_k w_k$ [by definition]

$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) w_k = \lambda_k w_k$

$w_k = \sum_{i=1}^{n} \left(\frac{x_i^T w_k}{n\lambda_k}\right) \cdot x_i$

$w_k = X\alpha_k$ — w_k is a LINEAR COMBINATION of data points!

$K\alpha_k = (n\lambda_k)\alpha_k$ ←

$K\beta_k = (n\lambda_k)\beta_k$

$1 = \alpha_k^T K \alpha_k$

$O(n^3)$

LINEAR ALGEBRA FACT : The non-zero eigenvalues of XX^T ($\in \mathbb{R}^{d \times d}$) and X^TX ($\in \mathbb{R}^{n \times n}$) are exactly the same!

Activity Question - 2

1) The k^{th} principal component w_k can be written as a linear combination of data points that is $w_k = X\alpha_k$, where X is the data matrix of shape $d \times n$ and $\alpha_k \in \mathbb{R}^n$. Let $K = X^TX$. Which of the following expressions must be satisfied by the α_k? ***1 point***

- ( ) $\alpha_k^T\alpha_k = 1$
- (•) $\alpha_k^T K\alpha_k = 1$ → directly from lecture
- ( ) $\alpha_k^T\alpha_k = KK^T$
- ( ) $\alpha_k\alpha_k^T = K^TK$

2) Consider a matrix X of shape $d \times n$ containing n examples belonging to d-dimensional space with $d > n$. Which of the following statements is/are correct? → all the non-0 eigenvalues are same.

Statement 1: All the eigenvalues of XX^T and X^TX are the same.

Statement 2: In general, the number of eigenvalues of XX^T is always less than the number of eigenvalues of X^TX.

- ( ) Only statement 1 is correct.
- ( ) Only statement 2 is correct.
- ( ) Both statements are correct.
- (•) Both statements are incorrect.

3) What is the time complexity of computing Eigen decomposition of $K_{n\times n} = X^TX$?

- ( ) $O(n^2)$
- ( ) $O(d^2)$
- (•) $O(n^3)$
- ( ) $O(d^3)$

$O(n^3)$ for $K_{n \times n}$ which better than $O(d^3)$ for $d >> n$

4) Let X be the data matrix of shape $d \times n$ with $d > n$. The eigenvector corresponding to the largest eigenvalue λ of X^TX is α_1. What will be the first principal component of the dataset?

- ( ) α_1
- ( ) $\frac{\alpha_1}{\sqrt{\lambda}}$
- ( ) $X\alpha_1$
- (•) $\frac{X\alpha_1}{\sqrt{\lambda}}$

$\alpha_1 = \frac{\beta_1}{\sqrt{\lambda}}$

$w_1 = X\alpha_1 = \frac{X\beta_1}{\sqrt{\lambda}}$

$$R(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad \text{for some } \sigma > 0$$

↑ RADIAL BASIS FUNCTION.

$x \in \mathbb{R}^2 \xrightarrow{\phi} \phi(x)$

$[f_1 \;\; f_2] \rightarrow [1 \;\; f_1^2 \;\; f_2^2 \;\; f_1 f_2 \;\; f_1 \;\; f_2]$

$\mathbb{R}^2 \qquad \mathbb{R}^6$

→ Can be shown to be a "valid" map.

→ Interestingly, ϕ in this case maps x to an "infinite" dimensional space.

[Technicalities aside, can think of this as mapping a data point to a "function" and dot-products between functions become integrals.]

d features, $\leq p^{th}$ power

$$\sum_{i=0}^{p} {}^{d}C_i \approx O(d^p)$$

Any function $R: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ which is a "valid" map is called a kernel function

$R(x,x') = (x^T x' + 1)^p$ → POLYNOMIAL KERNEL

$R(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ → RADIAL BASIS / GAUSSIAN KERNEL

2 things to check for validity of a kernel.

(a) R is symmetric i.e., $R(x,x') = R(x',x)$.

(b) For any dataset $\{x_1, \ldots, x_n\}$, the matrix $K \in \mathbb{R}^{n \times n}$ where $K_{ij} = R(x_i, x_j)$ is POSITIVE SEMI DEFINITE

↳ eigen values of K are non-negative

KERNEL PCA

- Input - $\{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$, Kernel $R: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$.
- Step 1: Compute $K \in \mathbb{R}^{n \times n}$ where $K_{ij} = R(x_i, x_j)$ $\forall i,j$
  → "Centre the kernel"
- Step 2: Compute $\beta_1, \ldots, \beta_p$ eigen vectors & $n\lambda_1 \geq \cdots n\lambda_p$ eigen values of K. and normalize to get $\forall u$ $\alpha_u = \frac{\beta_u}{\sqrt{n\lambda_u}}$

Modified Step 3: Compute $\sum_{j=1}^{n} \alpha_{kj} K_{ij}$ $\forall k$

$x_i \in \mathbb{R}^d \rightarrow \left[\sum_{j=1}^{n} \alpha_{1j} K_{ij}, \sum_{j=1}^{n} \alpha_{2j} K_{ij}, \ldots, \sum_{j=1}^{n} \alpha_{kj} K_{ij}\right]$

Create a new kernel K^c ← centered

$$K^c_{ij} = K_{ij} - \theta_i 1_j^T - 1_i \theta_j^T + P\, 1_i 1_j^T$$

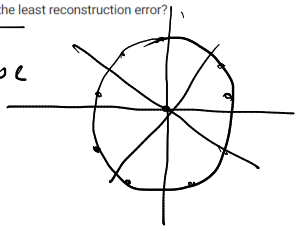
where $\theta_i = \frac{1}{n}\sum_{k=1}^{n} K_{ik}$ $\forall i$, $P = \frac{1}{n^2}\sum_{ij} K_{ij}$

Activity Question - 3

1) Consider 100 data points lying uniformly (distances between neighbors are the same) on a circle with the center at the origin. Along which of the following lines, the data points will have the least reconstruction error? **1 point**

- x-axis
- y-axis
- $y = x$
- Can not be determined using the given information.

This can't be determined



2) Consider the data points x_i ; $i = 1, 2, \ldots n$ lying on a circle given by $(f_1 - a)^2 + (f_2 - b)^2 = r^2$, where a, b and r are the integers and f_1, f_2 are the two features of the data points. Consider a point $u = [a^2 + b^2 - r^2, 1, 1, 0, -2a, -2b] \in \mathbb{R}^6$. Which of the following transformations ϕ will result in $\phi(x_i)^T u = 0$ for all i? **1 point**

- $[f_1, f_2] \rightarrow [1, f_1^2, f_2^2, f_1, f_2, f_1 f_2]$
- $[f_1, f_2] \rightarrow [f_1^2, f_2^2, f_1 f_2, f_1, f_2, 1]$
- $[f_1, f_2] \rightarrow [1, f_1^2, f_2^2, f_1 f_2, f_1, f_2]$

$r^2 = f_1^2 + a^2 - 2f_1 a + f_2^2 + b^2 - 2f_2 b$

$f_1^2 + f_2^2 - (a^2 + b^2 - r^2) - 2f_1 a - 2f_2 b = 0$

$f_1 = 1, \; f_2 = 1$

3) Consider the data points x_i ; $i = 1, 2, \ldots n$ lying on a curve given by $4(f_1 - a) - (f_2 - b)^2 = 0$, where a and b are the integers. Consider a point $u = [-4a - b^2, -1, 4, 2b] \in \mathbb{R}^4$. Which of the following transformations ϕ will result in $\phi(x_i)^T u = 0$ for all i? **1 poi**

- $[f_1, f_2] \rightarrow [1, f_2^2, f_1, f_2]$
- $[f_1, f_2] \rightarrow [1, f_1^2, f_2^2, f_1]$
- $[f_1, f_2] \rightarrow [f_1^2, f_2^2, f_1 f_2, f_1]$
- $[f_1, f_2] \rightarrow [1, f_2^2, f_1^2, f_2]$

$4f_1 - 4a - f_2^2 - b^2 + 2f_2 b = 0$

$-4a - b^2, -1, 4, 2f_2 b = 0$

4) If we transform the d-dimensional feature space to a higher D-dimensional space, which of the following issues may arise with the standard PCA on transformed data points? **1 poi**

- D may become larger than the number of data points.
- D may be less than the number of data points.

5) Assume that the data points lying in three-dimensional space (3 features) have been transformed into the higher dimensional space so that it captures all the cubic relationships between all the three features. What will be the dimension of transformed space?

↳ $q = 4$

dim. of transformed space $= {}^{n+q-1}C_{q-1}$

$n = 3$

$q = 4$

$\{0, 1, 2, 3\}$

$= {}^{6}C_3 = \frac{6!}{3!\,3!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 3 \times 2 \times 1}$

$= 20$

Activity Question - 4

1) let ϕ be the transformation that maps n d-dimensional data points to D-dimensional data points. What will the covariance matrix shape be in the transformed feature space?

- $n \times n$
- $d \times d$
- [x] $D \times D$
- $n \times D$

$C \in d \times d \Rightarrow D \times D$

2) What is the ij^{th} element of the matrix K in the transformed space if the transformation mapping is ϕ?

- $x_i^T x_j$
- [x] $\phi(x_i)^T \phi(x_j)$
- $\phi(x_i)\phi(x_j)^T$
- $\phi(x_i)\phi(x_j)$

$K_{ij} = \phi(x_i)^T \phi(x_j)$ kernel func^n

3) A transformation ϕ is given by

$\phi : \mathbb{R}^2 \to \mathbb{R}^6$

$\phi([a,b]^T) = [a^2,\ b^2,\ 1, \sqrt{2}ab,\ \sqrt{2}a,\ \sqrt{2}b]^T$

Which of the following expressions will be the same as $\phi(x_1)^T \phi(x_2)$, where $x_1, x_2 \in \mathbb{R}^2$?

- $(x_1^T x_2)^2 + 1$
- $(x_1^T x_2) + 1$
- $(x_1 x_2 + 1)^2$
- [x] $(x_1^T x_2 + 1)^2$

$x = [f_1, f_2]$, $x' = [g_1, g_2]$

$(x^T x' + 1)^2 = [f_1^2,\ f_2^2,\ 1,\ \sqrt{2} f_1 f_2,\ \sqrt{2} f_1,\ \sqrt{2} f_2]\,[g_1^2,\ g_2^2,\ 1,\ \sqrt{2} g_1 g_2,\ \sqrt{2} g_1,\ \sqrt{2} g_2]^T$

$= \phi(x)^T \phi(x')$

$x \to x_1$, $x' \to x_2$

4) How would you argue that the kernel function should be symmetric?

- [x] A kernel function is a dot product of two vectors and the dot product is commutative.
- A kernel function is a dot product of two vectors and the dot product is Associative.
- Non-zero eigenvalues of the matrices $\phi(X)^T\phi(X)$ and $\phi(X)\phi(X)^T$ are always the same.
- A function must always be symmetric.

an eg. can be seen as above

5) $\exists$ a function $\phi : \mathbb{R} \to \mathbb{R}$ such that $k(x_1, x_2) = \phi(x_1^T x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$

- $\exists$ a function $\phi : \mathbb{R} \to \mathbb{R}$ such that $k(x_1, x_2) = \phi(x_1^T x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$
- $\exists$ a function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $k(x_1, x_2) = \phi(x_1 x_2^T)$ for all $x_1, x_2 \in \mathbb{R}^d$
- [x] $\exists$ a function $\phi : \mathbb{R}^d \to \mathbb{R}^D$ such that $k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$
- $\exists$ a function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $k(x_1, x_2) = \phi(x_1) + \phi(x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$

$\phi(x_1^T x_2) = \phi(x_1)^T \phi(x_2)$

6) What are the necessary and sufficient conditions for a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to be a valid kernel?

- [x] k is symmetric that is $k(x_i, x_j) = k(x_j, x_i)$.
- [ ] k is a constant function that is $k(x_i, x_j) = c$ for some constant c.
- [ ] The matrix $K \in \mathbb{R}^n \times \mathbb{R}^n$, where $k_{ij} = k(x_i, x_j)$, is positive definite.
- [x] The matrix $K \in \mathbb{R}^n \times \mathbb{R}^n$, where $k_{ij} = k(x_i, x_j)$, is positive semi-definite.

① sym.
② +ve semi def.

7) The dimension of the transformed feature vectors ϕ whose dot product the kernel function computes, can be infinite.

- [x] True
- False

as in the case of radial basis func^n / gaussian func^n

$k(x, x') = \exp\left(\frac{-\|x - x'\|^2}{2\delta^2}\right) \quad \delta > 0$

$\Rightarrow \phi$ in this maps to an ∞ dimensional space

Activity Question - 5

1) Consider that 1000 data points belonging to d-dimensional space have a non-linear relationship. We apply kernel PCA to reduce the dimension of the data points and take the first k principal components. Can the value of k be larger than d?

- Yes
- [x] No

no, k never exceeds d

$k \leq d$

2) Why finding principal components (eigenvectors) in kernel PCA is not appreciated?

directly from lecture

- [x] It requires finding the transformation map ϕ and therefore, defeats the purpose of not finding ϕ.
- It cannot be calculated as ϕ can never be found.
- Principal components in kernel PCA do not capture the variance of the data.

$\phi(x_i)$ $\phi(x_j)$ $\to \square \to$ $\phi(x_i)^T(x_j) = K_{ij}$

x_i x_j ? same - lieresome process

3) Assume that we have a centered dataset. We apply a transformation ϕ on the same dataset. Is the dataset necessarily centered after applying ϕ?

- Yes
- [x] No

No, but we can create a centred kernel matrix K^c

4) What is the ij^{th} element of the matrix K in which the features are transformed to the higher dimensional space using a valid kernel k?

- [ ] $k(x_i^T x_j)$
- [x] $k(x_i, x_j)$
- [ ] $k(x_j, x_i)$
- [ ] $k(x_j^T x_i)$

$K_{ij} = k(x_i, x_j)$

5) Consider that kernel PCA was applied to two-dimensional data points. If w_k is the k^{th} principal component, what will be the projection of the point x_i on w_k, where $x_i \in \mathbb{R}^2$? Let ϕ be the mapping that transforms the data points into the higher dimension.

- $x_i^T w_k$
- $w_k^T x_i$
- [x] $\phi(x_i)^T w_k$
- $\phi(x_i) w_k^T$

$w_k = \phi(x)\,\alpha_k$

$\phi(x_i)^T w_k = \sum_{j=1}^{n} \alpha_{kj} K_{ij}$

Practice Assignment - 1

1) Assume that $w_k;\ k = 1, 2, ...d$ are d principal components corresponding to nonzero eigenvalues of the d-dimensional centered data points $x_i;\ i = 1, 2, ..., n$.

Statement 1: each x_i can be written as a linear combination of $w_k s$.

Statement 2: each w_k can be written as a linear combination of $x_i s$.

- Statement 1 is correct but statement 2 is incorrect.
- Statement 1 is incorrect but statement 2 is correct.
- Both statements are correct.
- Both statements are incorrect.

2) A transformation mapping ϕ is defined as

$\phi : \mathbb{R} \to \mathbb{R}^4$

$\phi(x) = [x^3, \sqrt{3}x^2, \sqrt{3}x, 1]^T$

Which of the following options are the same as $\phi(x_1)^T \phi(x_2)$ for two points $x_1, x_2 \in \mathbb{R}$?

Hint: Rather than doing the calculation, try to figure out the appropriate kernel function.

- [x] $(x_1x_2 + 1)^3$
- [x] $(x_2x_1 + 1)^3$
- [ ] $(x_1x_2 + 1)^4$
- [x] $\phi(x_2)^T \phi(x_1)$

3) Let C be the covariance matrix of n data points in d-dimensional space. Assume that the data points are mean-centered. If $2, 5$, and 7 are the only non-zero eigenvalues of C, what will be the non-zero eigenvalues of $X^T X$, where X is the matrix of shape (d, n) containing the data points? *1 poi*

- $2, 5, 7$
- $2d, 5d, 7d$
- $2n, 5n, 7n$

4) **Common Data for Questions 4 & 5** *1 poin*

Consider an image dataset matrix X of shape (d, n) with $d > n$. The kth principal component of the dataset can be written as $w_k = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$, where, the vector x_i is the ith data point. The k^{th} largest eigenvalue and the corresponding eigenvector of $X^T X$ are 4 and $[\frac{1}{\sqrt{51}}, \frac{3}{\sqrt{51}}, \frac{4}{\sqrt{51}}, \frac{5}{\sqrt{51}}]^T$, respectively.

What will be the value of α_1?

- 4
- $\frac{1}{\sqrt{51}}$
- $\frac{1}{2\sqrt{51}}$
- $[\frac{1}{\sqrt{51}}, \frac{3}{\sqrt{51}}, \frac{4}{\sqrt{51}}, \frac{5}{\sqrt{51}}]^T$
- $[\frac{1}{2\sqrt{51}}, \frac{3}{2\sqrt{51}}, \frac{4}{2\sqrt{51}}, \frac{5}{2\sqrt{51}}]^T$

5) What will be the k^{th} largest eigenvalue of the covariance matrix $\frac{1}{4}XX^T$?

Note that $n = 4$ as the length of the eigenvector of $X^T X$ is 4.

1

6) A function k is defined as

$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

$k([x_1, x_2]^T, [y_1, y_2]^T) = x_1^2 y_1^2 + x_2^2 y_2^2$

Is k a valid kernel?

Hint: Try to find out the appropriate ϕ.

- Yes
- No

7) A dataset of 1000 second-hand cars has four features: kilometers driven (x_1), mileage (x_2), the present price of the car (x_3), and the selling price (x_4). The selling price seems to have the following relationship (approximate) with the other three features. *1 p*

$x_4 = x_1^2 x_3 + 2x_2$

If we want to project the dataset into a lower dimensional space, which of the following task would be most appropriate?

- Standard PCA
- Kernel PCA with a polynomial kernel of degree 2
- Kernel PCA with a polynomial kernel of degree 3
- Kernel PCA with a polynomial kernel of degree 4

8) Abhishek runs a kernel PCA on a dataset containing n examples with d features. Which of the following strategy he should follow to center the data points?

strategy 1: First center the dataset using the mean and then apply the kernel.

Strategy 2: First apply the kernel and then center the matrix.

- Strategy 1
- Strategy 2
- Both strategies are the same

9) A dataset containing 1000 points in 3-dimensional space is run through the kernel PCA with the polynomial kernel of degree p. If the transformed dataset lives in a ten-dimensional space, what will be the value of p?

2

10) A dataset containing 1000 examples in 10-dimensional space is projected into other dimension space using kernel PCA with the following kernel.

$k(x_1, x_2) = \exp\left(\frac{-||x_1 - x_2||^2}{4}\right)$

What will be the dimension of the projected dataset?

- 10
- 40
- ∞
- Can not be determined

Graded Assignment - 1

From ques^n 1, k corresponds to identity transformation $\Rightarrow$ PCA = kernel + PCA

1) A function k is defined as follows.

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
$k(x_1, x_2) = x_1^T x_2$

$R^{d\times d} \to R$, $\phi(x) = x$, $k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$

Is k a valid kernel?

- [x] Yes — $k(x_1, x_2) = x_1^T x_2$
- [ ] No

2) If k is the valid kernel, we apply it to the three-dimensional dataset to run the kernel PCA. Select the correct options.

- [ ] We cannot run the PCA as k is not a valid kernel.
- [x] It will be the same as PCA with no kernel.
- [ ] It will be the same as the polynomial transformation of degree 2 and then run the PCA.
- [ ] It will be the same as the polynomial transformation of degree 3 and then run the PCA.

3) Consider ten data points lying on a curve of degree two in a two-dimensional space. We run a kernel PCA with a polynomial kernel of degree two on the same data points. Choose the correct options.

- [x] The transformed data points will lie on a 5-dimensional subspace of $\mathbb{R}^6$.
- [ ] The transformed data points will lie on a 6-dimensional subspace of $\mathbb{R}^{10}$
- [x] There will be some $w \in \mathbb{R}^6$ that all of the data points are orthogonal to.
- [ ] There will be some $w \in \mathbb{R}^{10}$ that all of the data points are orthogonal to.

$R^2 \to R^6$, $(x_1, x_2) \to (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$ [From lecture]

2D dataset $\Rightarrow$ 6D dataset (linear subspace)

so all pts. are orthogonal

4) Which of the following matrices can not be appropriate matrix $K = X^T X$ for some data matrix X?

- [x] $\begin{bmatrix} 1 & 8 \\ 8 & -1 \end{bmatrix}$
- [x] $\begin{bmatrix} 1 & 8 \\ 8 & 1 \end{bmatrix}$
- [ ] $\begin{bmatrix} -1 & 8 \\ -8 & 1 \end{bmatrix}$
- [x] $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

- non-zero values / +ve semi definite matrix is needed
- all are sym.

$\begin{bmatrix} 1 & 8 \\ 8 & 1 \end{bmatrix}$ $A - \lambda I = 0$ $\begin{bmatrix} 1-\lambda & 8 \\ 8 & 1-\lambda \end{bmatrix} = 0$

$(1-\lambda)^2 - 64 = 0$

$(1-\lambda)^2 = \sqrt{64}$ $\quad 1-\lambda = 8 \Rightarrow \lambda = -7$

↳ 2 ↳ 3

5) A function k is defined as

$k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$
$k(x_1, x_2) = (x_1^T x_2)^2$

Is k a valid kernel? $\phi(x) = x^2$

Verify with $x_1 = [a_1, a_2]^T$, $x_2 = [b_1, b_2]^T$

- [x] Yes
- [ ] No

6) Kernel PCA was run on the four data points $[1, 2]^T$, $[2, 3]^T$, $[2, -3]^T$, and $[4, 4]^T$ with the polynomial kernel of degree 2. What will be the shape of the matrix K? Notations are used as per lectures.

- [ ] 2×2
- [x] 4×4
- [ ] 6×6
- [ ] None of the above

$K = X^T X$ $\quad X = \text{shape}(d, n)$

$(n\times n) \quad (n \times d) \quad (d \times n)$

$n = 4$ (given)

$k(x_1, x_2) = (x_1^T x_2 + 1)^2$

7) Find the element at the index $(2, 3)$ of the matrix K defined in Question 6. Take the points in the same order.

- [ ] -4
- [x] 16

$2, 3 \Rightarrow k(x_2, x_3) = \left(\begin{bmatrix} 2 & 3 \end{bmatrix}_{1\times 2} \begin{bmatrix} 2 \\ -3 \end{bmatrix}_{2\times 1} + 1 \right)^2$

$= (-5 + 1)^2 = (4)^2 = 16$

8) A dataset containing 200 examples in four-dimensional space has been transformed into higher dimensional space using the polynomial kernel of degree two. What will be the dimension of transformed feature space?

15

degree 1 $\Rightarrow$ $f_1 / f_2 / f_3 / f_4 = 4$

degree 0 $\Rightarrow$ 1

degree 2

$f_1 f_2$ $f_2 f_3$ $f_1 f_3$ $f_1 f_4$ $f_2 f_4$ $f_3 f_4$ $\quad {}^4C_2 = \frac{4\times 3}{2} = 6$

f_1^2 f_2^2 f_3^2 f_4^4 $= 10$

9) Let $x_1, x_2, \ldots, x_n$ be d-dimensional data points $(d > n)$ and X be the matrix of shape $d \times n$ containing the data points. The k^{th} largest eigenvalue and corresponding unit eigenvector of $X^T X$ is λ and α_k, respectively. What will be the projection of x_i on the k^{th} principal component?

- [ ] $x_i^T \alpha_k$
- [ ] $\frac{x_i^T \alpha_k}{\lambda}$
- [x] $\frac{x_i^T X \alpha_k}{\sqrt{\lambda}}$
- [ ] $\frac{x_i^T X \alpha_k}{\sqrt{n\lambda}}$

k^{th} component $= \frac{\alpha_k X}{\sqrt{\lambda}}$

proj. of x_i of $k = \frac{x_i^T \alpha_k X}{\sqrt{\lambda}}$

10) Let k_1 and k_2 be two valid kernels. Is $3k_1 + 5k_2$ a valid kernel?

- [x] Yes
- [ ] No

for valid kernels, check
→ k is symmetric
→ matrix K is +ve semi-definite

MLT - Week3

- Clustering, partition the data into cluster, but there can be too many possibilities (k^n), so this is a np-hard problem, which means there is no optimum solution to it.

$$F(z_1, \dots, z_n) = \sum_{i=1}^{n} \| x_i - \mu_{z_i} \|_2^2$$

↳ Mean/average of z_i^{th} cluster

$$\mu_k = \frac{\sum_{i=1}^{n} x_i \mathbb{1}(z_i = k)}{\sum_{i=1}^{n} \mathbb{1}(z_i = k)}$$

$$\mathbb{1}(u) = \begin{cases} 1 & \text{if } u \text{ is true} \\ 0 & \text{o/w} \end{cases}$$

LLYOD'S ALGORITHM / K-MEANS ALGORITHM

INITIALIZATION (ITERATION) $z_1^0, z_2^0, \dots, z_n^0 \in \{1, \dots, k\}$

→ UNTIL CONVERGENCE

- COMPUTE MEANS $\forall k \quad \mu_k^t = \frac{\sum_{i=1}^{n} x_i \mathbb{1}(z_i^t = k)}{\sum_{i=1}^{n} \mathbb{1}(z_i^t = k)}$
- RE-ASSIGNMENT STEP $\forall i \quad z_i^{t+1} = \arg\min_k \| x_i - \mu_k^t \|_2^2$ [mean of if the current assignment is smallest, then don't reassign]

$$F(z_1^{t+1}, \dots z_n^{t+1}) < F(z_1^t, \dots, z_n^t)$$

- Lloyd algo always converges, but the converged solution may not always be optimum. Generally, in practice, it produces reasonable clusters.

$$\sum_{i=1}^{n} \| x_i - \mu_{z_i^{t+1}}^t \|^2 < \sum_{i=1}^{n} \| x_i - \mu_{z_i^t}^t \|^2$$ [By Algorithm's choice]

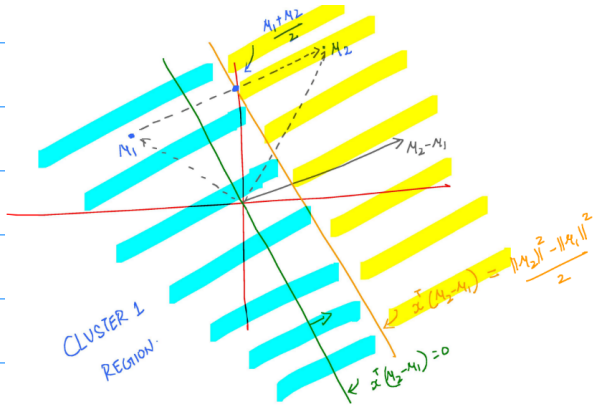
Mean of Cluster where x_i wants to go to

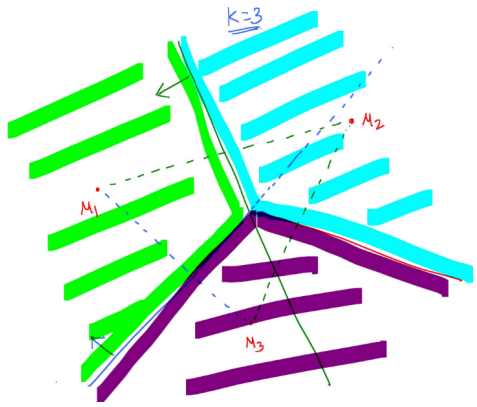
$F(z_1^t, \dots, z_n^t)$

Mean of Current Cluster where x_i is assigned to

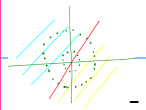
$$\sum_{i=1}^{n} \| x_i - \mu_{z_i^{t+1}}^{t+1} \|^2 \leq F(z_1^{t+1}, \dots, z_n^{t+1}) \leq \sum_{i=1}^{n} \| x_i - \mu_{z_i^t}^t \|^2$$

- The obj. func strictly reduces after each re-assignment. There are only finite number of assignments. Algo. must converge.





- Cluster regions are intersection of half spaces, they are also called as Voronoi regions.



- How to fix such clusters? - Kernelize K means - Spectral clustering

- Now initialization has 2 possible ways -
  1. pick k means uniformly at random from the dataset
  2. K-means ++ - choose first mean u1 uniformly at random from {x1,.....,xn}

- So basically, we want k to be as small as possible and we penalize large values of k.

Some common Criterion

A.I.C - Akaike Information Criterion $[2k - 2\log(L(\theta^*))]$

B.I.C - Bayesian Information Criterion $[k\log(n) - 2\log(L(\theta^*))]$

- CONVERGENCE - YES
- NATURE OF CLUSTERS - VORNOI REGIONS
- INITIALIZATION - K-MEANS++
- CHOICE OF K - OBJ + PENALTY(k). → this should be minimized

GUARANTEE

$$\mathbb{E}\left[\sum_{i=1}^{n} \| x_i - \mu_{z_i} \|^2\right] \leq O(\log k) \left[\min_{z_1, \dots, z_n} \sum_{i=1}^{n} \| x_i - \mu_{z_i} \|^2\right]$$

over randomness of algorithm

Activity Question - 1

1) How many configurations are possible for partitioning 100 data points into 10 clusters if these clusters are allowed to be empty?

- 10^{100}
- 100^{10}
- $100 * 10$
- $\binom{100}{10}$

n data pts → k clusters

possible clusters = k^n

100 data pts → 10 clusters

Config. = 10^{100}

2) A cluster will be good if:

- [ ] The points in the cluster are far apart.
- [x] The points in the cluster are close to the mean of the cluster.
- [x] The points in the cluster are homogeneous.
- [ ] The points in the cluster are heterogeneous.

3) If $z_1, z_2, ..., z_n$ represent the clusters assigned to data points $x_1, x_2, ..., x_n$ respectively, then which of the following would represent the mean (μ_k) of cluster k?

- $\mu_k = \sum_{i=1}^{k} z_i$
- $\mu_k = \sum_{i=1}^{k} x_i$
- $\mu_k = \sum x_i 1(z_i = k)$
- $\mu_k = \frac{\sum 1(z_i = k)}{\sum x_i 1(z_i = k)}$
- $\mu_k = \frac{\sum x_i 1(z_i = k)}{\sum 1(z_i = k)}$

$\{x_1 \dots x_n\}$
$\{z_1 \dots z_n\}$

$\mu_k = \frac{\sum_{i=1}^{n} x_i \mathbb{1}(z_i = k)}{\sum_{i=1}^{n} \mathbb{1}(z_i = k)}$

4) If $z_1, z_2, ..., z_n$ represent the clusters assigned to data points $x_1, x_2, ..., x_n$ respectively, and μ_k represents the mean of cluster k, then which of the following would be a good measure for 'goodness' of partitions?

- $\min_{z_1,z_2,...,z_n} \sum_i^n ||x_i - \mu_i||^2$
- $\min_{z_1,z_2,...,z_n} \sum_i^n ||x_i - z_i||^2$
- $\min_{z_1,z_2,...,z_n} \sum_i^n ||x_i - \mu_{z_i}||^2$

$\{x_1 \dots x_n\}$
$\{z_1 \dots z_n\}$

μ_k (from above)

$\min_{z_1 \dots z_n} \sum_{i=1}^{n} ||x_i - \mu_{z_i}||^2$

Activity Question - 2

1) Does k-means always converge?

- Yes
- No

Yes k means / Lloyds algo. always converges

2) The solution produced by k-means algorithm is always optimal.

- True
- False

it converges, but it's not necessary that soln is optimum

3) Which of the following is/are True? *1 point*

- [ ] K-means algorithm automatically determines the optimal value of K.
- [x] K-means algorithm can not automatically determine the value of K. It has to be provided as an input parameter.
- [x] Finding optimal clusters is an NP-hard problem.
- [ ] K-means helps us find a solution to an NP-hard problem.

$\{x_1 \dots x_n\}$ } input
$\{z_1 \dots z_n\}$ } parameters are necessary

& there is no optimum soln to np-hard problems.

4) If in an iteration t, the distance of a data point (x_i) from the mean of its currently assigned cluster (z_i^t) is the least as compared to its distance from the means of the other clusters, we do not change the cluster assigned to this data point (x_i) in this iteration. *1 point*

In this context, which of the following is true?

- This implies that z_i^t will be the final cluster assigned to x_i and it's cluster will never change in any subsequent iteration.
- The cluster assigned to x_i may change in some of the subsequent iteration(s).

it may Δ

Activity Question - 3

1) Consider a bunch of points $x_1, x_2, ..., x_l \in \mathbb{R}^d$. Which of the following would represent a point v whose sum of squared distances from the points $x_1, x_2, ..., x_l$ is the minimum?

- $v = \sum_{i=1}^{l} x_i$
- $v = \frac{1}{l} \sum_{i=1}^{l} x_i^2$
- $v = \sum_{i=1}^{l} x_i^2$
- $v = \frac{1}{l} \sum_{i=1}^{l} x_i$

$v = \frac{1}{l} \sum_{i=1}^{n} x_i$

↳ the avg. of all data pts.

2) For data points $x_1, x_2, ..., x_n$, partition configuration $z_1, z_2, ..., z_n$ and mean of the k^{th} cluster μ_k, which of the following represents the objective function of Lloyd's algorithm?

- $\max \sum_{i=1}^{n} ||x_i - \mu_{z_i}||^2$
- $\min \sum_{i=1}^{n} ||x_i - \mu_{z_i}||^2$
- $\max \sum_{i=1}^{n} ||x_i - \mu_k||^2$

func^n $\min \sum_{i=1}^{n} ||x_i - \mu_{z_i}||^2$

3) It the value of objective function strictly reduces, it indicates that,

- The partitions can not repeat.
- The partitions may repeat.

↳ partition won't repeat, the value obtained now can be considered as final answer

4) Choose the correct statement.

- There are only a finite number of partition configurations.
- There are an infinite number of partition configurations.

(fact)

Activity Question - 4

1) If μ_1 and μ_2 are means of two clusters c_1 and c_2 in k-means, then for all the data points settling in c_1,

- The means μ_1 and μ_2 are equidistant.
- The mean μ_2 is closer than μ_1
- The mean μ_1 is closer than μ_2
- Insufficient Information.

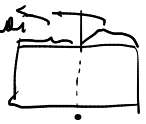
for all these pts

dist. of pt. to μ_1 < dist. of pts. to μ_2

2) For every cluster, it's Voronoi region will be formed from intersection of $k - 1$ half spaces if k is the total number of clusters.

- True
- False

Voronoi eg.



k = 2

k = 1 < 1

we use kernel k means → for this k means ✗

3) Choose the correct statement(s).

- [ ] K-means can efficiently cluster data points that are not linearly separable. ✗
- [x] K-means can not efficiently cluster data points that are not linearly separable.
- [x] Kernel K-means is used to cluster data points that are not linearly separable.

4) Which of the following is true with respect to optimality of clusters? *1 point*

- ○ K-means may not guarantee obtaining the optimal clusters, but there are other techniques that guarantee optimal clusters. ✗
- ◉ Getting optimal clusters is an NP-hard problem, so no technique exists that can guarantee optimal clusters

as discussed earlier

Activity Question - 5

1) To obtain good clusters, how should the initial means of clusters be?

- ◉ Means should be as far apart as possible. to get best results
- ○ Means should be as close to each other as possible.
- ○ A random placement of means is the best placement.

2) K-means++ initializes the clusters' means

- [ ] Randomly (Uniformly at random) } way 1 → pick any random means from the dataset
- [ ] Deterministically.
- [x] Probabilistically, by assigning a different probability to each data point. } this is 2nd way of initialization
- [x] In an iterative fashion

3) During different iterations of the initialization step, k-means++ assigns highest probability to those points to be chosen as clusters' means for which

- ◉ The closest cluster mean (of already chosen clusters) is the farthest. to get finer clustering results
- ○ The farthest cluster mean (of already chosen clusters) is the closest.

4) Due to a smart initialization of cluster means, K-means++ is able to produce optimal clusters.

- ○ True
- ◉ False — no whatever we do, we can get better clusters, but never optimum clusters.

Activity Question - 6

1) K-means algorithm automatically determines the value of K as per the given data.

- ○ True
- ◉ False — no, no automatic process

2) Which of the following is correct with respect to the value of k?

- ○ The value of K should be as large as possible.
- ○ The value of k should be as small as possible.
- ◉ The value of k should neither be very small nor very large. should be of med. size

3) If *obj* represents the value of objective function and P represnts penalty, then the best value of k will be that for which

- ○ (obj*P) is minimum.
- ◉ (obj+P) is minimum. best → minimize (obj + P) ↓
- ○ (obj-P) is minimum.
- ○ (obj/P) is minimum.

4) Which of the following techniques help(s) in fixing a suitable value of penalty?

- [ ] Elbow method → this is also a method but not considered in this course.
- [x] Akaike Information Criterion
- [x] Bayesian Information Criterion

Practice Assignment - 1

1) Which of the following sequences is correct for K-Means algorithm?

- Assign each data point to the nearest cluster centres.
- Re-assign each point to nearest cluster centres.
- Assign cluster centres randomly.
- Re-compute cluster centres.
- Specify the number of clusters.

- ○ 3, 5, 1, 4, 2
- ○ 5, 3, 1, 2, 4
- ◉ 5, 3, 1, 4, 2
- ○ 3, 5, 2, 4, 1
- ○ None of these

k means →
① specify k
② assign cluster randomly
③ assign each data pt. to it's best cluster center
④ recompute clusters
⑤ re-assign to new cluster center

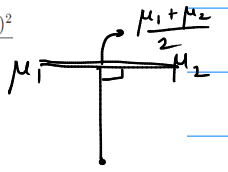
2) If $F(z_1^t, z_2^t, ..., z_n^t)$ represents the value of objective function in iteration t of Lloyd's algorithm, then which of the following is true?

- ○ $F(z_1^{t+1}, z_2^{t+1}, ..., z_n^{t+1}) > F(z_1^t, z_2^t, ..., z_n^t)$
- ◉ $F(z_1^{t+1}, z_2^{t+1}, ..., z_n^{t+1}) < F(z_1^t, z_2^t, ..., z_n^t)$
- ○ $F(z_1^{t+1}, z_2^{t+1}, ..., z_n^{t+1}) = F(z_1^t, z_2^t, ..., z_n^t)$

cluster is fed only if the new dist. is less than the previous one. After each re-assignment, obj. func^n ↓

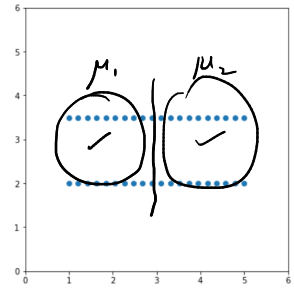
3) If μ_1 and μ_2 are means of two clusters in k-means, then the boundary between the two clusters will be (avg.)

- ○ Perpendicular to the line joining μ_1 and μ_2 and at the point $\frac{(\mu_1+\mu_2)^2}{2}$
- ○ Parallel to the line joining μ_1 and μ_2 and at the point $\frac{(\mu_1+\mu_2)^2}{2}$
- ◉ Perpendicular to the line joining μ_1 and μ_2 and at the point $\frac{\mu_1+\mu_2}{2}$
- ○ Parallel to the line joining μ_1 and μ_2 and at the point $\frac{\mu_1+\mu_2}{2}$



smart clustering } probabilistic model

4) Consider the following data points:



Assume K-means is run on these points with k = 2. Which of the following are expected to be the clusters formed out of K-means?

7) For 1000 data points, out of k = 1, 10 and 100, which value of k is likely to result in the maximum value of the objective function?

- 1 — k=1 ⇒ all data pts. will be in same 1 cluster. & the obj func^n will be maximum in this case.
- 10
- 100
- Insufficient information. Depends on data.

8) For 100 data points, if k = 100, what will be the value of the objective function?

- 100
- 0 — 100 n → 100 k, the obj func^n will be least the dist. of each pt. from its mean will be 0
- 100*100

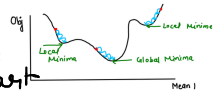
5) In the initialization step of k-means++, the squared distances from the closest mean for 5 points x_1, x_2, x_3, x_4, x_5 are: 1 point
25, 67, 89, 24, 56. In this context, which of the following is true? — x_3 is max — the prob. of being chosen is highest

- Any point out of x_1, x_2, x_3, x_4, x_5 may be chosen uniformly at random as next mean.
- Certainly x_3 will be chosen as its distance from closest mean is largest.
- x_3 will be chosen with the highest probability, but we are not sure whether this point will definitely be chosen.

6) With respect to Lloyd's algorithm, choose the correct statements: 1 point

- The partition configurations can not repeat themselves. — exp^n ✗
- After doing the reassignments, we might get the same partition configuration again. — not possible
- Objective function after making the re-assignments strictly reduces. ↓
- Objective function after making the re-assignments may increase. ✗ — occurs only if obj. func^n is used. vice-versa is ✓.
- Change of value of objective function indicates that the partition configuration has changed.

9) Choose the correct statements:



- In k-means algorithm, all cluster initializations lead to the same result — it all depends on where u start
- One initialization might get stuck in local minima, while another may lead to global minima.
- One initialization may converge while another may not.
- The initialization of cluster centres may affect the number of iterations K-means takes to converge. — if we initialize at global minima, then it takes 0 steps to converge, etc.

10) Outliers are data points that deviate significantly from the rest of data points. Knowing the way Lloyd's algorithm works, do you think it is sensitive to outliers?

- Yes — k means → means & euclidean distance is calculated, hence, outliers may affect the computa^n
- No,

Graded Assignment - 1

1) What would be the correct relationship among the following three quantities?: 1

(1) $\sum_{i=1}^{n} ||x_i - \mu_{z_i}^{t}||^2$, — value of obj func^n at iter^n t

(2) $\sum_{i=1}^{n} ||x_i - \mu_{z_i}^{t}||^2$ and — the intermediate value when centers are moving.

(3) $\sum_{i=1}^{n} ||x_i - \mu_{z_i}^{t+1}||^2$ — value of obj. func^n at iter^n t+1

where $\mu_{z_i}^{t}$ and $\mu_{z_i}^{t+1}$ refer to means of cluster z_i in iterations t and $t+1$ respectively. And $\mu_{z_i}^{t+1}$ is the mean of the cluster z_i where x_i is going to move in the next (i. e., $(t+1)^{th}$) iteration.

- (1) > (2) < (3)
- (1) < (2) < (3) — 1 > 2 > 3 this is occuring since the better clusters were possible, hence re-assignment took place & every time obj. func^n strictly reduced.
- (1) > (2) > (3)
- (1) < (2) > (3)

2) Consider that in an iteration t of Lloyd's algorithm, the partition configuration (P^t) is $z_1^t, z_2^t, ..., z_n^t$ where each $z_i^t \in$ 1, 2, ..., k. Assume that the algorithm does not converge in iteration t, and hence some re-assignment happens, thus updating the partition configuration in the next iteration (P^{t+1}) to $z_1^{t+1}, z_2^{t+1}, ..., z_n^{t+1}$. How can we say that partition configuration P^{t+1} is better than P^t? 1 point

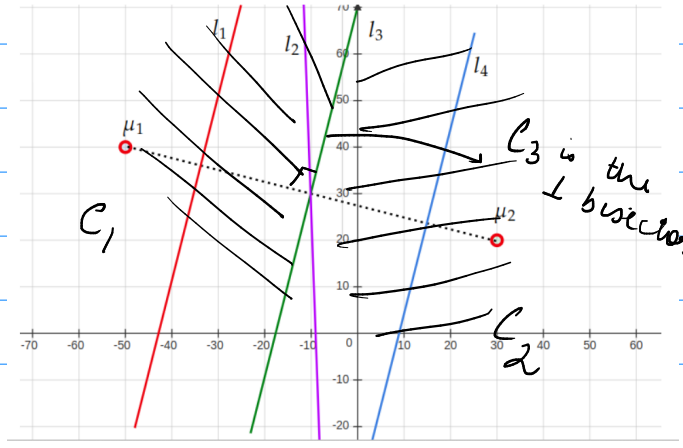
- The value of the objective function for P^{t+1} should be more than that for P^t
- The value of the objective function for P^{t+1} should be lesser than that for P^t
- The value of the objective function for P^{t+1} and P^t should be same.

reason same as before (reassignment happens if the data pt. found a closer mean)

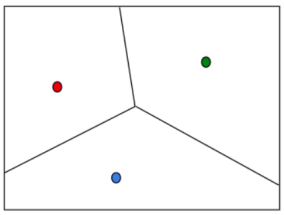
3) With respect to Lloyd's algorithm, choose the correct statements: 1 point

- At the end of k-means, the objective function settles in a local minima and reaching global minima may not be guaranteed. — sometimes it is local too
- At the end of k-means, the objective function always settles in the global minima. ✗
- The clusters produced by K-means are optimal. ✗ — NP-hard problem
- If the resources are limited and the data set is huge, it will be good to prefer K-means over K-means ++.
- In practice, k should be as large as possible. ✗ — as small as possible) elaborate initializa^n step of k means++ will take huge amt. of time

4) Consider two cluster centres μ_1 and μ_2 corresponding to two clusters C_1 and C_2 as shown in the below image. 1
Consider four half spaces represented by lines l_1, l_2, l_3 and l_4. Where would the data points falling in cluster C_1 lie?

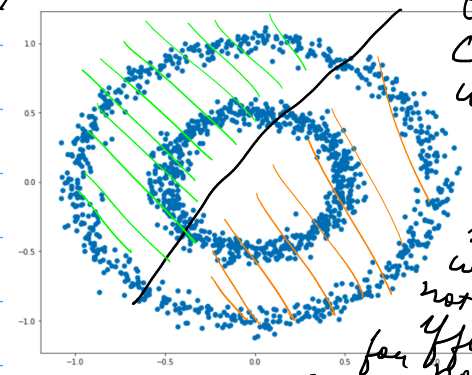


5) Which of the following best represents a valid voronoi diagram for K-means algorithm with K = 3? (The dots represent the cluster centres of respective clusters.)



half spaces are ⊥ bisectors of the line joining the cluster centers

6)



these clusters will be formed by k-means which is not much efficient for non-linear models

Assume that K-means is applied on this data with $k = 2$. Which of the following are expected to be the cluster centres found?

7) Assume that in the initialization step of k-means++, the squared distances from the closest mean for 10 points **1 point**
$x_1, x_2, \ldots, x_{10}$ are: 25, 67, 89, 24, 56, 78, 90, 85, 35, 95. Which point has the highest probability of getting chosen as the next mean and how much will that probability be?

- x_4, 0.24 ✗
- x_4, 0.037 ✓
- x_{10}, 0.95
- x_{10}, 0.1475 ✓

Sum = 644

$x_4 = \frac{24}{644} = 0.037$ $x_{10} = \frac{95}{644} = 0.147$

since, it's probab. is highest b/tw. these 10 pts.

8) Consider 7 data points $x_1, x_2, \ldots, x_7$: {(0, 4), (4, 0), (2, 2), (4, 4), (6, 6), (5, 5), (9, 9)}. Assume that we want to form 3 **1 pc**
clusters from these points using K-Means algorithm. Assume that after first iteration, clusters $C1, C2, C3$ have the following data points:

$C1$: {(2,2), (4,4), (6,6)}

$C2$: {(0,4), (4,0)}

$C3$: {(5,5), (9,9)}

After second iteration, which of the clusters is the data point (2, 2) expected to move to?

I → II

$C_1 : (4,4)$

$C_2 : (2,2)$ ←

$C_3 : (7,7)$

since the distance of data pt. (2,2) from mean (2,2) is 0.

- C_1
- $C2$ ✓
- C_3
- Can't say, it is not deterministic.

9) Which of the following statements are True?

since (means + euclidian dist.)

- K-means is extremely sensitive to cluster center initializations. ✓
- Bad initialization can lead to poor convergence speed. ✓ → yes, as discussed
- Bad initialization can lead to bad overall clustering. ✓

since, it's the 1st step to clustering

- 1 and 3
- 1 and 2
- 2 and 3
- 1, 2, and 3 ✓

10) If the data set has two features x_1 and x_2, which of the following are true for K means clustering with k = 3?

If x_1 and x_2 have a correlation of 1, the cluster centres will be in a straight line. ✓

If x_1 and x_2 have a correlation of 0, the cluster centres will be in straight line.

- 1
- 2
- None of these. Correlation does not affect cluster centres' position.

2 features
k = 3

x_1, x_2 have correla as 1 then data pts. lie along the line & hence cluster centers will also form a straight line.

MLT - Week 4

- Estimation - probabilistic assumption, observe data, assume a probabilistic model that generates the data, estimate unknown parameters using data assumption - iid sample

Activity Question - 1

1) Sequentially order the following steps in an estimation problem: *1 poin*

(1) Estimate the unknown parameters of the model using data

(2) Observe the data

(3) Assume a probabilistic model that generates the data

- ○ (1) → (2) → (3)
- ○ (2) → (1) → (3)
- ○ (3) → (2) → (1)
- ◉ (2) → (3) → (1)

Observe 2 → model 3 → find unknowns

2) Consider the following observation of data-points:

$\{0, \ 1, \ 0, \ 0, \ 0, \ 1, \ 0, \ 0, \ 1, \ 1, \ 1, \ 0, \ 0, \ 1\}$

If you assume the "coin-within-box" mechanism as the probabilistic model with parameter p, what is a good estimate of p? Enter your answer correct to two decimal places.

0.40

6/15 = 0.40 = no. of 1s

3) Consider a "coin-within-box" mechanism with true parameter $p = 0.8$. What can you say about the possibility and probability of the following points being generated by this model?

$\{0, \ 0, \ 0, \ 0, \ 1, \ 0, \ 0, \ 0, \ 0, \ 1\}$

- ○ It is impossible.
- ○ It is possible and highly probable.
- ◉ It is possible but highly improbable.
- ○ It is possible, but we can't comment much on the probability.

0.8 → 8/10, 0.2 → we can't say impossible since 0.8 is also a estimate only

4) Consider a dataset set of n observations each of which could be a zero or a one:

$x_1, \cdots, x_n$

Which of the following equations expresses the independence assumption?

- ○ $P(x_i) = P(x_j), \quad \forall\, i, j \in 1, \cdots, n$
- ○ $P(x_i \mid x_j) = P(x_j), \quad \forall\, i, j \in 1, \cdots, n, \text{ such that } i \neq j$
- ◉ $P(x_i \mid x_j) = P(x_i), \quad \forall\, i, j \in 1, \cdots, n, \text{ such that } i \neq j$

$P(x/y) = P(x)$

5) Consider a sequence of observations, all of which lie in $0, 1$:

$x_1, \cdots, x_n$

We assume the "coin-within-box" mechanism as the probabilistic model that generates this data. Given the independence assumption, is the following equation true?

$P(x_n \mid x_1, \cdots, x_{n-1}) = P(x_n)$

- ◉ Yes
- ○ No

this part doesn't matter in the case of independence.

6) Assume that you have two sets of observations X and Y that are generated from two different "coin-within-box" setups, with parameters p_x and p_y respectively, where $p_x \neq p_y$:

$\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$

Which of the following statements are true?

- ☑ The observation x_i is independent of the observation y_j.
- ☐ The observations x_i and y_j are identically distributed.
- ☐ The observation x_i is not independent of the observation y_j.
- ☑ The observations x_i and y_j are not identically distributed.

p_x p_y since they are not from the same box

7) Consider a generative story where there are two biased coins within a box, with different probabilities. To generate a data-point, one of these two coins is picked uniformly at random and this coin is tossed. What can you say about the observations? Note that you don't know which coin was chosen as that happens within the box and is not a part of the observation.

- ◉ They are independent and identically distributed.
- ○ They are independent but not identically distributed.

independent identically distributed

8) The parameters of the two coins in the previous question are 0.3 and 0.9. Can we replace the two coin-setup with a single coin? If the answer is in the negative, enter 0. If the answer is in the affirmative, enter the parameter of this coin.

0.6

$\frac{0.3 + 0.9}{2} = \frac{1.2}{2} = 0.6$

$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$ ← $\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Activity Question - 2

1) What is the expression for the likelihood function for the following observations if we assume a Bernoulli distribution with parameter p as our model? *1 p*

$\{1, \ 1, \ 0, \ 0, \ 1, \ 0, \ 0\}$

- ○ p^3
- ○ $(1-p)^4$
- ◉ $p^3(1-p)^4$
- ○ $p^3 + (1-p)^4$
- ○ $p^4(1-p)^3$

$1 \to p$, $0 \to (1-p)$

$L = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

$= p \times p \times (1-p) \times (1-p) \times p \times (1-p) \times (1-p)$

$= p^3(1-p)^4$

2) What is the maximum likelihood estimate of the parameter p based on the observed data from the previous question? Enter your answer correct to two decimal places.

0.43

$p = \frac{\text{no. of 1s}}{} = \frac{3}{7} = 0.42857 \sim 0.43$

3) Why do we work with log-likelihoods rather than just the likelihood? *1 po*

- ☑ log is a monotonically increasing function
- ☑ Since the likelihood involves the product of quantities less than 1, the likelihood will be a very small number, making it a hard quantity to work with computationally. log of fractions is computationally easier to handle.
- ☑ log transforms products to sums and thus making the objective function more friendly for optimization, especially when taking derivatives.

it becomes a sum

it's easier to make summations

4) Which of the following is the expression for the likelihood function for a Gaussian distribution with parameters (μ, σ^2) given the dataset $x_1, \cdots, x_n$? Here, f is the density of the Gaussian. *1 po*

- ○ $\prod_{i=1}^{n} P(x_i \,;\, \mu, \sigma^2)$
- ◉ $\prod_{i=1}^{n} f(x_i \,;\, \mu, \sigma^2)$
- ○ Both (a) and (b)
- ○ Neither (a) nor (b)

$\prod_{i=1}^{n} \int (x_i ; \mu, \sigma^2)$

5) Which of the following is the correct formulation of the optimization problem for determining the ML estimate of the mean of a Gaussian distribution given a variance σ^2 which has already been determined? The dataset is $x_1, \cdots, x_n$. *1 point*

- $\max_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$
- ✓ $\max_{\mu} \sum_{i=1}^{n} -(x_i - \mu)^2$
- Both (a) and (b) are wrong as the objective function depends on σ

$\max_{\mu} \sum_{i=1}^{n} -(x_i - \mu)^2$

$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$= \frac{1}{10} \times 18.8 = 1.88$

6) What is the maximum likelihood estimate for the mean of a Gaussian distribution that generates the following data? Assume that the variance of the distribution is a constant. Enter your answer correct to two decimal places.

$\{2.9, \quad 1.9, \quad 0.9, \quad 3.2, \quad 0.5, \quad 1.6, \quad 2.1, \quad 1.3, \quad 1.7, \quad 2.7\}$

✓ 1.88

HUNCH → codified using a probability distribution over θ → $P(\theta)$ PRIOR

⇓ DATA

UPDATED HUNCH → codified using a prob. distribution → $P(\theta/\text{DATA})$ POSTERIOR

A → Parameters θ

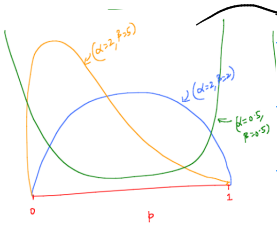
B → DATA $\{x_1, \cdots x_n\}$

$P(\theta/\{x_1, \cdots, x_n\}) = \frac{P(\{x_1 \cdots x_n\}/\theta)}{P(\{x_1, \cdots, x_n\})} \cdot P(\theta)$

LIKELIHOOD; PRIOR; POSTERIOR; EVIDENCE (does not depend on θ)

BETA PRIOR

$f(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Z}$



this shows how for diff. values of α, β the curve is

BETA PRIOR (α, β) —DATA, Bernoulli→ BETA POSTERIOR $(\alpha + n_h, \beta + n_t)$

$\mathbb{E}_{\text{posterior}} = \frac{\alpha + n_h}{(\alpha+\beta) + n}$ $n_h + n_t$

$\hat{p}_{map}$ = MAP = Max. aposteriori estimation

Activity Question - 3

1) Consider the following hunch/belief about the parameter of a Bernoulli distribution: "the parameter p is neither too small nor too large". In the context of Bayesian estimation, is the following statement true?

This hunch is formed **after** analyzing the dataset. That is, the hunch is a function of the dataset.

- True
- ✓ False

such statements make sense in Bayesian analysis & not after

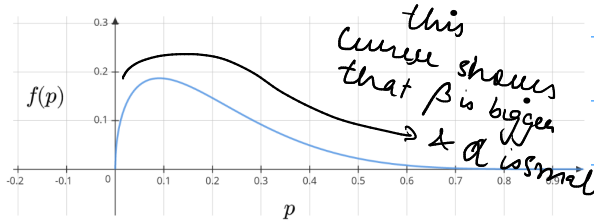
2) Which of the following is the correct relationship between the prior, posterior, likelihood and evidence?

- ✓ $\text{posterior} = \frac{\text{likelihood}}{\text{evidence}} \cdot \text{prior}$ ✓ formula
- $\text{posterior} = \frac{\text{evidence}}{\text{likelihood}} \cdot \text{prior}$
- $\text{prior} = \frac{\text{likelihood}}{\text{evidence}} \cdot \text{posterior}$

3) Consider the beta distribution: *1 point*

$f(p; \alpha, \beta) = \frac{1}{Z} \cdot p^{\alpha-1}(1-p)^{\beta-1}, \quad \forall p \in [0, 1]$

For what values of α and β does the distribution take the following form? Use a tool like Desmos or Geogebra to arrive at the answer. Ignore the values on the y-axis, just focus on the shape of the distribution.



this curve shows that β is bigger & α is smaller

- $\alpha = 2, \beta = 2$
- ✓ $\alpha = 1.5, \beta = 6$
- $\alpha = 0.5, \beta = 5$
- $\alpha = 6, \beta = 1.5$

$p^{(\alpha-1)} \; 1-p^{(\beta-1)}$ put these values & see the answer

4) Based on the previous question, how would you describe the prior in English?

- ✓ There is a high chance for the parameter of the Bernoulli distribution to be a small value.
- There is a high chance for the parameter of the Bernoulli distribution to be a large value.
- The parameter of the Bernoulli distribution is neither too high nor too low.
- The parameter of the Bernoulli distribution is either too high or too low.

6) In the context of the previous problem, what is $\hat{p}$, a point estimate for the parameter of the Bernoulli distribution, if we use the expectation of the posterior as the method of estimation? Enter your answer correct to two decimal places.

✓ 0.61

$\mathbb{E} = \frac{\alpha + n_h}{(\alpha+\beta) + n} = \frac{3+8}{5+13} = \frac{11}{18}$

$= 0.61$

5) Consider the following prior for the parameter p of a Bernoulli distribution:

$p \sim \text{Beta}(3, 2)$

The dataset observed is as follows:

$\{1, \quad 0, \quad 1, \quad 0, \quad 1, \quad 1, \quad 1, \quad 0, \quad 0, \quad 0, \quad 1, \quad 1, \quad 1\}$

What is the posterior?

Beta (α, β)

Post. $= (\alpha + n_h, \beta + n_t) = (3+8, 2+5)$

$n_h = 8$

$n_t = 5$

$\alpha = 3$

$\beta = 2$

Beta $= (11, 7)$

- $\text{Beta}(3, 2)$
- ✓ $\text{Beta}(11, 7)$
- $\text{Beta}(8, 10)$

STEP 1: Generate a mixture component among $\{1, \cdots, K\}$ $z_i \in \{1, \cdots, K\}$

$P(z_i = k) = \pi_k$ $\left[\sum_{i=1}^{K} \pi_i = 1,\ 0 \le \pi_i \le 1\ \forall i\right]$

STEP 2: Generate $x_i \sim N(\mu_{z_i}, \sigma^2_{z_i})$

$L(\theta) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}\right]$ (θ: all parameters)

$\log L(\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}\right)$ — (*)

- Not possible to solve this analytically.
- Need an alternate way to solve this efficiently!

Activity Question - 4

1) In the generative story for a mixture of K Gaussians, what are the two steps? *1 point*

- [ ] Step-1: Pick a mixture uniformly at random.
- [x] Step-2: Generate a data-point from the mixture picked in step-1
- [x] Step-1: Pick a mixture based on the categorical distribution over the mixtures specified by the π_ks.

π_ks

2) If there are three mixtures in a Gaussian mixture model, what can you say about the following expression for some point x_i?

$P(z_i = 1) + P(z_i = 2) + P(z_i = 3)$

1

3 mixtures; probabs. always add upto 1.

3) Consider the second step of the generative story. We wish to generate a point x_i from one of the mixtures. If z_i is the mixture index obtained from step-1, which of the following is the distribution that we have to sample from? This question is to test if you have understood the notations properly. *1 point*

- $\mathcal{N}(\mu_i, \sigma_i^2)$
- $\mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2)$

z_i

$N(\mu_{z_i}, \sigma^2_{z_i})$

4) For a Gaussian mixture models with five mixtures, how many free parameters do we need to estimate?

14

$K = 5$

$3K - 1$ or $2K + K - 1$

$15 - 1 = 14$ $\quad 10 + 4 = 14$

Activity Question - 5

1) What does the following expression correspond to in the context of a GMM? Here z_i is the latent variable corresponding to the point x_i.

$\frac{1}{\sqrt{2\pi}\sigma_k} \cdot \exp\left[\frac{-(x_i - \mu_k)^2}{2\sigma_k^2}\right]$

Here $f(.)$ and $P(.)$ are the pdf and pmf of the random variable in question.

- $P(z_i = k \mid x_i)$
- $f(x_i \mid z_i = k\ ;\ \mu_k, \sigma_k^2)$
- $f(x_i \mid z_i = k\ ;\ \mu_k, \sigma_k^2)$
- $f(x_i)$

$f(x_i \mid z_i = k\ ;\ \mu_k, \sigma_k^2)$

2) Which of the following expressions is the likelihood of a dataset $x_1, \cdots, x_n$ for a GMM with K mixtures and parameters $(\pi_k, \mu_k, \sigma_k^2)$ for the k^{th} mixture? f is the density of a Gaussian.

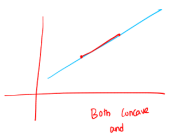
- $\prod_{i=1}^{n}\left[\prod_{k=1}^{K} \pi_k \cdot f(x_i;\ \mu_k, \sigma_k^2)\right]$
- $\prod_{i=1}^{n}\left[\sum_{k=1}^{K} \pi_k \cdot f(x_i;\ \mu_k, \sigma_k^2)\right]$

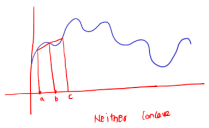
$\prod_{i=1}^{n}\left[\sum_{k=1}^{K} \pi_k \cdot f(x_i;\ \mu_k, \sigma_k^2)\right]$

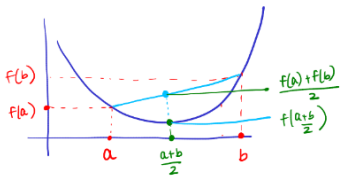
3) What is the contribution of the i^{th} data-point to the log-likelihood function of a GMM with K mixtures? The density of the k^{th} mixture is given by $f(x\ ;\ \mu_k, \sigma_k^2)$.

- $\sum_{k=1}^{K} \log\left[\pi_k f(x_i\ ;\ \mu_k, \sigma_k^2)\right]$
- $\prod_{k=1}^{K} \log\left[\pi_k f(x_i\ ;\ \mu_k, \sigma_k^2)\right]$
- $\log\left[\sum_{k=1}^{K} \pi_k f(x_i\ ;\ \mu_k, \sigma_k^2)\right]$

$\log\left[\sum_{k=1}^{K} \pi_k f(x_i;\ \mu_k, \sigma_k^2)\right]$

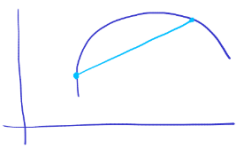






$\forall a, b$ $\quad f\left(\frac{a+b}{2}\right) \le \frac{f(a)+f(b)}{2}$

CONVEX FUNCTION



$\forall a, b$ $\quad f\left(\frac{a+b}{2}\right) \ge \frac{f(a)+f(b)}{2}$

CONCAVE FUNCTION

For concave $\log$ is a concave func^n

$f(\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_K a_K) \ge \lambda_1 f(a_1) + \cdots + \lambda_K f(a_K)$

$\left[\sum_{i=1}^{K} \lambda_i = 1,\ 0 \le \lambda_i \le 1\right]$

JENSEN'S INEQUALITY.

$f\left(\sum_{k=1}^{K} \lambda_k a_k\right) \ge \sum_{k=1}^{K} \lambda_k f(a_k)$

$\log\left[\sum_{k=1}^{K} \pi_k f(x_i;\ \mu_k, \sigma_k^2)\right]$

Fixing λ, we get

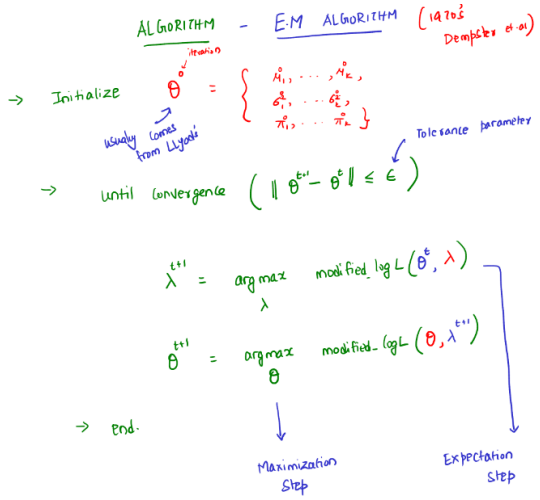
$\hat{\mu}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i x_i}{\sum_{i=1}^{n} \lambda_k^i}$ $\quad \hat{\sigma}_k^{2\,MML} = \frac{\sum_{i=1}^{n} \lambda_k^i (x_i - \hat{\mu}_k^{MML})^2}{\sum_{i=1}^{n} \lambda_k^i}$

$\hat{\pi}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i}{n}$

Fixing θ, we get

$\lambda_k^{i\,MML} = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}\right) \cdot \pi_k}{\sum_{k=1}^{K}\left(\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}} \cdot \pi_k\right)} \leftarrow P(z_i)$

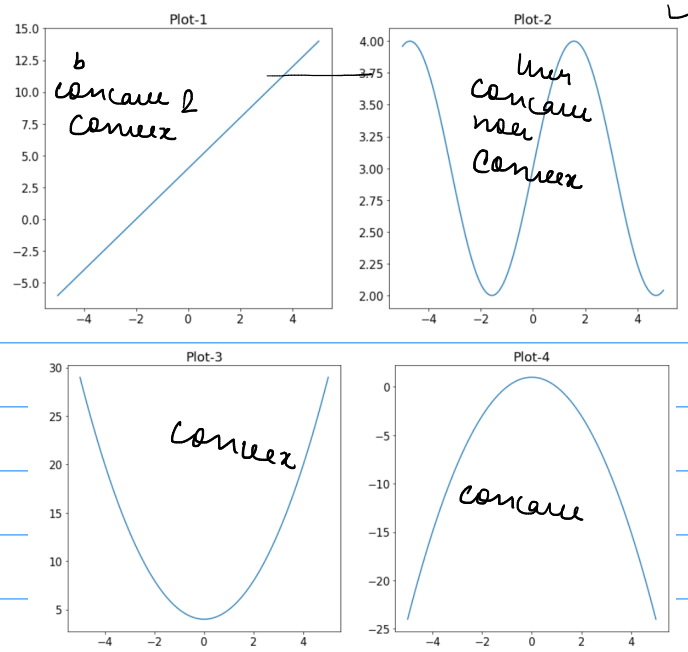
$\cdot P(z_i)$ / $P(z_i)$



- EM produces soft clustering, EM takes variances into account, EM clusters need not be voronoi regions.
- EM always converges to a local maximum of log likelihood.

Activity Question - 6

1) Match the function plotted in these figures to their corresponding types:



1 point

2) If $f(x)$ is a convex function, which of the following relations are true if a and b are two points in the domain of f?

- [ ] $f\left(\frac{a+b}{2}\right) = \frac{f(a)+f(b)}{2}$
- [ ] $f\left(\frac{a+b}{2}\right) > \frac{f(a)+f(b)}{2}$
- [x] $f\left(\frac{a+b}{2}\right) \leq \frac{f(a)+f(b)}{2}$
- [x] $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b), \quad \lambda \in [0,1]$

3) If f is a concave function and λ_ks are some parameters, select all the options that are true in the case of the Jensen's inequality, this includes the statement of the inequality as well as the constraints on the parameters.

- [x] $\sum_{k=1}^{K} \lambda_k = 1$
- [ ] $0 \leq \lambda_k \leq 1, \quad 1 \leq k \leq K$
- [ ] $\lambda_k > 0, \quad 1 \leq k \leq K$
- [x] $f\left(\sum_{k=1}^{K} \lambda_k x_k\right) \geq \sum_{k=1}^{K} \lambda_k f(x_k)$
- [ ] $f\left(\sum_{k=1}^{K} \lambda_k x_k\right) \leq \sum_{k=1}^{K} \lambda_k f(x_k)$

4) Is the following statement true or false?

$\log(x)$ is a concave function.

- (x) True
- ( ) False

Activity Question - 7

1) For a GMM with 5 mixtures and a dataset of 100 points, how many new parameters are introduced by the Jensen's inequality in the modified log-likelihood?

500

1 point

2) How do we interpret the parameter λ_k^i?

1 point

- (x) The probability of the point x_i belonging to mixture k, that is, $P(z_i = k \mid x_i)$
- ( ) The density of the point x_i given that it is from mixture k, that is, $f(x_i \mid z_i = k\ ;\ \mu_k, \sigma_k^2)$
- ( ) This parameter cannot be interpreted either as a probability or as a density.

3) What is the contribution of the i^{th} data-point to the modified log-likelihood that arises from applying the Jensen's inequality?

- ( ) $\sum_{k=1}^{K} \log\left[\pi_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \cdot \exp\left[\frac{-(x_i-\mu_k)^2}{2\sigma_k^2}\right]\right]$
- ( ) $\log\left[\sum_{k=1}^{K} \pi_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \cdot \exp\left[\frac{-(x_i-\mu_k)^2}{2\sigma_k^2}\right]\right]$
- ( ) $\sum_{k=1}^{K} \log\left[\frac{\pi_k}{\lambda_k^i} \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \cdot \exp\left[\frac{-(x_i-\mu_k)^2}{2\sigma_k^2}\right]\right]$
- (x) $\sum_{k=1}^{K} \lambda_k^i \cdot \log\left[\frac{\pi_k}{\lambda_k^i} \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \cdot \exp\left[\frac{-(x_i-\mu_k)^2}{2\sigma_k^2}\right]\right]$

4) In the context of GMMs, the modified log-likelihood is:

- ( ) an upper-bound for the log-likehood
- (x) a lower-bound for the log-likelihood

5) Consider a dataset with n points and a GMM with K mixtures. If we fix λ and maximize for θ, what is our estimates for the mean of the k^{th} mixture? Exactly two options are correct.

- [x] It is the weighted mean of the n points, where the weight for point i in mixture k is given by λ_k^i
- [ ] It is the mean of the n points
- [ ] $\hat{\mu}_k^{MML} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
- [x] $\hat{\mu}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i x_i}{\sum_{i=1}^{n} \lambda_k^i}$

6) If we fix θ and solve for λs, which of the following statements are true?

- [x] $\hat{\lambda}_k^i$ is the probability that the point x_i belongs to mixture k, that is, $P(z_i = k \mid x_i)$
- [x] $\hat{\lambda}_k^i = \frac{f(x_i \mid z_i = k\ ;\ \mu_k, \sigma_k^2) \cdot P(z_i = k)}{f(x_i)}$
- [ ] $\hat{\lambda}_k^i$ is dependent on $\hat{\lambda}_k^j$ for $i \neq j$

Activity Question - 8

1) Consider two steps in the EM algorithm at time-step $t+1$:

Step-X

$\lambda^{t+1} = \underset{\lambda}{\operatorname{argmax}} \quad \text{modified } \log L(\theta^t, \lambda)$

Step-Y

$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \quad \text{modified } \log L(\theta, \lambda^{t+1})$

Which one is the E step (expectation) and which one is the M step (maximization)?

- [x] Step-X is the E-step
- [x] Step-Y is the M-step
- [ ] Step-Y is the E-step
- [ ] Step-X is the M-step
- [ ] Both are M-steps, as there is a maximization happening in both steps

2) Is the following statement true or false?

The EM algorithm converges to the global maximum of the log-likelihood function of the GMM.

- [ ] True
- [x] False

3) After convergence of the EM algorithm (K mixtures) on a dataset with n points, let λ represent the last u parameters. Consider the following assignment to the cluster indicators:

$z_i = \underset{k}{\operatorname{argmax}} \quad \lambda_k$

Which of the following statements are true?

- [x] This is an instance of hard-clustering.
- [ ] This is an instance of soft-clustering.
- [ ] The clusters formed are Voronoi regions.
- [ ] The clusters formed are identical to the ones formed if Lloyd's algorithm is run on this dataset with K

4) What is the right orders for initializing the parameters $\theta^{(0)}$ in the EM algorithm?

(1) Compute $\pi_k^{(0)}$ as the proportion of points in cluster-k

(2) Compute $\mu_k^{(0)}$ as the mean of cluster-k

(3) Compute $\sigma_k^{2\,(0)}$ as the variance of cluster-k

(4) Run the Lloyd's algorithm on the dataset with K clusters

- [ ] (1) → (4) → (2) → (3)
- [ ] (4) → (1) → (2) → (3)
- [ ] (4) → (2) → (3) → (1)
- [ ] (1) → (2) → (3) → (4)

Practice Assignment - 1

1) Consider a dataset that has only 100 points, out of which 20 points have the value 1, 50 have value 2 and 30 have value 3. We use a categorical distribution to model this data. The parameters of the distribution are:

$p = P(x=1), \quad q = P(x=2), \quad r = P(x=3)$

If the distribution seems unfamiliar to you, think about an imaginary dice with three faces. What is the likelihood function for this data under this distribution?

- [ ] $p^{30} \cdot q^{50} \cdot r^{20}$
- [ ] $(p+q+r)^{100}$
- [x] $p^{20} \cdot q^{50} \cdot r^{30}$

2) What is the value of $p+q+r$? Enter your answer correct to two decimal places.

1.00

3) What is the maximum likelihood estimate of p? Enter your answer correct to two decimal places.

0.20

4) Consider a dataset of n data-points, $D = x_1, \cdots, x_n$. If we assume these points to have been generated from a Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, what is the expression for the log-likelihood after removing constant terms?

(1) Constant terms are those that don't depend on either μ or σ^2

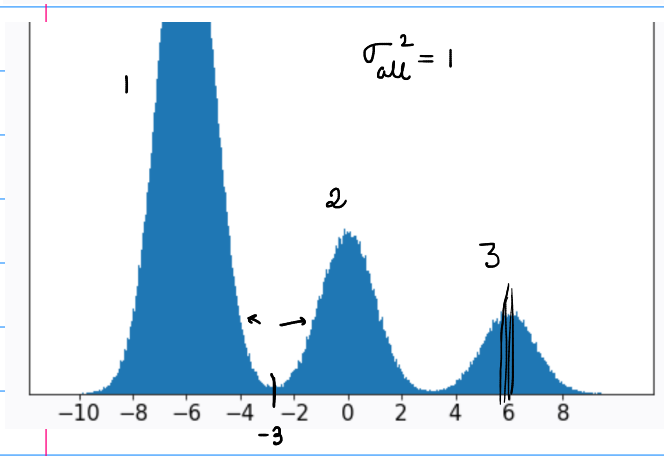
(2) $\log$ always means $\log_e$ unless otherwise specified.

- [ ] $\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$
- [ ] $-\log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$
- [x] $-n \log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$
- [ ] $n \log \sigma + \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$

5) Consider a dataset of heights of 300 individuals. The first 100 are drawn from the active pool of basketball players in the NBA. The next 100 are drawn from the list of chess grand masters. The last 100 are drawn randomly from the city of Chennai. All 300 individuals are in the age-group of 20 to 25. If we use a GMM to understand this data, what is a good choice of K, the number of mixtures?

2

Consider the histogram of one million points sampled from a GMM with three mixtures as shown in the figure below. The mixtures are labeled from left to right as 1, 2 and 3. The mean for each mixture is one of the ticks displayed on the x-axis. All the mixtures have unit variance:



6) What is the mean of mixture-3? Note that the mean is an integer here.

6

1 point

7) Which of the following could be the values of π_1, π_2 and π_3? 1 point

- [ ] $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$
- [ ] $\pi_1 = 0.4, \pi_2 = 0.1, \pi_3 = 0.4$
- [x] $\pi_1 = 0.7, \pi_2 = 0.2, \pi_1 = 0.1$
- [ ] $\pi_1 = 0.4, \pi_2 = 0.4, \pi_3 = 0.1$

8) If the point -3 is observed, what is the probability that it has come from mixture-2? Use the values of π_1, π_2, π_3 obtained from the previous question. Enter your answer correct to two decimal places.

0.22

9) Assume that you are given a set of one 10000 data-points in $\mathbb{R}$. You fit a GMM with $K = 2$ for this dataset using the EM algorithm to estimate the parameters. The EM algorithm was initialized as follows:

(1) $\mu_1 = -1, \mu_2 = 1$

(2) $\pi_1 = \pi_2 = 0.5$

(3) $\sigma_1^2 = \sigma_2^2 = 1$

The estimated means are $\hat{\mu}_1$ and $\hat{\mu}_2$ for the the two mixtures. A little while later, a domain expert comes and tells you that the dataset given to you was actually sampled from a Gaussian with mean 0 and variance 1. Which of the following options is true? Code the EM algorithm and observe what happens.

- [ ] $\hat{\mu}_1$ is very close to $\hat{\mu}_2$ but both are not close to 0
- [x] $\hat{\mu}_1$ is not close to $\hat{\mu}_2$ and neither of them is close to 0
- [ ] $\hat{\mu}_1$ is very close to $\hat{\mu}_2$ and both are close to 0

Graded Assignment - 1

1) Consider a dataset that has 10 zeros and 5 ones. What is the likelihood function if we assume a Bernoulli distribution with parameter p as the probabilistic model?

- p^{15}
- $(1-p)^{15}$
- $p^{10} \cdot (1-p)^5$
- $p^5 \cdot (1-p)^{10}$ ✓

$1 \to p$, $0 \to (1-p)$

10 0s $\to (1-p)^{10}$

5 1s $\to p^5$ $\Rightarrow p^5 \times (1-p)^{10}$

$L = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

2) In the previous question, what is the estimate of $\hat{p}_{ML}$? Enter your answer correct to two decimal places.

0.33 ✓

$\hat{p}_{ML} = \frac{\text{no. of ones}}{n} = \frac{5}{15} = 0.33$

3) Consider a dataset that has a single feature (x). The first column in the table below represents the value of the feature, the second column represents the number of times it occurs in the dataset.

x	Frequency
-1	1
0	1
2	4
4	2
5	2

$x_i f_i$: -1, 0, 8, 8, 10

Mean: $\frac{\sum x_i f_i}{\sum x_i} = \frac{25}{10} = 2.5$

If we use a Gaussian distribution to model this data, find the maximum likelihood estimate of the mean.

- 2
- 0
- 2.5 ✓
- The mean cannot be computed as the variance of the Gaussian is not explicitly specified.

4) Consider a beta prior for the parameter p of a Bernoulli distribution:

$p \sim \text{Beta}(3, 2)$ (α, β)

The dataset has 15 ones and 10 zeros. What is the posterior? (n_h, n_t)

- $\text{Beta}(3, 2)$
- $\text{Beta}(13, 17)$
- $\text{Beta}(18, 12)$ ✓
- $\text{Beta}(17, 11)$

Post. $= \text{Bet}(\alpha + n_h, \beta + n_t)$

$= \text{Bet}(3+15, 2+10)$

$= \text{Beta}(18, 12)$

5) In the previous question, we use the expected value of the posterior as a point-estimate for the parameter of the Bernoulli distribution. What is $\hat{p}$? Enter your answer correct to two decimal places.

0.60 ✓

$E = \frac{\alpha + n_h}{(\alpha+\beta)+n} = \frac{3+15}{30} = \frac{18}{30} = 0.60$

We wish to fit a GMM with $K = 2$ for a dataset having 4 points. At the beginning of the t^{th} time step of the EM algorithm, we have $\theta^{(t)}$ as follows:

$\pi_1 = 0.3, \quad \pi_2 = 0.7$

$\mu_1 = 2, \quad \sigma_1^2 = 1$

$\mu_2 = 3, \quad \sigma_2^2 = 1$

The density of the points given a particular mixture is given to you for all four points. f is the density of a Gaussian.

x_i	$f(x_i \mid z_i = 1)$	$f(x_i \mid z_i = 2)$
1	0.242	0.054
2	0.399	0.242
3	0.242	0.399
4	0.054	0.242

Use three decimal places for all quantities throughout the questions.

7) What is the value of λ_k^i for $i = 1$ and $k = 2$ after the E-step? Enter your answer correct to three decimal places.

0.342 ✓

λ_k^i for $i = 1$, $k = 2$

$\lambda_k^i = P(z_i = k \mid x_i) = \frac{P(z_i = k) \cdot f(x_i \mid z_i = k)}{f(x_i)}$

$\lambda_2^1 = P(z_1 = 2 \mid x_1) = \frac{P(z_1 = 2) \times f(x_1 \mid z_1 = 2)}{f(x_i)}$

$= \frac{0.7 \times 0.054}{0.7 \times 0.054 + 0.3 \times 0.242} = 0.342$

8) If we pause the algorithm at this stage (after the E-step) and use the λ_k^i values to do a hard-clustering, what would be the cluster assignment? We use the following rule to come up with cluster assignments:

$z_i = \underset{k}{\text{argmax}}\, \lambda_k^i$ → hard clustering

The answer is in the form of a vector: $\mathbf{z} = [z_1 \quad z_2 \quad z_3 \quad z_4]^T$.

- $[1 \quad 1 \quad 1 \quad 1]^T$
- $[2 \quad 2 \quad 2 \quad 2]^T$
- $[1 \quad 1 \quad 2 \quad 2]^T$
- $[1 \quad 2 \quad 2 \quad 2]^T$ ✓

Sol^n →

We need to compute the table of λ_k^i values from which we can read of the cluster assignments.

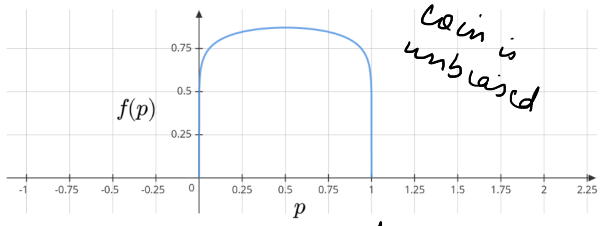
x_i	λ_1^i	λ_2^i	z_i
1	0.658	0.342	1
2	0.414	0.586	2
3	0.206	0.794	2
4	0.087	0.912	2

we calculated

~ 0.3

0.7

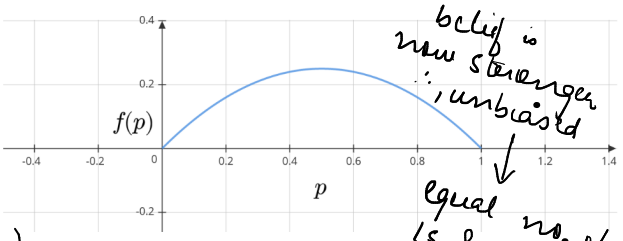
6) Consider the following prior distribution (Beta) of the parameter p of a Bernoulli distribution: *1 point*



Coin is unbiased

analyse

After observing 10 data-points, the following is the posterior distribution:



belief is now stronger, unbiased

equal no. of 1s & 0s.

Ignore the values on the Y-axis and just focus on the shapes of the distributions. Which of the following could correspond to the observed data?

- {1, 1, 1, 0, 1, 1, 0, 1, 1, 1}
- {0, 1, 0, 0, 0, 1, 0, 0, 0, 0}
- {1, 1, 0, 1, 0, 0, 0, 1, 1, 0} ✓

9) What is the value of μ_1 after the M-step? I

$$\mu_1 = \frac{\sum_{i=1}^{4} \lambda_1^i x_i}{\sum_{i=1}^{4} x_i} = \frac{0.658 \times 1 + 0.414 \times 2 + 0.206 \times 3 + 0.087 \times 4}{0.658 + 0.414 + 0.206 + 0.087} \approx 1.796$$

Values from the table (above)

✓ 1.797

10) A GMM is fit for a dataset with 5 points. At some time-step in the EM algorithm, the following are the values of λ_k^i for all points in the dataset for the k^{th} mixture after the E-step:

$\lambda_k^1 = 0.3$
$\lambda_k^2 = 0.1$
$\lambda_k^3 = 0.4$
$\lambda_k^4 = 0.8$
$\lambda_k^5 = 0.2$

What is the estimate of π_k after the M-step? Enter your answer correct to two decimal places.

$$\pi_k = \frac{1}{5} \sum_{i=1}^{n} \lambda_k^i = \frac{1}{5}(0.3 + 0.1 + 0.4 + 0.8 + 0.2) = \frac{1.8}{5}$$

$$= 0.36$$

✓ 0.36

11) What is the value of the following expression after the E-step at time-step t in the EM algorithm? There are 100 data-points and 3 mixtures.

$\sum_{i=1}^{100} \sum_{k=1}^{3} \lambda_k^i \rightarrow 1$ → 100

$$\sum_{i=1}^{n} \lambda_k^i = 1 \text{ for each data pt.}$$

since, there are 100 data pts.

- 3
- ✓ 100
- 103
- 300
- 1
- The answer depends on the time-step t we are at