Review Questions

Lecture 2

What is the angle between the two lines characterized by

$$\mathbf{w}_1 = [1, 1]^T$$

$$\mathbf{w}_2 = [1, -1]^T$$

(a) 0 (b) 45 (c) 90 (d) 120

C

Lecture 2

We stop two people at random. What is the probability that they were born on the same day of the week?
(a) $\frac{1}{7}$ (b) $\frac{1}{7^2}$ (c) $\frac{1}{7+7}$ (d) $\frac{1}{2}$ (e) None of the above

A

Lecture 3

We are given a set of 2D points from two classes, as shown in the Figure.
Q: To make the computations efficient, we want to do a dimensionality reduction from 2D to 1 D with the help of a 1×2 matrix $\mathbf{W}$

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

What should be the $\mathbf{W}$ matrix be in this case?
(Indeed the goal is to get good classification performance in the new feature space $\mathbf{x}'$, while the computations could be efficient)

(a) $[1, 0]$ (b) $[-1, 0]$ (c) $[2, 0]$ (d) $[1, 1]$ (e) $[0, 1]$ (f) $[0, -1]$ (g) $[1, 0]^T$

A,B,C

Lecture 3

We are given a set of 2D points from two classes, as shown in the Figure. We want to "rorate" the data so that points are spread across first (x) axis (i.e., something like rotate clockwise by 45°)

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

What should be the 2×2 matrx $\mathbf{W}$ be in this case?

(a) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$

(f) $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ (g) Any one of the above (h) None of the above

F,E

Lecture 3

What is the probability that, in a room of 30 people, there is a pair of people who have the same birthday. (use calculator, pick the closest)
(a) 0.11 (b) 0.31 (c) 0.51 (d) 0.71 (e) 0.91

D

Lecture 4

Confusion matrix is:

(a) Square (b) Always Diagonal (c) Can never be diagonal (d) Can be diagonal (e) Always Symmetric (f) Can never be symmetric (g) Can be Symmetric

A,D,G

Lecture 4

$$\frac{TP}{P}$$

is known as:
(a) Accuracy (b) Precision (c) Recall (d) None of the above

C

Lecture 4

A disease occurs with a probability of 0.4 (i.e., it is present in 40% of the population). You have a test that detects the disease with a probability 0.6, and produces a false positive with probability of 0.1. What is the (posterior) probability that the test comes back positive.

Hint: S is the event that you are sick; P is the event that test comespositive.

$$P(S|P) = \frac{P(P|S)P(S)}{P(P)} = \frac{P(P|S)P(S)}{P(P|S)P(S) + P(P|\bar{S})P(\bar{S})}$$

(a) 0.6 (b) 0.7 (c) 0.8 (d) 0.9 (e) 0.95

C

Lecture 4

Two SMAI students (Raju and Sheela) worked on the same problem with the same measurements/features and samples, except that their feature orderings were different. (i.e., $\mathbf{x}$ and $\mathbf{x}'$ were permutations.) Identify correct statement(s).

- (a) Both got the same accuracy with KNN (same K and Eucli. distance)
- (b) Both got different accuracy with KNN (same K and Eucli. distance)
- (c) Their confusion matrices were different i.e., elements (cells) were swapped.
- (d) Both had the same Covariance Matrices (Hint: $\Sigma = \frac{1}{N}\sum_{i=1}^{N}[\mathbf{x}_i - \mu][\mathbf{x}_i - \mu]^T$)
- (e) Both had covariance matrices of the same Rank.
- (f) Both had covariance matrices where cells (elements) were swapped.

A,E,F

Lecture 5

We know that the rank of a 3×3 matrix formed by first 9 numbers arranged sequentially is 2.

What is the rank of a 5×5 matrix formed by first 25 numbers arranged sequentially?

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5 (f) none of the above

B

Lecture 5

A certain test for disease is known to have True positive of 0.6 and False Positive of 0.1.
A population of 100 people (where 60 of them are infected) undergoes this test.

What could be the confusion matrix?

(a) $\begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}$ (b) $\begin{bmatrix} 0.6 & 0.4 \\ 0.9 & 0.1 \end{bmatrix}$ (c) $\begin{bmatrix} 0.6 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}$ (d) $\begin{bmatrix} 0.58 & 0.42 \\ 0.15 & 0.85 \end{bmatrix}$
(e) None of the above

E(A)

Covariance Matrix:

- (a) Can neve be diagonal.
- (b) Can be diagonal
- (c) Always Positive Semi Definite
- (d) Always full rank
- (e) Never full rank.
- (f) Always Symmetric
- (g) Not guaranteed to be symmetric

B,C,F

Lecture 5

Let $\mu = \sum_{i=1}^{N} \mathbf{x}_i$ be the mean of $\mathbf{x}_1, \ldots, \mathbf{x}_N$. $\mathbf{x}_i \in R^d$

- (a) μ is also $\in R^d$
- (b) μ is always one of the N samples.
- (c) μ can not be one of the N samples
- (d) μ can be one of the N samples.
- (e) μ is equidistant from all the N samples
- (f) μ has the least sum of square error from all the samples (i.e., μ is $\arg\min_{\mathbf{y}} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}||_2^2$)

A,D,F

Lecture 5

You are planning a picnic today, but the morning is cloudy

- 50% of all rainy days start off cloudy.
- But cloudy mornings are common (about 40% of days start cloudy)
- This is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

(a) 10% (b) 12.5% (c) 15% (d) $>$ 20% (e) $<$ 20%

Hint: Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

B,E

Lecture 6

If

$$Ax = \lambda x$$

then What are the eigen values and eigen vectors of A^2 i.e., Find:

$$A^2x = ?$$

- (a) λ, x
- (b) λ^2, x
- (c) λ, $2x$
- (d) λ^2, $2x$
- (e) none of the above

B,D

Lecture 6

Consider

$$\mathbf{x}'_i = \mathbf{W}\mathbf{x}_i$$

$\mathbf{W}$ is constructed as below:

- We start with a $d \times d$ identity matrix
- We randomly permute (rearrange) the columns
- We remove half of the rows and create a $\frac{d}{2} \times d$ matrix $\mathbf{W}$

The process of creation of new representation is:

- (a) Feature subset selection; A random subset of the original features will be in $\mathbf{x}'$
- (b) Feature extraction; New features are linear combination of old ones, and not really a subset.
- (c) Dimensionality Reduction; New representation has smaller dimension than the original one.
- (d) This can not be done since these operations are illegal (or mathematically not defined).

A,C

Lecture 6

Consider three sets

$$A = \{1, 3, 4, 5, 6, 7, 8\}$$

$$B = \{2, 4, 6, 8\}$$

$$C = \{1, 2, 3, 4, 5\}$$

We know Jacard index ($J = \frac{|A \cap B|}{|A \cup B|}$) as a good measure of similarity. Let us use $1 - J$ as a distance.

Does $1 - J$ obey triangular inequality for this set? [4]

(Hint Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$)

- (a) YES
- (b) NO
- (c) Triangular inequality is not applicable for this problem.
- (d) Can not be computed.

A

Lecture 6

Consider the problem of feature transformation as

$$\mathbf{x}'_i = \mathbf{W}\mathbf{x}_i$$

If $\mathbf{x}_i \in R^2$ and $\mathbf{W}$ is 2×2 matrix with rank as 1, then

the new points $\mathbf{x}'_i$

- lie on a line in 2D
- are also R^2
- undefined
- One coordinate (dimension) of all the $\mathbf{x}'_i$ will be always the same
- all points in 2D will collapse into a single point. (i.e., $\mathbf{x}'_i = \mathbf{x}'_j$ for all i, j)
- none of the above.

A,B

In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize,and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether the you want to SWITCH their choice to the other unopened door, or STICK to their original choice.
[5]

What should we advise? What is the prob. of win if the candidate switch:

- (a) $\frac{1}{3}$ since all the doors are equally likely. Don't switch
- (b) $\frac{1}{2}$ since there are only two left, both are equally likely, no advantage in switching.
- (c) $\frac{2}{3}$. Prefer switching. Bayes says so.
- (d) $\frac{1}{3}$. Don't switching. Bayes says so.
- (e) None of the above.

[5] A very popular problem on internet from khan academy to mit lecture notes!. Appreciate the role of evidence, specially if the answer is not intuitive.

C

Lecture 7

Consider a square matrix $\mathbf{A}$ with eigen values λ_i and eigen vectors $\mathbf{v}_i$. Then for $\mathbf{A}^T$,

- (a) Eigen values and eigen vectors are the same as that of $\mathbf{A}$.
- (b) Eigen values are the same. Eigen vectors are $\mathbf{v}^T$.
- (c) Eigen values are $\frac{1}{\lambda_i}$
- (d) We can comment for symmetric matrix $\mathbf{A}$. But not for other square matrices.
- (e) None of the above.

D

Lecture 7

Consider the problem of finding the minima of $f(x) = x^2 - 2x + 4$

$$\min_x x^2 - 2x + 4$$

The optimum value of the function f^* is:

- (a) 1
- (b) 2
- (c) 3
- (d) 4
- (e) -2
- (f) none of the above

C

Consider a data matrix $\mathbf{X}$ where every column is mean subtracted samples. We compute $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathsf{T}}$. Then $\mathbf{A}$

- is a $d \times d$ matrix
- is a $N \times N$ matrix
- is a symmetric matrix
- is a scaled version of the covariance matrix
- none of the above

A,C,D

The inverse of an upper triangular matrix is:

- (a) Upper triangular
- (b) Lower triangular
- (c) Need not be triangular.
- (c) Will not exist
- (d) None of the above

A

Lecture 7

The the eigen values of $\mathbf{A}$ are λ_i, what are the eigen values of $\alpha\mathbf{A}$, where α is a scalet.

- (a) λ_i itself.
- (b) $\frac{\alpha}{\lambda_i}$
- (c) $\alpha\lambda_i$
- (d) $\frac{1}{\alpha}\lambda_i$
- (e) None of the above.

C

Lecture 8

A man is known to speak truth 2 out of 3 times. He throws a die and reports that number obtained is a four. Find the probability that the number obtained is actually a four.

(A) $\frac{1}{7}$

(B) $\frac{2}{7}$

(C) $\frac{3}{7}$

(D) $\frac{4}{7}$

B

Lecture 8

We know that the optimal classifier for two equally probable (equal Prior probability) classes (days of months) $N(23, \sigma^2)$ and $N(33, \sigma^2)$ is 28.

If the variance of the second class becomes double, then the the optimial classification threshold:

(A) increase from 28

(B) decrease from 28

(C) remain unchanged

(D) can not be predicted from the data

B (D)

Lecture 8

Suppose a disease is prevalent in 1% of the population. Its medical diagnosis is 90% accurate in both directions. Given that a person tested positive, what is the chance, he actually has the disease (rounded to nearest integer)?

(A) 98%

(B) 88%

(C) 8%

(D) 1%

C

The system of linear equations

$$4x + 2y = 7$$

$$2x + y = 6$$

has

(A) a unique solution

(B) no solution

(C) an infinite number of solutions

(D) exactly two distinct solutions

B

Rank of the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 0 & 1 & 2 \end{bmatrix}$$

is

(A) 0

(B) 1

(C) 2

(D) 3

C

Lecture 9

$X \sim N(0,1)$, $Y \sim N(1,1)$ and $Z = X + Y$. Then,

(A) $Z \sim N(0,2)$

(B) $Z \sim N(0,1)$

(C) $Z \sim N(1,1)$

(D) $Z \sim N(1,2)$

(E) None of the above

D

Lecture 9

Product of Eigen values of a real square matrix is:

(A) Determinant

(B) Rank

(C) Trace

(D) non-Negative

(E) None of the above

A

Lecture 9

A and B are two independent events such that $P(\overline{A}) = 0.4$ and $P(A \cap B) = 0.2$ Then $P(A \cap \overline{B})$ is equal to

(A) 0.4

(B) 0.2

(C) 0.6

(D) 0.8

(E) None of the above

A

Lecture 9

If $\mathbf{A}$ is a $n \times n$ matrix, with every pair of columns orthogonal i.e., $\mathbf{a_i} \cdot \mathbf{a_j} = \mathbf{0}$ $\forall i, j$ and $||\mathbf{a_i}|| = 1$. Then:

(A) $\mathbf{A}^{-1} = \mathbf{A}^T$.

(B) $\mathbf{A}\mathbf{A}^T = \mathbf{I}$

(C) $\mathbf{A}\mathbf{A}^T$ has only one 1 in every column and all others zero.

(D) $\mathbf{A}^{-1}$ has only one 1 in every column and all others zero.

(E) none of the above

A,B,C

Lecture 9

Consider a matrix A of size $m \times n$. Rank of A is (choose one one most correct answer)

(A) $\leq \min(m, n)$

(B) $\leq \max(m, n)$

(C) $\geq \min(m, n)$

(D) $\geq \max(m, n)$

(E) $\frac{m+n}{2}$

A

Lecture 10

Consider a set of general vectors $\mathbf{a}_i \in R^d$. (assume all elements are some random numbers in the range of $[0, 1]$) $\mathbf{b}$ is another such vector. Consider the matrix:

$$\mathbf{A} = \sum_{i=1}^{k} 10^i \mathbf{a}_i \mathbf{a}_i^T + \sum_{i=k+1}^{d} 10^{-i} \mathbf{b}\mathbf{b}^T$$

What is the effective rank of $\mathbf{A}$

(A) d

(B) none of the above

(C) k

(D) 1

(E) $k+1$

C

Lecture 10

Consider a set of general vectors $\mathbf{a}_i \in R^d$. (assume all elements are some random numbers in the range of $[0, 1]$) $\mathbf{b}$ is another such vector. Consider the matrix:

$$\mathbf{A} = \sum_{i=1}^{k} 10^{-k} \mathbf{a}_i \mathbf{a}_i^T + \sum_{i=k+1}^{d} \mathbf{b}\mathbf{b}^T$$

What is the effective rank of $\mathbf{A}$

(A) 1

(B) none of the above

(C) k

(D) $k+1$

(E) d

C,(A)

Lecture 10

If $A = UDV^T$, then A^TA is:

(A) A square matrix

(B) UD^2U^T

(C) VD^2V^T

(D) none of the above

(E) is always full rank

A,C

Lecture 10

Consider a set of general vectors $\mathbf{a}_i \in R^d$. (assume all elements are some random numbers in the range of $[0, 1]$) $\mathbf{b}$ is another such vector. Consider the matrix:

$$\mathbf{A} = \sum_{i=1}^{k} \mathbf{a}_i \mathbf{a}_i^T + \sum_{i=k+1}^{d} \mathbf{b}\mathbf{b}^T$$

What is the effective rank of $\mathbf{A}$

(A) none of the above

(B) k

(C) $k+1$

(D) 1

(E) d

D(C)

Lecture 10

QUESTION DETAILS

Back

Instructions Bulk uploaded assignment for class review

Type Multiple choice Questions (MCQ)

Question Answer the question
Click to zoom.

Consider a set of general vectors $\mathbf{a}_i \in R^d$. (assume all elements are some random numbers in the range of $[0, 1]$) $\mathbf{b}$ is another such vector. Consider the matrix:

$$\mathbf{A} = \sum_{i=1}^{k} 10^{-i} \mathbf{a}_i \mathbf{a}_i^T + \sum_{i=k+1}^{d} 10^{i} \mathbf{b}\mathbf{b}^T$$

What is the effective rank of $\mathbf{A}$

(A) k

(B) 1

(C) none of the above

(D) d

(E) $k+1$

C(B)

Lecture 11

Consider a vocabulary of size d. One hot representation of a word i is "1" at the location (index) corresponding to that word and zero else where. Given a document that contains P words, $\mathbf{w_1}, \ldots, \mathbf{w_P}$, we compute

$$\mathbf{x} = \sum_{i=1}^{P} \mathbf{w}_i$$

Then,

(A) $\mathbf{x}$ is in R^P independent of the vocabulary size.

(B) $\mathbf{x}$ is in R^d independnet of the number of words in the document.

(C) $\sum_i x_i$ is P (x_i is the i th element of $\mathbf{x}$)

(D) $\mathbf{x}$ is the histogram of the words, with x_i as the frequency of i th word.

(E) $\mathbf{x}$ is the probability distribution with x_i as the probability of words in that document.

B,C,D

Lecture 11

Consider a document is represented by a histogram of the words in the document. **h**. i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P1). P1 is defined as permuting sentences in a document and rewriting a sentence by permuting the words.

(A) **h** is not invariant to the P1

(B) **h** is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

(C) **h** is invariant to the P1

(D) a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a".)

C,D

Lecture 11

Consider a document is represented by a histogram of the words in the document $\mathbf{h}$ i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P2). P2 is defined as replacing a set of words by their synonyms.

(A) a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a")

(B) $\mathbf{h}$ is invariant to the P2

(C) $\mathbf{h}$ is not invariant to the P2

(D) $\mathbf{h}$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

C,A

Lecture 11

A professor suspected that students while submiting home works are doing the paraphasing operations i.e., both P1 and P2. This resulted in failure of some similarity tests.

Professor designs a $d \times d$ word similarity matrix $\mathbf{S}$ such that $\mathbf{S}_{ij} = \mathbf{S}_{ji} = 1$ if words i and j are synonyms and zero else. (Note: d is the size of vocabulary).

Now to compare two documents, professor multiplies the histogram representations by $\mathbf{S}$.

$$\mathbf{h}'_i = \mathbf{S}\mathbf{h}_i$$

(Note: $\mathbf{h}'_i$ is the new representation. Also, note, after multiplying with the $\mathbf{S}$, the dimension does not change)

(A) the idea is worth, but then $\mathbf{S}$ should not have made symmetric. with only one of $\mathbf{S}_{ij}$ or $\mathbf{S}_{ji}$ as 1. The method could have worked as expected.

(B) the new representation is invariant under the operation $P1$ and $P2$. (i.e., All the plagiarism now will be detected.)

(C) the new representation helps for detecting people who have paraphrased with P2. But now it fails for the documents that were not paraphrased (like the original ones/sincere students!).

(D) the new representation is not invariant for $P2$ and it does not help.

)ption A

B

Lecture 11

We want to compare two documents i and j which are represented as histogram (popular known as bag of words) of words h_i and h_j.

Here is what four students argued:

(A) Cosine distance is a popular distance to compare two documents using this representation.

(B) we should remove the stop words (common words in the language) from the sentence so that the comparison will be more useful. Two documents have the same number of 'the' does not mean any useful similarity between them.

(C) Since these are probability distributions, KL-Divergence (may be a symmetric one) is an ideal candidate to compare.

(D) histograms should be normalized by dividing by the number of words in the document so that the comparison operation becomes "some what invariant" to another linguistic operation: "summarization".

A,B,C,D(a,b,d)

Lecture 12

Class Review 8, Question 1

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathcal{L}_1 = \mathcal{L}_2$ for all $\mathbf{w}$, then

(A) $A_{ii} = \frac{1}{\alpha_i}$ else zero

(B) $A_{ii} = \alpha_i$ else zero

(C) A is a diagonal matrix

(D) $A_{ij} = \alpha_i \cdot \alpha_j$

(E) none of the above

B,C

Lecture 12

Class Review 8, Question 2

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

See $\mathcal{L}_2$ closely,

(A) When A is PD, we can do cholesky decomposition of A as LL^T and an equivalent formulation is possible in $\mathcal{L}_1$ is each sample getting transformed as $\mathbf{L}^T\mathbf{x}_i$ (as in LMNN/Metric Learning)

(B) When A is a diagonal matrix, this is equivalent to weighing each sample independently.

(C) When A is not a diagonal matrix, this loss does not make any sense. Don't use.

(D) When A is a rank deficient matrix, an equivalent formulation is possible in $\mathcal{L}_1$ with a dimensionality reduction (this could be proved with eigen decomposition).

(E) None of the above

A,B,D

Class Review 8, Question 3

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 \neq \lambda_2 \neq 0$, then

(A) The smaller the lambda, the better the solution

(B) The optimal parameters $\mathbf{w}^*$ is independent of λ_i.

(C) The larger the lambda, the better the solution.

(D) When lambda is nonzero (positive), loss will increase (since $g(w)$ is also positive in practice), better to use $\lambda = 0$.

(E) None of the above.

E

Lecture 12

Class Review 8, Question 4

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 = \lambda_2 = 1$, then

(A) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(B) At the optima, value of the losses are same. i.e., $\mathcal{L}_1^* = \mathcal{L}_2^*$

(C) The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(D) Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(E) none of the above

B,C,D

Lecture 12

Class Review 8, Question 5

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 2$ for all i, and $\lambda_1 = \lambda_2 = 0$, then

(A) Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(B) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(C) At the optima, value of the losses are same. $\mathcal{L}_1^* = \mathcal{L}_2^*$

(D) The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(E) none of the above

D

Class Review 9, Question 1

Consider the problem of finding a solution to the following equation:

$$3w_1 + 4w_2 = 12$$

the line crosses the axes $w1$ and $w2$ respectively at:

(A) 6 and 6

(B) 3 and 4

(C) 12 and 12

(D) 4 and 3

(E) None of the above

D

Lecture 13

Class Review 9, Question 2

Consider the vector $\mathbf{w} = [w_1, w_2]^T$ and the objective function to be minimized as:

$$\min_{\mathbf{w}} (3w_1 + 4w_2 - 12)^2 + \lambda g(\mathbf{w})$$

If $g(\mathbf{w})$ is L0 norm of $\mathbf{w}$, and $\lambda = 1$, what is the optimal value of $\mathbf{w}$

(A) $[1, 1]^T$

(B) $[4, 0]^T$

(C) $[0, 3]^T$

(D) $[3, 4]^T$

(E) None of the above

B,C

Lecture 13

Class Review 9, Question 3

Consider the vector $\mathbf{w} = [w_1, w_2]^T$ and the objective function to be minimized as:

$$\min_{\mathbf{w}} (3w_1 + 4w_2 - 12)^2 + \lambda g(\mathbf{w})$$

If $g(\mathbf{w})$ is L1 norm of $\mathbf{w}$ and $\lambda = 2$, what is the optimal value of $\mathbf{w}$ (if the true answer is very close to one given, do round/approximate for simplifying the answer here)

(A) $[0, 3]^T$

(B) $[4, 0]^T$

(C) $[3, 4]^T$

(D) $[1, 1]^T$

(E) None of the above

A

Lecture 13

Class Review 9, Question 4

Consider the vector $\mathbf{w} = [w_1, w_2]^T$ and the objective function to be minimized as:

$$\min_{\mathbf{w}} (3w_1 + 4w_2 - 12)^2 + \lambda g(\mathbf{w})$$

If $g(\mathbf{w})$ is L1 norm of $\mathbf{w}$ and $\lambda = 1$, what is the optimal value of $\mathbf{w}$ (if the true answer is very close to one given, do round/approximate for simplifying the answer here)

(A) $[4, 0]^T$

(B) $[1, 1]^T$

(C) $[0, 3]^T$

(D) $[3, 4]^T$

(E) None of the above

C

Lecture 13

Class Review 9, Question 5

Consider the vector $\mathbf{w} = [w_1, w_2]^T$ and the objective function to be minimized as:

$$\min_{\mathbf{w}} (3w_1 + 4w_2 - 12)^2 + \lambda g(\mathbf{w})$$

If $g(\mathbf{w})$ is L2 norm of $\mathbf{w}$ and $\lambda = 1$, what is the optimal value of $\mathbf{w}$

(A) $[0, 3]^T$

(B) $[3, 4]^T$

(C) $[4, 0]^T$

(D) $[1, 1]^T$

(E) None of the above

E

Class Review 10, Question 1

Given a set of 2D points X on a line that makes 45 degree to the x-axis:

$$X = \{[1,1]^T, [2,2]^T, [3,3]^T, [4,4]^T, [5,5]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) $\lambda_1 = -1$

(B) $\lambda_1 = \lambda_2$

(C) Σ is singular

(D) $\lambda_2 = 0$

(E) none of the above

C,D

Lecture 14

Class Review 10, Question 2

Given a set of 2D points X on a line that makes 45 degree to the x-axis:

$$X = \{[-2,2]^T, [-3,3]^T [-4,4]^T, [-5,5]^T [-6,6]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) $\lambda_2 = 0$

(B) $\lambda_1 = -1$

(C) Σ is singular

(D) $\lambda_1 = \lambda_2$

(E) none of the above

A,C

Class Review 10, Question 3

Given a set of 2D points X on the vertical line $x_2 = 5$,

$$X = \{[1,5]^T, [2,5]^T, [3,5]^T, [4,5]^T, [5,5]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) $\lambda_1 \geq \lambda_2$

(B) μ is on the same line.

(C) Σ is singular

(D) Σ is diagonal

(E) None of the above

A,B,C,D

Class Review 10, Question 4

Set X has 10 points. 5 of them are on a line that makes 45 degrees with the x_1 axis and another 5 from on a line that makes 135 degrees with the x_1 axis. We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) Σ is singular

(B) Σ is diagonal

(C) $\lambda_1 = \lambda_2 \neq 0$

(D) μ is on either of these lines.

(E) None of the above

E

Class Review 10, Question 5

Given a set of 2D points X on the vertical line $x_1 = 5$,

$$X = \{[5,1]^T, [5,2]^T, [5,3]^T, [5,4]^T, [5,5]^T\}$$

We now add an additional point $[4,3]^T$ to X. We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) $\lambda_1 \geq \lambda_2$

(B) Σ is singular

(C) Σ is diagonal

(D) $\mathbf{u}_1$ and $\mathbf{u}_2$ are nearly orthogonal, but not perfectly orthogonal.

(E) None of the above.

A,C

Class Review 11, Question 1

Consider X to be a square matrix of size $n \times n$ and $X = UDV^T$.

(A) If rank(X) $= k$, D has k zeros in diagonal

(B) If rank(X) $= n$, D has all non-zero entries in diagonal.

(C) if rank(X) $= n$ but $|A|$ is a very small number then, D takes the form $D = diag(d_1, d_2, .., \epsilon)$ where ϵ is a very small number

(D) If rank(X) $= k$, D has $n - k$ zeros in diagonal

(E) None of these

B,C,D

Class Review 11, Question 2

Consider X to be a square matrix of size $n \times n$ and $X = UDV^T$.

(A) Both X^TX and XX^T have the same eigenvectors

(B) D^2 contains the eigenvalues of X^TX on its diagonal

(C) D contains the eigenvalues of X^TX on its diagonal

(D) X, XX^TX and XX^T have the same eigenvalues

(E) Both X^TX and XX^T have the same eigenvalues

(F) None of these

B,E

Lecture 15

Class Review 11, Question 3

Suppose you want to apply PCA to your data X which is in 2D and you decompose X as UDV^T. Then,

(A) PCA can be useful if all elements of D are equal

(B) D is not full-rank if all points in X lie on a circle

(C) D is not full-rank if all points in X lie on a straight line

(D) V is not full-rank if all points in X lie on a straight line

(E) PCA can be useful if all elements of D are not equal

(F) None of these

E,C

Class Review 11, Question 4

Let $X = UDV^T$. Then

(A) Columns of U are eigenvectors of X^TX

(B) Columns of V are eigenvectors of X^TX

(C) Rows of V are eigenvectors of X^TX

(D) Rows of U are eigenvectors of X^TX

(E) None of these

B

Class Review 11, Question 5

Let $X = UDV^T$. Then

(A) Rows of V are eigenvectors of XX^T

(B) Rows of U are eigenvectors of XX^T

(C) Columns of V are eigenvectors of XX^T

(D) Columns of U are eigenvectors of XX^T

(E) None of these

D

Class Review 12, Question 1

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

In each iteration, we modify the learning rate as $\eta^{k+1} \leftarrow 0.8\eta^k$.

Consider a training set of 5 positive and 5 negative samples. They are not linearly separable. Then:

(A) This algorithm will converge.

(B) Error (number of Mis-classification) on the training data is guaranteed to be zero.

(C) This algorithm will oscillate.

(D) This algorithm will not converge.

(E) none of the above

A

Class Review 12, Question 2

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

Consider a training set of 5 positive and 5 negative samples. They are not linearly separable.

(A) Irrespective of the initalization, implementation will oscillate/cycle.

(B) Assume the termination criteria was "if there is no change in the mis-classification rate across two iterations, terminate". Even then the implementation will never converge.

(C) Assume the termination criteria was "if there is no change in the misclassification rate across two iterations, terminate". Then the implementation could have converged.

(D) Since the problem is non-convex, if we can find a right initialization, we will get the best solution very fast.

(E) None of above.

A,C

Lecture 16

Class Review 12, Question 3

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

Consider a training set of 5 positive and 5 negative samples. They are linearly separable.

We run this implementation 10 times as:

(A) This algorithm will converge to the same solution irrespective of the relative ordering of the data. (i.e., data set was shuffled across runs) (assume initialization is same in all cases and η is fixed as 0.1)

(B) This algorithm will converge to the same solution irrespective of the learning rate (say in the range 0.05 to 0.2). (assume the initialization is same in all cases.)

(C) This algorithm will converge to the same solution irrespective of the initialization.(assume learning rate fixed at 0.1)

(D) This algorithm will converge to a valid solution irrespective of the initialization. (assume learning rate fixed at 0.1)

(E) This algorithm will converge to a valid solution irrespective of the learning rate (say in the range 0.05 to 0.2). (assume the initialization is same in all cases.)

A,D,E

Lecture 16

Class Review 12, Question 4

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

Consider a training set of 5 positive and 5 negative samples. They are not linearly separable. We also have 10 test samples (5 each from both classes). They are linearly separable.

(A) This algorithm will converge.

(B) Error (number of Mis- classification) on the test data is guaranteed to be zero.

(C) Error (number of Mis-classification) on the training data is guaranteed to be zero.

(D) This algorithm will not converge.

(E) none of the above

D

Lecture 16

Class Review 12, Question 5

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

Consider a training set of 5 positive and 5 negative samples. They are linearly separable. We also have 10 test samples (5 each from both classes). They are also linearly separable.

(A) Error (number of Mis- classification) on the test data is guaranteed to be zero.

(B) Error (number of Mis-classification) on the training data is guaranteed to be zero.

(C) Perceptron algorithm will converge.

(D) Perceptron algorithm will not converge.

(E) none of the above

B,C

Lecture 17

Class Review 13, Question 1

We want to do PCA using gradient descent. Assume that Σ is the covariance matrix, η is the learning rate. Then the update rule is

(A) $u_{k+1} = (I - \eta\Sigma)u_k$

(B) $u_{k+1} = (I + \eta\Sigma)u_k$

(C) $u_{k+1} = \eta\Sigma u_k$

(D) None of these

Ans - B

Class Review 13, Question 2

PCA solves this problem:

$$\max_u u^T \Sigma u - \lambda(u^T u - 1)$$

where Σ is the covariance matrix. Which of the following are true regarding PCA

(A) Sum of variances captured by all eigenvectors is $\text{tr}(\Sigma)$

(B) If all data points are on a line then at least one of the eigenvalues is 1

(C) If all data points are on a line then at least one of the eigenvalues is 0

(D) λ is the variance captured by the eigen vector u

Ans – A,C,D

Lecture 17

Class Review 13, Question 3

Consider we are dong PCA to go from R^2 data to R^1. Consider each point is denoted by (X_i, Y_i). Then in which of these situations will PCA work reasonably well:

(A) $X_i^2 + Y_i^2 = 10$

(B) $Y_i = X_i + 10$

(C) $Y_i = X_i + 10 + \epsilon_i$ where $\epsilon_i \sim N(0,1)$

(D) $X_i^2 + Y_i^2 <= 10$

(E) None of these

Ans – B,C

Lecture 17

Class Review 13, Question 4

Consider we have data in R^2. Then the linear regression line and the PCA line

(A) will always be the same

(B) will never be the same

(C) can sometimes be the same

(D) None of these

Ans - C

Lecture 17

Class Review 13, Question 5

Consider we are using PCA to compress face images using top K eigenvectors and then we do the reconstruction. Then

(A) Reconstruction will be good for non-face images (say buildings)

(B) Compression (for face images) is lossless

(C) Reconstruction will be bad for non-face images (say buildings)

(D) Compression (for face images) is lossy

(E) None of these

Ans – C,D

Lecture 18

Class Review 14, Question 1

Consider the sigmoid function $g(\alpha, z) = \frac{1}{1+e^{-\alpha z}}$ where α is a positive real number.

(A) if $\alpha_1 > \alpha_2$, then $g(\alpha_1, z) \geq g(\alpha_2, z)$ for all z in the range $[1, 2]$

(B) if $\alpha_1 > \alpha_2$, then $g(\alpha_1, z) \geq g(\alpha_2, z)$ for all z

(C) if $\alpha_1 > \alpha_2$, then $g(\alpha_1, z) \geq g(\alpha_2, z)$ for all z in the range $[-2, -1]$

(D) if $\alpha_1 > \alpha_2$, then $g(\alpha_1, z) \geq g(\alpha_2, z)$ for all z in the range $[-1, 1]$

(E) if $\alpha_1 > \alpha_2$, then $g(\alpha_1, z) \leq g(\alpha_2, z)$ for all z

A

Lecture 18

Class Review 14, Question 2

You know the popular sigmoid function $g(z) = \frac{1}{1+e^{-z}}$, and also the $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

(A) when $z = 0$, $tanh(z)$ is 0.5.

(B) $tanh(z) = 2g(2z) - 1$

(C) $tanh(z)$ is in the range of $[-1, +1]$

(D) $tanh(z)$ is in the range of $[0, 1]$

(E) when $z = 0$, $tanh(z)$ is 0.

B,C,E(,D)

Lecture 18

Class Review 14, Question 3

Consider the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$ Then $g'(z)$ i.e., derivative of $g(z)$ with respect to z

(A) $g(z)(1 - g(z))$

(B) is always positive for all values of z

(C) $\frac{e^{-z}}{(1+e^{-z})^2}$

(D) is constant, i.e., derivative is independent of z.

(E) $\frac{1}{1+e^{z}}$

A,B,C

Lecture 18

Class Review 14, Question 4

Consider the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$ Then $1 - g(z)$ is

(A) is in the range of $[-1, +1]$.

(B) $\frac{e^{-z}}{1+e^{-z}}$

(C) is in the range of $[-1, 0]$.

(D) is in the range of $[0, 1]$.

(E) $\frac{1}{1+e^{z}}$

A,B,D,E

Lecture 18

Class Review 14, Question 5

Consider the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$

(A) when z is negative, $g(z)$ is also negative.

(B) $g(z)$ is always in the range of $[0, 1]$

(C) $g(z)$ is always in the range of $[-1, 1]$

(D) when $z = 0$, $g(z) = 0.5$

B,C,D

Lecture 19

Class Review 15, Question 1

Consider a two input (i.e., x_1 and x_2 neuron with an activation $\phi(z)$.

Consider $\{-1, +1\}$ logic, and $\phi(z) = +1$ if $\mathbf{w}^T\mathbf{x} \geq -1$ else -1

For what values of w_1 and w_2, the neuron will act as AND?

(A) $w_1 = -1$ and $w_2 = -1$

(B) $w_1 = -2$ and $w_2 = -2$

(C) $w_1 = 1$ and $w_2 = 1$

(D) $w_1 = 2$ and $w_2 = 2$

(E) None of the Above

E

Lecture 19

Class Review 15, Question 2

Consider a two input (i.e., x_1 and x_2 neuron with an activation $\phi(z)$.

Consider $\{-1, +1\}$ logic, and $\phi(z) = +1$ if $\mathbf{w}^T\mathbf{x} \geq 1$ else -1.

For what values of w_1 and w_2, the neuron will act as AND?

(A) $w_1 = 2$ and $w_2 = 2$

(B) $w_1 = -1$ and $w_2 = -1$

(C) $w_1 = 1$ and $w_2 = 1$

(D) $w_1 = -2$ and $w_2 = -2$

(E) None of the Above

A,C

Lecture 19

Class Review 15, Question 3

Consider a two input (i.e., x_1 and x_2 neuron with an activation $\phi(z)$.

Consider $\{-1, +1\}$ logic, and $\phi(z) = +1$ if $\mathbf{w}^T\mathbf{x} \geq 1$ else -1.

For what values of w_1 and w_2, the neuron will act as OR?

(A) $w_1 = -2$ and $w_2 = -2$

(B) $w_1 = 2$ and $w_2 = 2$

(C) $w_1 = 1$ and $w_2 = 1$

(D) $w_1 = -1$ and $w_2 = -1$

(E) None of the Above

E

Lecture 19

Class Review 15, Question 4

Consider a two input (i.e., x_1 and x_2 neuron with an activation $\phi(z)$.

Consider $\{0, +1\}$ logic, and $\phi(z) = +1$ if $\mathbf{w}^T\mathbf{x} \geq 1$ else 0

For what values of w_1 and w_2, the neuron will act as AND?

(A) $w_1 = 1$ and $w_2 = 1$

(B) $w_1 = 2$ and $w_2 = 2$

(C) $w_1 = -2$ and $w_2 = -2$

(D) $w_1 = -1$ and $w_2 = -1$

(E) None of the Above

E

Lecture 19

Class Review 15, Question 5

Consider a two input (i.e., x_1 and x_2 neuron with an activation $\phi(z)$.

Consider $\{-1, +1\}$ logic, and $\phi(z) = +1$ if $\mathbf{w}^T\mathbf{x} \geq 1$ else -1

For what values of w_1 and w_2, the neuron will act as NAND?

(A) $w_1 = -1$ and $w_2 = -1$

(B) $w_1 = 2$ and $w_2 = 2$

(C) $w_1 = 1$ and $w_2 = 1$

(D) $w_1 = -2$ and $w_2 = -2$

(E) None of the Above

E

Lecture 20

Class review 16 Question 1

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve. Start $\mathbf{w}^0 = [-1,-1,4]^T$. Find $\mathbf{w}^1$?

(A) Algorithm has converged. $\mathbf{w}^2$ will be the same as $\mathbf{w}^1$

(B) $\mathbf{w}^1$ is independent of η

(C) $\mathbf{w}^1$ is parallel to $\mathbf{w}^0$, but different.

(D) $\mathbf{w}^1$ will be the same as $\mathbf{w}^0$

(E) None of the above

Using [x,y,1] A

Using [1,x,y] E

Lecture 20

Class review 16 Question 2

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.

Start $\mathbf{w}^0 = [1, -1, 0]^T$. Find $\mathbf{w}^1$?

(A) $\mathbf{w}^1$ is independent of η

(B) $\mathbf{w}^1$ is parallel to $\mathbf{w}^0$, but different.

(C) $\mathbf{w}^2$ will be the same as $\mathbf{w}^1$

(D) $\mathbf{w}^1$ will be the same as $\mathbf{w}^0$

(E) None of the above

Using [x,y,1] E

Using [1,x,y] A,C,D

Lecture 20

Class review 16 Question 3

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.

Start $\mathbf{w}^0 = [-1, -1, 2]^T$. Find $\mathbf{w}^1$?

(A) $\mathbf{w}^2$ will be the same as $\mathbf{w}^1$

(B) $\mathbf{w}^1$ is parallel to $\mathbf{w}^0$, but different.

(C) $\mathbf{w}^1$ is independent of η

(D) $\mathbf{w}^1$ will be the same as $\mathbf{w}^0$

(E) None of the above

Using [x,y,1] A,C,D

Using [1,x,y] E

Lecture 20

Class review 16 Question 4

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.

Start $\mathbf{w}^0 = [-1, -1, 1.9]^T$. Find $\mathbf{w}^1$?

(A) $\mathbf{w}^1$ is independent of η

(B) $\mathbf{w}^2$ will be the same as $\mathbf{w}^1$

(C) $\mathbf{w}^1$ will be the same as $\mathbf{w}^0$

(D) $\mathbf{w}^1$ is parallel to $\mathbf{w}^0$, but different.

(E) None of the above

Using [x,y,1] B

Using [1,x,y] E

Lecture 20

Class review 16 Question 5

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.

Start $\mathbf{w}^0 = [1, -1, 0]^T$. $\mathbf{w}^2$ is:

(A) $[1.2, -0.8, 0.1]^T$

(B) $[0.8, -1.2, -0.1]^T$

(C) $[1, -1, 0]^T$

(D) $[1.4, -0.6, 0.2]^T$

(E) None of the above

Using [x,y,1] E

Using [1,x,y] C

Lecture 21

Class review 17 Question 1

Consider a multi-class classification problem with 8 classes. Let us compare the following three:

A DDAG with pairwise

B Fully balanced Binary Hierarchical Classifier (BHC)

C Majority voting on pairwise classification.

(A) A, B, and C will require exactly the same number of classifiers.

(B) B is faster than A and C (B requires less compute than A and C) for evaluating/testing a sample.

(C) C is better suited for parallel evaluation than A and B

(D) A is faster than C (A requires less compute than C) for evaluating/testing a sample.

(E) All the above statements are true.

B,C,D

Lecture 21

Class review 17 Question 2

Consider an 8-class classification with Binary Hierarchical classification (BHC).

(A) If BHC is not balanced, number of classifiers will increase (compared to the balanced one)

(B) If BHC is not balanced, average time for classification (amount of compute) will increase (compared to the balanced one).

(C) An unbalanced BHC can be converted to a BHC with no loss in accuracy with some rotate operators (just like a rotate operations in AVT Tree in a typical data structure course)

(D) We prefer Balanced BHC, since balanced BHCs will have the highest accuracy.

(E) If BHC is not balanced, we will have multiple leaves with the same label.

(F) All the above.

c

Lecture 21

Class review 17 Question 3

Consider a multi-class classification problem with K classes.

We have now K one vs rest linear classifiers are designed as $\mathbf{w_1} \ldots, \mathbf{w_K}$

(A) Finding $\mathbf{w_1} \ldots, \mathbf{w_K}$ can be formulated and solved as K independent training problem.

(B) Finding $\mathbf{w_1} \ldots, \mathbf{w_K}$ has to be formulated and solved as a single training/optimization problem.

(C) We prefer "Classify as k if k is $\arg\max_k \mathbf{w}_k^T\mathbf{x}$".

(D) We prefer "Classify as k if $\mathbf{w_k^T x} \geq 0$". This will have unambiguous and correct classification.

(E) We used "Classify as k if k is $\arg\max_k \mathbf{w}_k^T\mathbf{x}$" and this resulted in all samples correctly classifying with no ambiguity. If this is the case, all the $\mathbf{w}_i$ (say in a 2D plane) geometrically define lines that intesect at a common point.

B,C (B,C,E)

Lecture 21

Class review 17 Question 4

Consider a K class multi-class classifier implemented with pair-wise classifier and majority voting.

Accuracy of samples in class ω_i is η_i.

(A) Overall accuracy is the average of accuracies of all the K classes. i.e., $\frac{1}{K}\sum_{i=1}^{K}\eta_i$

(B) Overall accuracy is the sum of accuracies of all the K classes. i.e., $\sum_{i=1}^{K}\eta_i$

(C) Overall accuracy is the weighted average of accuracies of all the K classes, where weights are the inverse of the prior probabilities of each of the classes i.e., $\frac{1}{K}\sum_{i=1}^{K}\frac{1}{P(\omega_i)}\eta_i$

(D) Final decision is the class that gets majority votes.

(E) Overall accuracy is the weighted average of accuracies of all the K classes, where weights are the prior probabilities of each of the classes i.e., $\frac{1}{K}\sum_{i=1}^{K}P(\omega_i)\eta_i$

D,E

Lecture 21

Class review 17 Question 5

Consider a multi-class classification problem with 6 classes.

(A) Binary Hierarchical Classification (BHC) is not applicable for this problem, since 6 is not a power of 2.

(B) DDAG requires 15 pairwise classifiers.

(C) DDAG can not be designed for 6 classes since 6 is not a power of 2.

(D) DDAG requires $6C_2$ pairwise classifiers.

(E) None of the above.

B,D

Lecture 22

Class review 18 Question 1

You know LDA/Fisher. Consider a two class problem with two samples $((2,3),+)$ and $((3,4),-)$. S_w is regularized as $S_w + \sigma I$. The LDA solution $\mathbf{w}^*$ will be:

(A) along the direction of $(2,3)$ and $(3,4)$

(B) orthogonal to the direction of $(2,3)$ and $(3,4)$

(C) neither along nor orthogonal to

(D) If we have not regularized, S_w would have become a NULL matrix.

(E) None of the above.

A,D

Lecture 22

Class review 18 Question 2

You know LDA/Fisher for two class classification problem. We know the problem as solving for:

$$S_b u = \lambda S_w u$$

and the solution as:

$$u^* = \alpha S_w^{-1}[\mu_1 - \mu_2]$$

(A) Both S_B and S_w are of $d \times d$.

(B) There is one and only one u that satisfy the problem. (or solution to the problem is unique).

(C) Given the problem statement, we can write $S_b = \lambda S_w$. Or one matrix is the scaled version of the other.

(D) u^* is obtained by solving our problem with an additional constraint of $||u||_2^2 = 1$

(E) None of the above.

A

Lecture 22

Class review 18 Question 3

You know LDA/Fisher. The goal is to:

(A) minimize inter class scatter

(B) maximize inter class scatter

(C) minimize within class scatter

(D) maximize within class scatter

(E) Any two of the above.

B,C

Lecture 22

Class review 18 Question 4

You know LDA/Fisher. There are two classes ω_1 and ω_2. $d = 4$. Both are multivariate Gaussians with $\Sigma = I$. There are 50 and 100 samples from these two classes respectively. i.e., $N = 150$. The rank of S_B is

(A) 4 (d)

(B) 1

(C) 2

(D) 150(N)

(E) None of the above

B

Lecture 22

Class review 18 Question 5

You know LDA/Fisher. There are two classes ω_1 and ω_2. $d = 4$. Botl multivariate Gaussians with $\Sigma = I$. There are 50 and 100 samples fro classes respectively. i.e., $N = 150$.

What is the rank of S_w is:

(A) 4

(B) 150

(C) 100

(D) 1

(E) 2

A

Quiz 1 Questions

Quiz 1, Question 1

In a MCQ a student randomly guesses from the options if she does not know. Given that there were 3 choices in a question and that 0.8409 is the chance she knows the answer, what is the probability that she knew the answer if she answered correctly?

(A) $\frac{2.5227}{3.5227}$

(B) $\frac{1}{2.6818}$

(C) $\frac{1}{3.5227}$

(D) $\frac{2.5227}{2.6818}$

D

Quiz 1, Question 2

Bag I contain 7 white and 2 black balls. Bag II contains 6 white and 10 black balls.

A ball is drawn at random from one of the bags, and it is found to be white. What is the probability that it was drawn from Bag I.

(A) $\frac{7}{9}$

(B) $\frac{63}{159}$

(C) $\frac{84}{156}$

(D) $\frac{112}{166}$

D

Quiz 1, Question 3

A man is known to speak truth 5 out of 10 times. He throws a die and reports that number obtained is a four. Find the probability that the number obtained is actually a four.

(A) $\frac{1}{6}$

(B) $\frac{5}{35}$

(C) $\frac{5}{25}$

(D) $\frac{5}{30}$

D,A

Quiz 1, Question 4

Given the following confusion matrix what is the precision?

	Predicted +ve	Predicted -ve
Actual +ve	1	10
Actual -ve	5	6

(A) $\frac{1}{11}$

(B) $\frac{1}{6}$

(C) $\frac{7}{22}$

(D) $\frac{6}{11}$

B

Quiz 1, Question 5

Consider that numbers from 1 to 25 are arranged in a 5 by 5 dimensional square matrix M in a way such that first 5 numbers are in row 1, next 5 numbers in row 2 and so on. The rank of M is

(A) 1

(B) 2

(C) 5

(D) None of these

B

Quiz 1, Question 6

Consider the covariance matrix Σ

(A) Σ can not be Diagonal if the distribution is Normal.

(B) Σ is symmetric

(C) Σ is Diagonal if the distribution is Normal.

(D) Σ is PSD

(E) None of the above

B,D

Quiz 1, Question 7

We are working with N samples each of d dimension. Consider $N \leq d$

(A) While computing Eigen values, we will see at least $d - N$ zero eigen values.

(B) PCA can not be computed

(C) While computing Eigen values, we will see d zero eigen values.

(D) While computing Eigen values, we will see at max $d - N$ zero eigen values.

(E) We can not use eigen value/vector computation. We need to use SVD.

(F) None of the above.

A

Quiz 1, Question 8

We are working with N samples each of d dimension. Consider $N < d$

(A) Solution to the problem of linear regression as a closed form can not be computed because the matrices are no longer compatble for multiplication.

(B) Solution to the problem of linear regression as a closed form can not be computed because the matrix can not be inverted.

(C) Solution to the problem of ridge regularized linear regression as a closed form can not be computed because the matrix can not be inverted.

(D) None of the above

B

Quiz 1, Question 9

We know the weighted Euclidean distance

$$\tau = [\mathbf{x} - \mathbf{y}]^T [\mathbf{A}][\mathbf{x} - \mathbf{y}]$$

Where $\mathbf{x}$ is a vector in d dimension $\mathbf{A}$ is a square matrix

(A) If $k < d$ and $A = \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^T$, with $\mathbf{u}_i$s as orthonormal. Then A is rank deficient and Euclidean τ can not be computed.

(B) When $\mathbf{A}$ is non-diagonal matrix, τ can not be a metric.

(C) If $\mathbf{A}$ is a non-identity diagonal matrix. Then "dist" is a scaled version of Euclidean distance or "$\tau = \alpha$ Euclidean distance"

(D) If $k < d$ and $A = \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^T$, with $\mathbf{u}_i$s as orthonormal. Then τ is equivalent to dimensioality reduction $\mathbf{x}' = \mathbf{W}\mathbf{x}$ with $\mathbf{W}$ as $k \times d$ matrix with $\mathbf{u}_i^T$ as the i th row.

(E) None of the above.

C,D

Quiz 1, Question 10

We know that solution to the problem of Maximize $\mathbf{w}^T\mathbf{A}\mathbf{w}$ subject to $||\mathbf{w}|| = 1$ is the eigen vector corresponding to the largest eigen value.

Note that we assume that a typical eigen value computation assumes to be returning (i) eigen values arranged in non-increasing order (ii) eigen vectors have unit L2 norm.

What is the solution to the problem of Maximize $\mathbf{w}^T\mathbf{A}\mathbf{w}$ subject to $||\mathbf{w}||_2^2 = 2$

(A) Eigen vector correspond to the second eigen value.

(B) $\sqrt{2} * \mathbf{u}$ where $\mathbf{u}$ is the eigen vector correspond to the first eigen value.

(C) Eigen vector correspond to the first eigen value.

(D) 2*$\mathbf{u}$ where $\mathbf{u}$ is the eigen vector correspond to the first eigen value.

(E) None of the above.

B

Quiz 1, Question 11

Consider the following maximum likelihood estimation(MLE) objective for linear regression:

$$\max_{\theta} \prod_{i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)$$

which leads to the following objective (by taking -ve log):

$$\min_{\theta} \sum_{i} (y_i - \theta^T x_i)^2$$

In the MLE objective,

(A) It is assumed that all the residuals $(y_i - \theta^T x_i)$ have the same standard deviation

(B) It is assumed that the residuals $(y_i - \theta^T x_i)$ have a mean of 0

(C) It is assumed that the residuals $(y_i - \theta^T x_i)$ have a standard deviation of 1

(D) None of these

A,B

Quiz 1, Question 12

Consider the function

$$f(w) = w^2 + w + 1$$

We want to find the minima of the function using gradient descent. We start at $w^0 = 5$. What should be the learning rate η so that we reach the minima in a single step?

Hint: There may be many ways to solve this. One of the easiest is to see that the derivative of the point after 1st update is 0:

(A) 0.5

(B) 0.1

(C) 1

(D) 0.05

(E) None of these

A

Quiz 1, Question 13

Let us say that we have computed the gradient of our cost function and stored it in a vector g. What is the cost of one gradient descent update given the gradient?

D is number of dimensions, N is the number of samples

(A) $O(N)$

(B) $O(ND)$

(C) $O(ND^2)$

(D) $O(D)$

D

Quiz 1, Question 14

We saw the loss function for linear regression as

$$J(\theta) = (Y - X\theta)^T (Y - X\theta)$$

We saw that we get a closed form solution for θ by solving $\frac{\partial J(\theta)}{\partial \theta} = 0$:

$$\frac{\partial}{\partial \theta} \left(Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta\right) = 0$$

$$\implies -2X^T Y + 2X^T X\theta = 0 \implies \theta = (X^T X)^{-1} X^T Y$$

Now find the closed form solution that minimizes this loss function (assume W is symmetric):

$$J(\theta) = (Y - X\theta)^T W (Y - X\theta)$$

(A) $\theta = (X^T W X)^{-1} X^T Y$

(B) $\theta = (X^T X)^{-1} X^T W Y$

(C) $\theta = (X^T W X)^{-1} X^T W Y$

(D) $\theta = (X^T W X)^{-1} X^T W^{-1} Y$

(E) None of these

C

Quiz 1, Question 15

Consider the dataset of 4 points in R^2

$$X = \begin{bmatrix} 7 & -3 \\ 6 & -4 \\ -2 & 6 \\ -3 & 5 \end{bmatrix}$$

Run PCA for this data (in your notebook) to go from R^2 to R^1. Then

(A) The projection of 1st and 2nd points in the new subspace is at same point

(B) 1st Principal component is $[1, 1]^T$

(C) The projection of 3rd and 4th points in the new subspace is at ame point

(D) The projection of 1st and 3rd points in the new subspace is at same point

(E) 1st Principal component is $[1, -1]^T$

(F) None of these

C,D,E

Quiz 1, Question 16

Consider a typical two class classification problem. We have a labelled set of 1000 samples. (say $N = 1000, d = 2$).

We train a classifier iterative (say using a GD) and minimize a loss function (eg. a mean square error loss).

We use 80% data for training (i.e., 800) and rest 20% (i.e., 200) for testing.

During the iteration k, loss on the training data and Test data are L^k_{Tr} and L^k_{Te}. Let the accuracy of the classifier at iteration i be η^i_{Tr} and η^i_{Te}.

(A) If $L^k_{Tr} \gg L^l_{Tr}$, then $l > k$.

(B) If $L^k_{Tr} > L^l_{Tr}$, then If $L^k_{Te} > L^l_{Te}$.

(C) If $L^k_{Tr} \gg L^l_{Tr}$, then $\eta^k_{Tr} > \eta^l_{Tr}$ (strictly greater).

(D) If $L^k_{Tr} > L^l_{Tr}$, then If $L^k_{Te} < L^l_{Te}$.

(E) None of the above.

A

Quiz 1, Question 17

Consider a two class classification problem. We have a labelled set of 1000 samples. (say $N = 1000, d = 2$).

We train a classifier iterative (say GD) and minimize a loss function (eg. a mean square error loss) by using all the samples.

In each iteration, we use a random 80% of the total data as training (i.e., 800) and rest 20% (i.e., 200) for testing.

(A) Since the training data is changing in every iteration (or regular basis), the loss will not come down.

(B) This iterative algorithm will not converge.

(C) Since the test data is changed on a regular basis, the solution will generalize well.

(D) This is perfectly fine way of training the ML solution.

(E) None of the above

E

Quiz 1, Question 18

In the context of supervised machine learning,

(A) Supervised learning is all about overfitting to the given data.

(B) Occam's Razor says: Suppose there exist two explanations for an occurrence. In this case the one that requires the smallest number of assumptions is usually correct.

(C) Regularization decrease the chance of overfitting.

(D) Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points.

(E) None of the above.

B,C,D

Quiz 1, Question 19

Consider a typical two class classification problem. We have a labelled set of 1000 samples. (say $N = 1000, d = 2$).

We train a classifier iterative (say using GD) and minimize a loss function (eg. a mean square error loss).

We use 80% data for training (i.e., 800) and rest 20% (i.e., 200) for testing.

(A) If both training and test loss are zero, then we are sure that the algorithm has overfit.

(B) If we allow to continue for enough time the the loss on training data will eventually become zero.

(C) To train the model, We iterate until the loss on the training data becomes zero.

(D) If at any point of time, the loss on training data is zero, then the loss on test data will also be zero.

(E) None of the above.

E

Quiz 1, Question 20

In the context of regularization:

(A) Any regularization will lead to sparse solution.

(B) L_0 regularization leads to sparse solution.

(C) L_2 regularization leads to sparse solution.

(D) L_∞ regularization leads to sparse solution.

(E) L_1 regularization leads to sparse solution.

B,E

Review Questions contd...

Lecture 23

Class review 19 Question 1

Let $\mathbf{p} = [p_1, p_2]^T$ and $\mathbf{q} = [q_1, q_2]^T$ be two vectors in 2D.

$\kappa(\cdot, \cdot)$ is a kernel and $\phi()$ is the corresponding feature map.

If $\phi(\mathbf{z}) = [z_1^2, z_2^2, \sqrt{2}z_1z_2, \sqrt{2}z_1, \sqrt{2}z_2, 1]^T$ then $\kappa(\mathbf{p}, \mathbf{q})$ is:

(A) $(1 + \mathbf{p}^T\mathbf{q})^2$

(B) $\mathbf{p}^T\mathbf{q}$

(C) $(\mathbf{p}^T\mathbf{q})^2$

(D) $(\mathbf{p}^T\mathbf{q})^{\sqrt{2}}$

(E) None of the above.

A

Lecture 23

Class review 19 Question 2

Let $\mathbf{p} = [p_1, p_2]^T$ and $\mathbf{q} = [q_1, q_2]^T$ be two vectors in 2D.

$\kappa(\cdot, \cdot)$ is a kernel and $\phi()$ is the corresponding feature map.

If $\phi(\mathbf{z}) = [z_1^2, z_2^2, \sqrt{2}z_1 z_2]^T$ then $\kappa(\mathbf{p}, \mathbf{q})$ is:

(A) $\mathbf{p}^T\mathbf{q}$

(B) $(\mathbf{p}^T\mathbf{q})^{\sqrt{2}}$

(C) $(1 + \mathbf{p}^T\mathbf{q})^2$

(D) $(\mathbf{p}^T\mathbf{q})^2$

(E) None of the above.

D

Lecture 23

Class review 19 Question 3

Let $\mathbf{p} = [p_1, p_2]^T$ and $\mathbf{q} = [q_1, q_2]^T$ be two vectors in 2D.

$\kappa(\cdot, \cdot)$ is a kernel and $\phi()$ is the corresponding feature map.

Let $\kappa() = \sum_{i=1}^{P} \kappa_i()$. Then the $\phi()$ is

(A) $\phi()$ is obtained by concatenating $\phi_i()$s.

(B) $\Pi_{i=1}^{P} \phi_i()$

(C) There is no analytical relationship between $\phi()$ and $\phi_i()$s.

(D) $\sum_{i=1}^{P} \phi_i()$

(E) None of the above.

A

Lecture 23

Class review 19 Question 4

Let $\mathbf{p} = [p_1, p_2]^T$ and $\mathbf{q} = [q_1, q_2]^T$ be two vectors in 2D.

$\kappa(\cdot, \cdot)$ is a kernel and $\phi()$ is the corresponding feature map.

If $\phi(\mathbf{z}) = [z_1^2, z_2^2, z_1 z_2, z_2 z_1]^T$ then $\kappa(\mathbf{p}, \mathbf{q})$ is:

(A) $(\mathbf{p}^T\mathbf{q})^2$

(B) $\mathbf{p}^T\mathbf{q}$

(C) $(1 + \mathbf{p}^T\mathbf{q})^2$

(D) $(\mathbf{p}^T\mathbf{q})^{\sqrt{2}}$

(E) None of the above.

A

Lecture 23

Class review 19 Question 5

Let $\mathbf{p} = [p_1, p_2]^T$ and $\mathbf{q} = [q_1, q_2]^T$ be two vectors in 2D.

$\kappa(\cdot, \cdot)$ is a kernel and $\phi()$ is the corresponding feature map.

Let $z = \kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T\mathbf{q})^2$.

(A) z is unique for the given $\kappa()$.

(B) $\phi()$ is unique given the $\kappa()$.

(C) $\phi() \in R^2$.

(D) z is scalar.

(E) All the above.

A,D

Class review 20 Question 1

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem. The decision boundary is:

(A) $-2x_1 - 2x_2 = 3$

(B) $2x_1 + 2x_2 = -3$

(C) $-2x_1 - 2x_2 = -3$

(D) $2x_1 + 2x_2 = 3$

(E) None of the above.

C,D

Class review 20 Question 2

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem.

(A) Addition of $([0,2]^T, +)$ will change the support vector set, but not the margin.

(B) Addition of $([1,2]^T, +)$ does not change the support vector set and the margin.

(C) Addition of $([0, \frac{3}{2}]^T, +)$ will change the support vector set, and the margin.

(D) Addition of $([0, \frac{3}{2}]^T, +)$ will change the support vector set, but the number of support vectors will not change.

(E) Addition of no sample can increase the margin.

A,B,C,E

Class review 20 Question 3

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem.

(A) Given that the problem is in 2D, and binary classification, addition of a new support vector sample will make one of the existing support vectors as non-support vector.

(B) If we remove any one of the support vectors from the training data and retrain the SVM, we will get a different solution.

(C) For this problem, there exists at least one sample, removal of it will lead to a different solution for the SVM.

(D) There exists at least one non-support vector in $\mathcal{D}$, such that removal of it from the training data lead to a different solution.

(E) None of the above.

C

Class review 20 Question 4

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem. The following is a support vector:

(A) $[0,2]^T$

(B) $[2,2]^T$

(C) $[0,0]^T$

(D) $[\frac{3}{2},0]^T$

(E) $[1,1]^T$

E

Class review 20 Question 5

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem.

(A) If we remove $[0,0]^T$ from $\mathcal{D}$, the margin increase.

(B) If we remove $[0,1]^T$ from $\mathcal{D}$, the margin increases.

(C) If we remove $[2,2]^T$ from $\mathcal{D}$, the margin increases.

(D) If we remove $[1,0]^T$ from $\mathcal{D}$, the margin increases.

(E) If we remove $[1,1]^T$ from $\mathcal{D}$, the margin increases.

D,E

Lecture 24

Class review 21 Question 1

Consider a training dataset with m samples in R^d space. Consider a kernel which projects any sample in R^k space.

We would like to find out the Kernel SVM prediction on a test sample. What is the time complexity of this prediction

(A) $O(k)$

(B) $O(mdk)$

(C) $O(m)$

(D) $O(mk)$

(E) $O(md)$

(F) $O(d)$

(G) None of these

E

Lecture 24

Class review 21 Question 2

Consider the dataset $\mathcal{D}$:

$\mathcal{X} = \{[1,0]^T, [0,1]^T, [.5,.5]^T, [1,1]^T\}$, $\mathcal{Y} = \{+1, +1, -1, -1\}$

We use the kernel $K(x, x') = \frac{x^T x'}{\|x\|\|x'\|}$ to peform kernel SVM (hard-margin). Then the decision boundary in the new feature space can be written as $w_1 x_1 + w_2 x_2 + w_3 = 0$

(A) $w_1 = 0, w_2 = 1$

(B) $w_1 = 1, w_2 = 0$

(C) $w_1 = 1, w_2 = -1$

(D) $w_1 = 1, w_2 = 1$

(E) $w_1 = 0, w_2 = 0$

(F) None of these

D

Lecture 24

Class review 21 Question 3

What does the kernel $K(x, x') = \frac{x^T x'}{\|x\|\|x'\|}$ do?

(A) Project all points on a unit circle

(B) Project all points on a line

(C) Project all points on the line $Y = X$

(D) Project all points on a non-unit circle

(E) None of these

A

Lecture 24

Consider the primal form of soft margin SVM:

$$min_w \frac{1}{2} w^T w + C \sum_i \xi_i$$

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \forall i$.

Now find its dual form as a function of α, x, y

(A) Dual form objective remains same as hard margin SVM

(B) $w = \sum_i \alpha_i x_i y_i$ similar to hard margin SVM

(C) $\sum_i x_i y_i = 0$ similar to hard margin SVM

(D) Dual form constraints remain same as hard margin SVM

(E) None of these

A,B,C

Lecture 24

Class review 21 Question 5

Consider the dataset $\mathcal{D}(\mathcal{X}, \mathcal{Y})$:

$\mathcal{X} = \{[1,0]^T, [0,1]^T, [.5,.5]^T, [1,1]^T\}$, $\mathcal{Y} = \{+1, +1, -1, -1\}$

We use the kernel $K(x, x') = \frac{x^T x'}{\|x\|\|x'\|}$ to peform kernel SVM (hard-margin)

(A) We see 4 unique points in the new feature space

(B) $\mathcal{D}$ is linearly separable in the new feature space

(C) $\mathcal{D}$ is linearly separable in the original feature space

(D) If wee add the point $\{[2,2]^T, -1\}$ to $\mathcal{D}$, then $\mathcal{D}$ is linearly separable in the new feature space

(E) If wee add the point $\{[2,2]^T, 1\}$ to $\mathcal{D}$, then $\mathcal{D}$ is linearly separable in the new feature space

(F) None of these

B,D

Class review 22 Question 1

This question is based on the brief review of Kernels you had seen at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0 (shared in the class last week).

We know $\kappa(\mathbf{p}, \mathbf{q})$ as $\phi(\mathbf{p})^T \phi(\mathbf{q})$.

Consider two vectors $\phi(\mathbf{p})$ and $\phi(\mathbf{q})$ and their L2 normalized version as $\phi(\mathbf{p})'$ and $\phi(\mathbf{q})'$. i.e.,

$$\phi(\mathbf{p})' = \frac{\phi(\mathbf{p})}{\|\phi(\mathbf{p})\|}$$

How do we compute $\kappa'(\mathbf{p}, \mathbf{q}) = (\phi(\mathbf{p})')^T (\phi(\mathbf{q})')$ in terms of $\kappa(\mathbf{p}, \mathbf{q}) = (\phi(\mathbf{p}))^T (\phi(\mathbf{q}))$.

i.e., $\kappa'(\mathbf{p}, \mathbf{q}) =$

(A) $\frac{\kappa(\mathbf{p},\mathbf{q})}{\sqrt{\kappa(\mathbf{p},\mathbf{p})\kappa(\mathbf{q},\mathbf{q})}}$

(B) $\frac{\kappa(\mathbf{p},\mathbf{q})}{\sqrt{\kappa(\mathbf{p},\mathbf{q})\kappa(\mathbf{p},\mathbf{q})}}$

(C) $\frac{\kappa(\mathbf{p},\mathbf{q})}{\kappa(\mathbf{p},\mathbf{q})\kappa(\mathbf{p},\mathbf{q})}$

(D) $\frac{\kappa(\mathbf{p},\mathbf{q})}{\kappa(\mathbf{p},\mathbf{p})\kappa(\mathbf{q},\mathbf{q})}$

(E) None of the above

A

Class review 22 Question 2

This question is based on the brief review of Kernels you had seen at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0 (shared in the class last week).

We know $\kappa(\mathbf{p}, \mathbf{q})$ as $\phi(\mathbf{p})^T \phi(\mathbf{q})$.

How we express the squared euclidean distance

$$d(\phi(\mathbf{p}), \phi(\mathbf{q})) = [\phi(\mathbf{p}) - \phi(\mathbf{q})]^T [\phi(\mathbf{p}) - \phi(\mathbf{q})]$$

in terms of the kernels

(A) $\kappa(\mathbf{p}, \mathbf{p}) + \kappa(\mathbf{q}, \mathbf{q}) - 2\kappa(\mathbf{p}, \mathbf{q})$

(B) $\kappa(\mathbf{p}, \mathbf{p}) + \kappa(\mathbf{q}, \mathbf{q}) + 2\kappa(\mathbf{p}, \mathbf{q})$

(C) $(\kappa(\mathbf{p}, \mathbf{p}) - \kappa(\mathbf{q}, \mathbf{q}))^2$

(D) $\kappa(\mathbf{p}, \mathbf{q})$

(E) None of the above.

A

Class review 22 Question 3

This question is based on the brief review of Kernels you had seen at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0 (shared in the class last week).

Let μ be the mean of samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

Also τ be the mean of samples $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)$.

Let $\phi()$ be the feature map corresponding to RBF Kernel (or your more familiar quadratic kernel).

(A) μ and τ are identical.

(B) $\tau \neq \phi(\mu)$

(C) μ and τ are different; but of same dimension.

(D) $\tau = \phi(\mu)$

(E) Two of the above are true.

B

Class review 22 Question 4

This question is based on the brief review of Kernels you had seen at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0 (shared in the class last week).

We know $\kappa(\mathbf{p}, \mathbf{q})$ as $\phi(\mathbf{p})^T\phi(\mathbf{q})$.

Consider a data matrix with a feature map i.e.,

$$\mathbf{X} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots \phi(\mathbf{x}_N)]$$

Then $\mathbf{X}^T\mathbf{X}$ is

(A) $N \times N$

(B) The kernel matrix $\mathbf{K}$ with $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$

(C) Symmetric

(D) Can not be computed since $\phi()$ can map to infinite dimension

(E) None of the above

A,B,C

Class review 22 Question 5

This question is based on the brief review of Kernels you had seen at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0 (shared in the class last week).

We know $\kappa(\mathbf{p}, \mathbf{q})$ as $\phi(\mathbf{p})^T \phi(\mathbf{q})$.

Kernel matrix $\mathbf{K}$ is

(A) Symmetric

(B) $d \times d$

(C) $N \times N$

(D) Depends on $\phi()$

(E) None of the above

A,C,D

Class review 23 Question 1

Kernel SVMs are:

(A) Always hard margin

(B) Always soft margin

(C) If there is a hard margin feasible, soft-margin K-SVMs eventually finds this.

(D) Can be either hard margin or soft margin

(E) None of the above

D

Class review 23 Question 2

In the context of K-SVM:

(A) If the kernel $\kappa(x, y) = x^T y$, then K-SVM is equivalent to Linear SVM

(B) Performance of the solution change with kernel.

(C) Formulation change with kernel.

(D) We can use different Kernels during training and testing.

(E) All the above four are true.

A,B

Class review 23 Question 3

In the context of Softmargin Linear SVMs:

(A) If there is a hard margin feasible, soft-margin linear SVMs can finds this with certain 'C'.

(B) The larger the C, the formulation become closer and closer to hard margin SVM.

(C) If the data is linearly separable, we should not use soft margin SVM.

(D) If there is a hard margin feasible, soft-margin linear SVMs always finds this.

(E) None of the above.

A,B

Class review 23 Question 4

Number of Support Vectors:

(A) can be as small as 1

(B) can be as large as N

(C) depends on the kernel we use.

(D) is $2 \times d$

(E) None of the above.

B,C

Class review 23 Question 5

For Support Vector Machines:

(A) minimization of $\mathbf{w}^T\mathbf{w}$ with the associated constraints is leading to the maximization of the margin.

(B) If the data is linearly separable, linear SVM and linear perceptron will give the same solution.

(C) A hard margin linear SVM yields a valid (constraint satisfying) solution for any data.

(D) maximization of $\mathbf{w}^T\mathbf{w}$ with the associated constraints is leading to the maximization of the margin.

(E) None of the above.

A

Class review 24 Question 1

You might have read the notes on Kernels and SVMs at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0

Look at the equation (94) related to the objective function:

(A) There is a typo. LHS will have to be

(B) There is a typo. ξ_i should be replaced as ξ_i^2

(C) This is an L1 softmargin SVM

(D) This is an L2 softmargin SVM

(E) None of the above.

C

Class review 24 Question 2

You might have read the notes on Kernels and SVMs at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0

See the pseudo-code for Kernel Perceptron (Algorithm 3). Assime the kernel to be $(\mathbf{x}^T\mathbf{y})^2$

(A) For data that is linearly separable, this algorithm will give you a linear decision boundary.

(B) The initialization $\alpha_i = 0$ is a must. With no other initialization, this algorithm will not work (say will not converge)

(C) Step of computing Kernel Matrix (step 2) should have been in side the loop (repeat structure).

(D) Since this is now Kernelized, with any data (irrespective of whether the data is linearly separable or not), this algorithm will converge.

(E) None of the above.

E

Class review 24 Question 3

You might have read the notes on Kernels and SVMs at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0

See the Sec 3.3 (equations)

Why is $a \in R^+$ written there?

(A) a negative does not lead to K being PSD

(B) It could have been $a \in R^-$

(C) It is a typo.

A

Class review 24 Question 4

You might have read the notes on Kernels and SVMs at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0

See the Sec 3.3. Assume $\alpha_i \in R^+$; $\beta_i \in R^+$; $\kappa_i(\mathbf{p}, \mathbf{q})$ being a valid kernel. Also K and L are some positive integers.

Then a new kernel $\kappa(\cdot, \cdot) =$

(A) $\sum_{i=1}^{K} \alpha_i \kappa_i(\mathbf{p}, \mathbf{q}) + \prod_{i=1}^{L} \beta_i \kappa_i(\mathbf{p}, \mathbf{q})$ is a valid kernel

(B) $\sum_{i=1}^{K} \kappa_i(\mathbf{p}, \mathbf{q})$ is a valid kernel.

(C) $\prod_{i=1}^{L} \kappa_i(\mathbf{p}, \mathbf{q})$ is a valid kernel.

(D) $\sum_{i=1}^{K} \alpha_i \kappa_i(\mathbf{p}, \mathbf{q})$ is a valid kernel.

(E) $\prod_{i=1}^{L} \beta_i \kappa_i(\mathbf{p}, \mathbf{q})$ is a valid kernel.

A,B,C,D,E

Class review 24 Question 5

You might have read the notes on Kernels and SVMs at: https://www.dropbox.com/s/qryziuo3u143q5e/KERNEL-REVIEW.pdf?dl=0

Consider the decision making rule. "one side of a line (in 2D) is +ve class and other side of a line is -ve class" Figure 7 shows that VC dimension of a class of functions (lines) in 2D is 3.

What is the VC dimension in 1D for a function class. If $x > \theta$, positive, else negative.

Write your answer in the space provided.

(Sample answer (possibly incorrect): 1)

VC dimension is 2

Class review 25 Question 1

An MLP has two inputs, two hidden layers of 3 neurons each and an output of two neurons. All the neurons have biases. The number of weights (or learnable parameters) is:

(A) 37

(B) 29

(C) 24

(D) 21

(E) None of the above

B

Class review 25 Question 2

We know that $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. What is the derivative of $tanh(x)$

(A) $1 + \tanh^2(x)$

(B) $1 - \tanh^2(x)$

(C) $\tanh(x)(1 - \tanh(x))$

(D) $1 + \tanh(x)$

(E) None of the above

B

Class review 25 Question 3

We know that the VC dimension of a set of lines in 2D is 3. What is the VC dimension of a set of planes in 3D?

(A) Remains the same. i.e., 3

(B) $3+1=4$

(C) $2+2=2$

(D) $2 \times \frac{3}{4}=6$

(E) None of the above

B

Class review 25 Question 4

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Backpropagation algorithm *can guarantee (always find) the optimal solution/weights* **for a Multilayer Perceptron**.

Answer Text	Backpropagation algorithm can not guarantee (doesn't always find) the optimal solution/weights for a Multilayer Perceptron.

Class review 25 Question 5

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

A Single Layer Perceptron **can solve ExOR problem**.

Answer Text A Multi Layer Perceptron can solve ExOR problem

Class review 26 Question 1

About Backpropagation Algorithm:

(A) When the algorithm is allowed to run for many (say ∞) iterations, the network will overfit.

(B) When the algorithm is allowed to run for many (say ∞) iterations, we will reach a local minima.

(C) When the algorithm is allowed to run for many (say ∞) iterations, we will reach the same local minima, irrespective of the initialization

(D) When the algorithm is allowed to run for many (say ∞) iterations, the loss becomes zero.

(E) All the above are true.

B

Class review 26 Question 2

About Backpropagation Algorithm:

Which may be a really **bad** termination crieria

(A) When loss is near zero, end.

(B) When all gradients are near zero, end.

(C) When no major change in loss, end.

(D) When learning rate is near zero, end.

(E) All the above are terrible termination crieria.

D/AD

Class review 26 Question 3

Consider an MLP getting used for a three class classification problem.

Output layer has three neurons and we use a cross entropy loss.

(A) If the loss is zero, implies that the MLP as a classifier has 100% accuracy on the training data.

(B) If the accuracy of the test data is 100%, it implies that the loss computed on the training data might have been zero.

(C) If the loss is zero, implies that the MLP as a classifier has 100% accuracy on the test data.

(D) If the accuracy of the training data is 100%, it implies that the loss might have been zero.

(E) None of the above.

A/ABD

Class review 26 Question 4

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

An MLP has no activation in the output. It has sigmoid activity in all the hidden layers. *It can not be used to output negative values* **because** *sigmoid outputs in [0,1] (no negative)*

An MLP has no activation in the output. It has sigmoid activity in all the hidden layers. *It can not be used to output negative values* **because** *sigmoid outputs in [0,1] (no negative)* can be used output=wx+b can be -ve even if $x \geq 0$ if w or b are -ve for last layer

Class review 26 Question 5

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Backpropagation algorithm *can be understood as an iterative optimization algorithm.*

Backpropagation algorithm *can be understood as an iterative optimization algorithm.*

Class review 27 Question 1

There is a popular problem called "parity". Consider **x** be d-dimensional, $d > 1$ which each $x_i \in \{-1, +1\}$ and y be $+1$ if the number of $+1$ in **x** is odd and else -1.0

(A) When $d = 2$, this problem reduced to *ExoR*.

(B) This problem is linearly separable.

(C) This is an example of a linearly non-separable problem.

A,C

Class review 27 Question 2

Remember the problem set we solved in the last class. (refer the questions and your answers if needed).

We solved the Separable SVM problem in 1D for

$$(-1, +1), (0, -1), (+1, -1)$$

We knew the solution: w as -2 and b as -1

Assume x was k times (say was measured in a different unit; remember the normalization of the data). For example, when $k = 2$, the data will look like:

$$(-2, +1), (0, -1), (+2, -1)$$

(A) w will remain the same, while b will become $\frac{b}{k}$.

(B) No such systematic change is possible for w and b. The problem will have to be solved afreash.

(C) b will remain the same, while w will become $\frac{w}{k}$

(D) w and b will remain the same.

(E) w and b will also become k times.

C

Class review 27 Question 3

Remember the problem set we solved in the last class. (refer the questions and your answers if needed).

We solved the Separable SVM problem in 1D for

$$(-1, +1), (0, -1), (+1, -1)$$

We knew the solution: w as -2 and b as -1

Consider an extension, we add some small zero mean Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each of the samples and create 10 variations each. (total of 30 samples). (Assume $\sigma = 0.5$.)

We solve the SVM problem.

(A) The optimal values of w and b will remain almost the same.

(B) Margin remains the same.

(C) The optimal values of w and b will become 10 times.

(D) We will have only two Support Vectors.

(E) We expect around 20 Support Vectors. (non zero αs)

(F) None of the above.

A

Class review 27 Question 4

Remember the problem set we solved in the last class. (refer the questions and your answers if needed).

Consider the ExOR problem. There are four samples and four αs and four support vectors.

We can generalize this observation as:

"**For any linearly non-separable problem**, *number of support vectors is same as number of samples.*"

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Answer Text

For any linearly non-separable problem, number of support vectors is not the same as the number of samples

Class review 27 Question 5

Remember the problem set we solved in the last class. (refer the questions and your answers if needed).

We solved SVM for a linearly in-separable data:

$$(-1, +1), (0, -1), (+1, +1)$$

and obtained: $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = 2$

"**Assume we had an additional (4th) sample** $(+2, +1)$ **in our data**, *the* α*s for the first three samples, will remain the same.*"

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Answer Text	Assume we had an additional 4th sample (+2,+1) in our data, the alphas for the first three samples will remain the same.

Class review 28 Question 1

Consider the ReLU activation function for a neuron.

(A) Output is always non-negative.

(B) Output is same as input.

(C) Output is always positive

(D) output is either one or zero.

(E) None of the above.

A

Class review 28 Question 2

Consider the ReLU activation function for a neuron. Derivative of the ReLu function:

(A) continuous

(B) differentiable

(C) can take two values.

(D) is Constant throughout

(E) can never be negative

C,E

Class review 28 Question 3

Consider an MLP with 2 inputs, 3 neurons in hidden and one output. Hidden neurons and output neuron uses ReLU Activation.

Let the input be x_1 and x_2 and output be y. We train this with MSE loss.

(A) If x_1, x_2 are negative, and y is positive for all the samples, this network can not be used for effective problem solving.

(B) This network can not be useful if either input or output is negative.

(C) This network can be effectively used irrespective of whether input or output is negative.

(D) If x_1, x_2 are positive, and y is negative for all the samples, this network can not be used for effective problem solving.

(E) None of the above.

D

Class review 28 Question 4

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Consider a deep neural network with ReLU activations. *Since the gradient is same as input (which can be very large quantity), there is a chance of vanishing gradient problem.*

Consider a deep neural network with ReLU activations.
Since the gradient is same as input (which can be very large quantity), there is a chance of vanishing gradient problem.
gradient is 0 or 1; there is no vanishing gradient

Class review 28 Question 5

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

For leaky ReLu, gradients are *either positive or negative.*

For leaky ReLu, gradients are *either positive or negative.*
gradients are always positive

Class review 29 Question 1

A Fully connected layer has 100 neurons at input and 100 neurons at output. The number of parameters to learn is:

(Assume there is no bias.)

(A) 50

(B) 1000

(C) 10000

(D) 200

(E) None of the above

C

Class review 29 Question 2

A convolutional layer has 100 inputs and 5 channels of 100 outputs there with sufficient zero padding. The number of learnable parameters is: Each output channel is computed with 7 learnable weights.

Total number of learnable parameters is:

(A) 5

(B) 12

(C) 35

(D) 7

(E) None of the above

C

Class review 29 Question 3

A convolutional layer has 100 inputs and 100 outputs there is sufficient zero padding. The number of learnable parameters is:

(A) 10

(B) 5

(C) 1

(D) 3

(E) Any of the above

E

Class review 29 Question 4

We know that if there is no zero padding, the convolution output is smaller than the original. Consider an input of size/length 100. Convolution is carried over a window of length 7 with stride of 1. What is the length/size of output?

94

Class review 29 Question 5

A convolution layer has 3 input channels of size 100 each. Output is computed over a 1-D window of length 5. There are 7 output channels. Stride is 2.

How many learnable parameters exist in this layer?

105

Class review 30 Question 1

Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

Network is initialized with all weights as non-zero but a small constant.

(A) With Back propagation, weights won't change.

(B) All derivatives ($\frac{\partial L}{\partial w}$) are zero.

(C) Output is zero.

(D) Loss is zero.

(E) None of the above.

E

Class review 30 Question 2

Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

We initialize the network with good practices available/reported. Starting from the same initialization, we train the network multiple times with BP/GD. Starting from the same initialization, we train the network multiple times with BP/GD with momentum term also.

(A) Starting from the same initialization, the solution at epoch 100 remains same with or without momentum.

(B) Starting from the same initialization, the solution at epoch 100 remains same in all the runs, when the implementation was SGD.

(C) Starting from the same initialization, the solution at epoch 100 remains same in all the runs, when the implementation was batch GD.

(D) With an appropriate but fixed convergence criteria (say early stopping), models trained with and without momentum could be different.

(E) With an appropriate but fixed convergence criteria (say early stopping), models trained with and without momentum will be the same.

C,E

Class review 30 Question 3

Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

What may be a better initialization, among the following?

(A) All the weights random and positive

(B) Some weights random and positive and some weights random and negative.

(C) All weights the same.

(D) All the weights random and negative.

(E) All of the above are equally good or equally bad.

A,B,D

Class review 30 Question 4

Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

Network is initialized with all weights as zero.

(A) With Back propagation, weights won't change.

(B) Output is zero.

(C) All derivatives ($\frac{\partial L}{\partial w}$) are zero.

(D) Loss is zero.

(E) None of the above.

A,B,C

Class review 30 Question 5

Consider an MLP: 2 input, one hidden layer (3 neurons) and one output. All neurons use ReLU activation. No bias. We use MSE loss.

We use this network for regression to predict, say the mean temperature of tomorrow in Hyderabad, which is always positive.

We train the network with sufficient amount of data, and follow good practices of training.

(A) At the end of training Loss will become zero.

(B) At the end of training, all our weights will be positive.

(C) At the end of training, we are at a local minima.

(D) At the end of training, we will be at a global minima.

(E) None of the above.

C

Class review 31 Question 1

Consider a Divisive Clustering Algorithm with two steps:

Create an MST

Successively remove the longest or largest edges.

Assume there are 100 samples, and all edges are of unique length.

(A) Every run of this algorithm can give different solution and therefore, this is sensitive to the ordering/indices of the samples in the set.

(B) This algorithm is yielding a globally optimal solution to a specific objective.

(C) Since there are better/other clustering algorithms, the final solution is only locally optimal.

(D) The objective function that this algorithm minimizes is the following:

$$\sum_{i} \sum_{x_i, x_j \in C_i} d(x_i, x_j)^2$$

(E) Since this is a global optima, there can not exist a better clustering algorithm.

B,D

Class review 31 Question 2

Consider a Divisive Clustering Algorithm with two steps:

Create an MST

Successively remove the longest or largest edges.

Assume there are 100 samples, and all edges are unique length.

What can be a **bad** termination criteria?

(A) Stop when there are no more edges to remove.

(B) Average length of leftout edges is more than average length of removed edges.

(C) Stop when the length of the next largest is less than half of the edge removed in the previous step?

(D) Stop when all the left out edges are less than p

(E) Stop when the length of the next largest is more than half of the edge removed in the previous step?

A,B,E

Class review 31 Question 3

Consider a Divisive Clustering Algorithm with two steps:

Create an MST

Successively remove the longest or largest edges.

Assume there are 100 samples, and all edges are of unique length. If we have removed 5 edges, the number of clusters is:

(A) 2^5

(B) 6

(C) 5

(D) 5^2

(E) None of the above.

B

Class review 31 Question 4

Among the following problems, a clustering algorithm is mostly appropriate for:

(A) Predicting the rainfall in HYD in 2022.

(B) To detect Credit Card transactions that are Frauds

(C) For a robot to decide the direction to travel to Himalaya 105 class room.

(D) To translate a sentence from Hindo to Telugu

(E) All the above

B

Class review 31 Question 5

Assume there are N samples in a data set, the number of distinct ways in which we can cluster this set is:

(A) 2^N

(B) ${}_N C_2$

(C) N

(D) M

(E) None of the above

E

Class review 32 Question 1

Computational complexity/effort of K-means algorithm depends on:

(A) N

(B) K

(C) d

(D) No of iterations to converge

(E) All the above

A,B,C,D,E

Class review 32 Question 2

In K-Means:

(A) K is often odd.

(B) K is often larger than N

(C) K is often much smaller than N.

(D) K is often equal to N.

(E) None of the above

C,A

Class review 32 Question 3

K-Means:

(A) reaches the same final answer irrespective of the initialization.

(B) maximizes the sum of within cluster variances

(C) never converges to the global optima.

(D) cluster assignments are mutually exclusive and collectively exhaustive.

(E) None of the above

D

Class review 32 Question 4

Consider a measure computed from the final answer of K-Means:

$$J_k = \frac{1}{k} \sum_{l} \sum_{x_i \in C_l} \|x_i - \mu_l\|_2^2$$

With increase in k

(A) J_k will monotonically increase

(B) It could increase first and then decrease.

(C) It could decrease and then increase

(D) J_k will monotonically decrease.

(E) None of the above.

D

Class review 32 Question 5

Consider a set of 10 2D points (i.e., $N = 10, d = 2$) $\{\mathbf{x}_i\}$ as

$$\{[-2,-1]^T, [-3,-2]^T, [0,-1]^T, [-1,0]^T, [2,3]^T, [-1,-2]^T, [3,2]^T, [3,3]^T, [1,1]^T, [2,2]^T\}$$

Cluster them into two clusters $K = 2$. Initialize the K Means such that the first five samples are in cluster A and the next 5 are in cluster B.

Write the final means as

(x, y) and (a, b)

[-1.4, -1.2] and [2.2, 2.2]

Class review 33 Question 1

Suppose instead of XOR data, we now want to work on NAND data. Model 1 is a MLP with a hidden layer with 2 neurons as we saw. Model 2 is a SLP.

(A) Model 2 can classify all 4 samples correctly but not model 1

(B) Neither model 1 nor model 2 can classify all 4 samples correctly.

(C) Both model 1 and model 2 can classify all 4 samples correctly.

(D) Model 1 can classify all 4 samples correctly but not model 2

(E) None of these

C

Class review 33 Question 2

We saw this implementation of MLP in PyTorch:

```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        # Initialize all the layers with learnable parameters
        self.fc1 = nn.Linear(2, 2, True)
        self.fc2 = nn.Linear(2, 1, True)

    def forward(self, x):
        # Write the forward pass
        x = self.fc1(x)
        x = torch.sigmoid(x)
        x = self.fc2(x)
        x = torch.sigmoid(x)
        return x
```

What happens if we remove the 2 lines with code x=torch.sigmoid(x) in the forward function

Assume that we are working with the XOR data as given in the notebook shared.

(A) Model remains multi layer perceptron

(B) Syntax error

(C) Model becomes Single layer perceptron

(D) Math error

(E) None of these

C

Class review 33 Question 3

We saw this implementation of MLP model in PyTorch:

```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        # Initialize all the layers with learnable parameters
        self.fc1 = nn.Linear(2, 2, True)
        self.fc2 = nn.Linear(2, 1, True)

    def forward(self, x):
        # Write the forward pass
        x = self.fc1(x)
        x = torch.sigmoid(x)
        x = self.fc2(x)
        x = torch.sigmoid(x)
        return x
```

The output of this model is

(A) Always less than one

(B) Always greater than zero

(C) Always zero

(D) Can be negative as well as positive

(E) None of these

A,B

Class review 33 Question 4

We saw an implementation of MLP in numpy and PyTorch. You may have noticed that the weights are initialized randomly. What happens if we we set all weights and biases to 0

Assume we are using ReLU activation

(A) Weights do not change while training

(B) Underfitting issue

(C) Overfitting issue

(D) No problem: the model will converge nicely

(E) None of these

A

Class review 33 Question 5

We saw the implementation of MLP in PyTorch with XOR data. What happens if we add 10 more hidden layers with 100 weights each with non-linear activation and train the model till loss is minimized.

(A) Can not classify all 4 samples correctly.

(B) Results in the same decision boundary

(C) Results in a different decision boundary but still able to classify all 4 samples correctly.

(D) None of these

C

Class review 34 Question 1

Which of these are valid ways to address the problem of overfitting?

(A) Add dropout to one of the layers

(B) Increase the number of layers in model

(C) Decrease the number of layers in model

(D) L1 regularization of weights

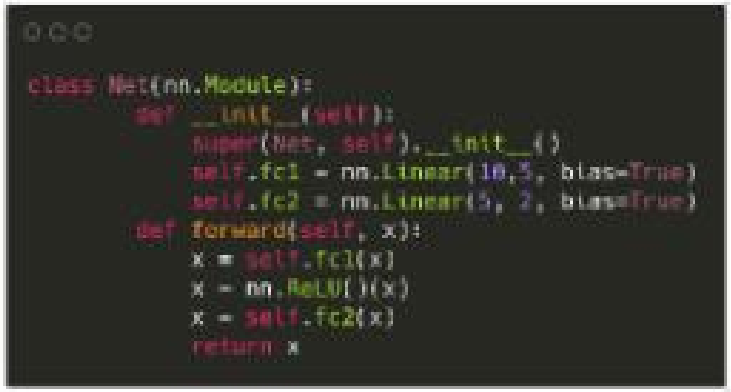
(E) Data augmentation

(F) None of these

A,C,D,E

Class review 34 Question 2

Consider this model:



```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.fc1 = nn.Linear(10,5, bias=True)
        self.fc2 = nn.Linear(5, 2, bias=True)
    def forward(self, x):
        x = self.fc1(x)
        x = nn.ReLU()(x)
        x = self.fc2(x)
        return x
```

Now consider that a tensor of shape $(5, 10)$ is given as input to this model. What is the shape of the output tensor?

(A) $(5, 5)$

(B) Cannot be answered with given information

(C) $(5, 2)$

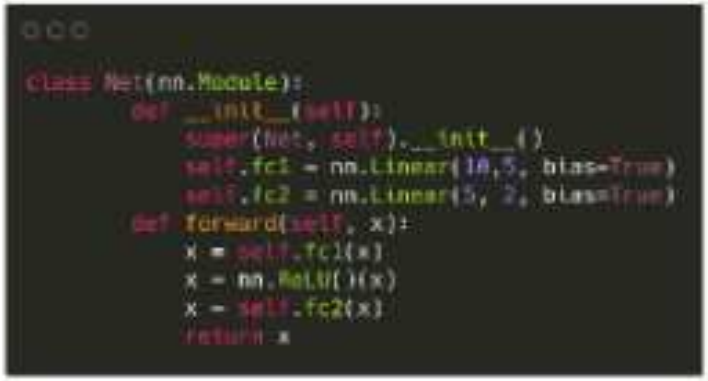
(D) $(10, 2)$

(E) None of these

C

Class review 34 Question 3

Consider this model:



```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.fc1 = nn.Linear(10,5, bias=True)
        self.fc2 = nn.Linear(5, 2, bias=True)
    def forward(self, x):
        x = self.fc1(x)
        x = nn.ReLU()(x)
        x = self.fc2(x)
        return x
```

What is the number of trainable parameters in this model

(A) 60

(B) 67

(C) Depends on the input

(D) 62

(E) None of these

B

Class review 34 Question 4

Consider this model:

```
class Net(nn.Module):
        def __init__(self):
            super(Net, self).__init__()
            self.fc1 = nn.Linear(10,5, bias=True)
            self.fc2 = nn.Linear(5, 2, bias=True)
        def forward(self, x):
            x = self.fc1(x)
            y = nn.ReLU()(x)
            z = self.fc2(y)
            return z
```

Now consider that a tensor of shape $(5, 10)$ is given as input to this model. What is the shape of the tensor y in the forward function ?

(A) Cannot be answered with given information

(B) $(5, 2)$

(C) $(10, 2)$

(D) $(5, 5)$

(E) None of these

D

Class review 34 Question 5

Consider this model with just 1 convolutional layer.

```
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv = nn.Conv2d(in_channels=3,out_channels=64,
                              kernel_size=(5,5), bias=False)
    def forward(self, x):
        x = self.conv(x)
        return x
```

Consider that we want to input a tensor of shape (10,3,512,512) to this model. Then

(A) The number of trainable parameters is $64 \times 3 \times 5 \times 5$

(B) The shape of output tensor is $(10, 64, 508, 508)$

(C) The number of trainable parameters is $64 \times 5 \times 5$

(D) The shape of output tensor is $(10, 3, 508, 508)$

(E) None of these

A,B

Class review 35 Question 1

While re-using a trained network for a new task:

(A) All the layers are equally useful.

(B) Which layer is more appropriate depends on the tasks.

(C) We always prefer to take the later (towards the end) layer

(D) We always prefer to take an early(in the beginning) layer

(E) None of the above.

D,(B)

Class review 35 Question 2

It is believed that adding noise is some sort of regularization.

(A) Lower the noise the better the regularization

(B) Adding noise to the weights is useful.

(C) Higher the noise the better the regularization.

(D) Adding noise to the output/labels is useful. (for simplicity, assume the task is regression!).

(E) Adding noise to the input is useful.

B,C,E

Class review 35 Question 3

Which of the following regularization in NN (implemented in PyTorch) lead to sparse

(A) L1 regularization

(B) Data Augmentation

(C) Dropout

(D) L2 regularization

(E) None of the above

A,C

Class review 35 Question 4

A sparse set of weights in a Deep MLP is preferred:

(A) it could lead to better generalization

(B) it is easy to train when the number of weights/parameters are less

(C) it has many zeros and lesser amount of operations in forward pass

(D) it is compact and fit in lesser memory

(E) All the above

A,B,C,D,E

Class review 35 Question 5

Consider a problem where we do data augmentation and early stopping.

(A) With data augmentation, training accuracy may decrease.

(B) With data augmentation, the iteration where we do early stop, will increase.

(C) With data augmentation, performance on the validation set is expected to increase.

(D) With data augmentation, training accuracy is expected to increase.

(E) With data augmentation, the iteration where we do early stop, will decrease.

A,E,C

Quiz 2 Questions

Quiz 3, Question 1

Consider an MLP with two inputs, three hidden neurons and one output neurons. Hidden neurons and output neurons have sigmoid activation. There is no bias. Output neuron has a MSE loss.

Consider a sample $([5,5]^T, 0.7)$ i.e., $x = [5,5]^T$ and $y = 0.7$. We would like to update all the weights based on the gradient of the loss ($\mathcal{L}$). Assume that $w_{ij}^{[k]}$ connects ith neuron of layer k with jth neuron of layer k+1. Thus weights betwee input and hidden layer are $w_{11}^{[1]}, w_{21}^{[1]}, w_{12}^{[1]}, w_{22}^{[1]}, w_{13}^{[1]}, w_{23}^{[1]}$ and those between hidden layer and output layer are $w_{11}^{[2]}, w_{21}^{[2]}, w_{31}^{[2]}$

Find the numerical value of $\frac{\partial \mathcal{L}}{\partial w_{11}^{[1]}}$. Answer upto 4 decimal places.

Quiz 3, Question 2

Consider a single layer perceptron with two input and one output. The weights from from first and second inputs are w_1 and w_2 respectively. Also assume a -1, +1 logic. Let w_0 be the weights associated with bias +1.

The activation at the output is:

$$\phi(x) = +1 \text{ if } x \geq 0 \text{ and } -1 \text{ else}$$

If $w_0 = 1, w_1 = -1, w_2 = -1$, then this perceptron is equivalent to:

(fill from the gates like: AND, OR, ExOR, NAND, NOR)

Quiz 3, Question 3

Remember the SVM problem from the problems we solved in the class. (1D samples)

$$(-1,+1),(0,-1),(+1,-1)$$

we geometrically solved the problem and saw the optimal primal solution as $w=-2$ and $b=-1$

Assume the samples were

$$(0,+1),(+1,-1),(+2,-1)$$

geometrically solve and give the answer as w=——,b=——

w=-2, b=1

Quiz 3, Question 4

Consider the following 10 samples used for training a Kernel SVM with $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T\mathbf{q})^2$. Labels are also given.

$$([-1,-1]^T,+1),([1,1]^T,+1)([+3,+4]^T,+1),([0,0]^T,-1)([10,10]^T,-1)$$

$$([0,1]^T,+1),([-10,-10]^T,-1)([1,0]^T,-1),([-2.5,-3.5]^T,+1),([4.5,6.5]^T,-1)$$

corresponding α are:

$$0, 1, 0, 1, 0, 2, 0, 1, 0, 0$$

(α *values are scaled/adjusted to make the numerical computation simpler!*)
Assume $b = 0$.

Consider at the test time, we have a sample $[-2,-2]^T$ Is this sample in positive class or negative class?

Quiz 3, Question 5

Consider an MLP with two input, one output and one hidden layer with two neurons. No bias. All weights are 1.0.

Hidden neurons have ReLu Activation and output has tanh activation.

Find the output of this MLP for an input of $[-2, -3]^T$

Quiz 3, Question 6

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

The optimization problem that **Logistic Regression** solves is *not convex*.

Quiz 3, Question 7

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

The deeper the decision tree **the better the decision tree as per Occam's razor.**

Quiz 3, Question 8

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

The optimal solution to PCA and LDA *are always orthogonal.*

Quiz 3, Question 9

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

Consider an MLP with 5 layers with all linear activations and MSE loss. The problem of training this MLP *is not possible* **with back propagation algorithm.**

Quiz 3, Question 10

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

ReLu is a *linear* activation function.

Quiz 3, Question 11

Consider the popular activation function Leaky-ReLu.

(A) its gradient can be either positive or. negative.

(B) its value can be either positive or negative

(C) it is an increasing function.

(D) it is a non-decreasing function

(E) all the above

Quiz 3, Question 12

For Kernel Percepron

(A) It can be used for linearly separable or non-separable data

(B) At test time, we evaluate it as:

$$sign(\mathbf{w}^T\mathbf{x})$$

(C) At the test time, we evaluate it as:

$$sign(\sum_{i=1}^{N} \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x})$$

(D) At the test time, we evaluate it as:

$$sign(\sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})$$

(E) when kernel is linear kernel, Kernel Perceptron reduces to the regular Perceptron.

Quiz 3, Question 13

Consider an MLP with one hidden layer. **x** is the input and **y** is the output. All neurons in the hidden and output have ReLU activation.

(A) This network is not appropriate for learning functions which can also take negative values as outputs.

(B) This network assumes **x** has only positive elements.

(C) While trained with BP, this network will have all weights positive.

(D) While trained with BP, this network will have all weights non-negative.

(E) All the above.

Quiz 3, Question 14

Consider an MLP which is getting trained with Back Propagation for a multiclass classification problem.

(A) The optimization problem we solve is convex if the number of classes is two.

(B) The optimization problem we solve is non-convex independent of the number of classes.

(C) We typically terminate the training when we reach a local minima (ie., GD can not change the solution)

(D) When we stop the training with an "early stopping criteria", the solution is often not a local minima.

(E) None of the above.

B,C,D

Quiz 3, Question 15

Consider a deep MLP and shallow MLP. Both gives the same loss and accuracy on the training data trained with the same number of samples.

(A) We prefer deep MLP (since deep neural networks are the best as of now)

(B) We prefer shallow MLP

(C) Both are equally good.

(D) Both neural networks then represent the same function. (since the loss is equal on both)

(E) None of the above.

Quiz 3, Question 16

Consider two quadratic kernels: $\kappa_1(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T\mathbf{q} + 1)^2$ and $\kappa_2(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T\mathbf{q})^2$.

Make the necessary minimal changes (if any required) and rewrite as true sentences in the space provided. Avoid changing the words in bold.

$\kappa_3() = \kappa_1() + \kappa_2()$ is also a valid kernel.

Quiz 3, Question 17

Consider a two class classification problem in 2-dimension with 6 data points.

$$\mathcal{D} = \{([0,0]^T, -), ([1,0]^T, -), ([0,1]^T, -), ([1,1], +), ([2,2]^T, +), ([2,0]^T, +)\}$$

We construct a hard margin SVM solution for this problem. The following is a support vector:

(A) $[0,0]^T$

(B) $[1,1]^T$

(C) $[2,2]^T$

(D) $[\frac{3}{2},0]^T$

(E) $[0,2]^T$

Quiz 3, Question 18

Consider an MLP with 3 inputs, two hidden layers of 5 neurons each and two output neurons. All neurons have sigmoid activation. All hidden neurons have bias.

How many learnable parameters are there in this network?

Quiz 3, Question 19

If there are 5 classes, a DDAG based multi class classifier will require —— binary classifiers to build the. DDAG.

Quiz 3, Question 20

Consider a two class classification problem in 2 dimensions. We know that both the classes can be modelled as multivariate Gaussians. We have 1000 samples each from both the classes (i.e., N=2000).

If means are always well separated and variances are always small:

We use a linear SVM.

(A) number of support vectors will be very small (say closer to d than closer to N)

(B) number of support vectors will be very larger (say closer to N than closer to d).

(C) in general, number of support vectors have nothing to do with the mean and variance of the classes.

(D) in general, number of support vectors depends on mean but not variance.

(E) in general, number of support vectors depends on variance and not mean.

A/AC