1 Longer Problem

1: MLP gradient

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 between 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.

   FIB 0.0228

2. 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 between 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 $\frac{\partial \mathcal{L}}{\partial w_{12}^{[1]}}$. Answer upto 4 decimal places.

   FIB 0.0228

3. 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 between 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 $\frac{\partial \mathcal{L}}{\partial w_{13}^{[1]}}$. Answer upto 4 decimal places.

   FIB 0.0228

4. 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 between 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 $\frac{\partial \mathcal{L}}{\partial w_{23}^{[1]}}$. Answer upto 4 decimal places.

   FIB 0.0228

5. 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 between 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 $\frac{\partial \mathcal{L}}{\partial w_{11}^{[2]}}$. Answer upto 4 decimal places.

FIB 0 (5.18e-6)

6. 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 between 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 $\frac{\partial \mathcal{L}}{\partial w_{31}^{[2]}}$. Answer upto 4 decimal places.

   FIB (5.18e-6)

2: SLP and implementation of Gates

1. 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 $w_0 = XX, w_1 = XX, w_2 = XX$, then this perceptron is equivalent to:

   (fill from the gates like: AND, OR, ExOR, NAND, NOR)

   FIB Examples: w0, w1, w2 -¿ Answer 0, +1, +1 -¿ OR -1, +1,+1 -¿ AND 1, -1, -1 -¿ NAND -1, -1, -1 -¿ NOR

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 $w_0 = 0, w_1 = 1, w_2 = 1$, then this perceptron is equivalent to:

   (fill from the gates like: AND, OR, ExOR, NAND, NOR)

   FIB OR

3. 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)

   FIB AND

4. 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)

FIB NAND

5. 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)

   FIB NOR

3: SVM

1. 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

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

   geometrically solve and give the answer as w=——,b=——

   FIB w=-1, b=-1

2. 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

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

   geometrically solve and give the answer as w=——,b=——

   FIB w=2, b=1

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

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

   geometrically solve and give the answer as w=——,b=——

   FIB w=-1,b=1

4. 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

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

geometrically solve and give the answer as w=——,b=——

FIB w=1, b=1

5. 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=——

FIB w=-2, b=1

4: Kernels

1. 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 $[-1,-1]^T$. Is this sample in positive class or negative class?

FIB positive 5

2. 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?

FIB positive 20

3. 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?

FIB positive 20

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 $[1,-2]^T$ Is this sample in positive class or negative class?

FIB positive 8

5. 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 $[-2,1]^T$ Is this sample in positive class or negative class?

FIB negative -1

5: Forward Pass

1. 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 $[1,-2]^T$

   FIB 0 tanh(relu(1-2))

2. Consider an MLP with two input, one output and one hidden layer with two neurons. No bias. All weights are 2.0.

   Hidden neurons have ReLu Activation and output has tanh activation.

   Find the output of this MLP for an input of $[1,-2]^T$

   FIB 0 tanh(2*relu(2-4))

3. 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 $[1, -2]^T$

   FIB -0.9640 tanh(-1*relu(-1+2)) = tanh(-1)

4. 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 $[-2, -3]^T$

   FIB 0 tanh(relu(-2-3))

5. Consider an MLP with two input, one output and one hidden layer with two neurons. No bias. All weights are unity.

   Hidden neurons have ReLu Activation and output has tanh activation.

   Find the output of this MLP for an input of $[2, -3]^T$

   FIB 0 tanh(relu(2-3))

2 Conceptual FIB

1 Logistic Regression

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

   Logistic Regression is a popular algorithm for *regression problem.*

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

   Logistic Regression outputs *probability of a sample to a class.* No change

3. 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 *convex.* No change

4. 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.* convex

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

   The Loss function that **Logistic Regression** uses *is hinge loss.* Cross entropy loss

2 Multi-Class Classifiers

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

   The number of leaves of an unpruned decision tree classifier with K classes with at least one sample per class *will be less/more/equal than K* equal to or more than

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

   The number of binary classifiers in a DDAG classifier with K classes where $K > 2$ *will be less/more/equal than* K equal to or more than [MORE IS CORRECT]

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

   The number of binary classifiers in a DDAG classifier with K **classes to be evaluated at the test time** *will be less/more/equal than* K less (always K-1)

4. 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.** shallower

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

   Deep decision trees *are prone to overfitting.* No change

3: FDA/Fisher

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

   Fisher discriminant *maximizes* the between class scatter and *minimizes* the within class scatter. No change

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

   Fisher discriminant *minimizes* the between class scatter and *maximizes* the within class scatter. minimizes within class, maximizes between class scatter

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

   Principal Component Analysis maximizes the between class scatter and minimizes the within class scatter. PCA maximizes overall variance/scatter

4. 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 not always orthogonal.* do not exhibit any relation

5. 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 never orthogonal.* do not exhibit any relation

4: Optimization

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

   While training, the optimization problem that MLP solves *is concave.* neither convex nor concave

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

   While training, an MLP with a hinge loss *solves a non convex optimization problem.* neither convex nor concave

3. 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 convex.* No change

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 an MLP with 5 layers with all linear activations and MSE loss. The problem of training this MLP *is non-convex.* convex

5. 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.** possible

5: Activations

1. 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 with linear activation *can solve ExOR problem* Cannot solve

2. 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 *non-linear* **activation function.** No change

3. 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.** Non linear

4. 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 *piece-wise linear* **activation function.** NO change

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

   Derivateive of ReLu at x *and* y *are same* **when** xy **is positive**. No change

3 Conceptual MCQ

1: Activation Functions

1. 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

   CD

2. Consider the popular activation function 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

CD

2 Kernels

1. 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.

   ADE /ACE/ACDE

2. Consider a set of N valid kernels $\kappa_i(\cdot,\cdot)$

   (a) $\sum_{i=1}^{N}\kappa_i()$ is also a valid kernel.

   (b) $\sum_{i=1}^{N}\alpha_i\kappa_i()$ is also a valid kernel for any $\alpha_i \in R$.

   (c) $\sum_{i=1}^{N}\alpha_i\kappa_i()$ is also a valid kernel for any $\alpha_i \in R^+$.

   (d) $\Pi_{i=1}^{N}\kappa_i()$ is also a valid kernel.

   (e) All the above.

   ACD

3 MLP

1. Consider an MLP with one hidden layer. $\mathbf{x}$ is the input and $\mathbf{y}$ 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 $\mathbf{x}$ 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.

   A

2. Consider an MLP with one hidden layer. $\mathbf{x}$ is the input and $\mathbf{y}$ is the output. All neurons in the hidden and output have ReLU activation.

   (a) This network can be reduced to $\mathbf{y} = \mathbf{W}\mathbf{x}$

   (b) This network can be modelled as: "Either $\mathbf{y} = \mathbf{W}_1\mathbf{x}$ or $\mathbf{y} = \mathbf{W}_2\mathbf{x}$"

   (c) If all elements of $\mathbf{x}$ are negative, $\mathbf{y} = \mathbf{0}$.

   (d) If $\mathbf{y} = \mathbf{0}$ imply that at least some of the elements of $\mathbf{x}$ are negative.

   (e) None of the above.

   B /E

   W1W2X if $\mathrm{W}_1 W_2 X > 0, W_1 X > 0 otherwise 0$

4 Backpropagation

1. Consider an MLP which is getting trained with Back Propagation for a multiclass classification problem.

   (a) The performance of the final model will depend on the initialization.

   (b) The performance of the final model will depend on the learning rate we use.

   (c) The performance of the final model will depend on the termination criteria we use.

   (d) The performance of the final model will depend on the loss function we use.

   (e) Exactly three of the above four are correct.

   ABCD

2. 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.

   BC/BCD

5 Deep

1. 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.

   B

2. Consider a deep MLP and shallow MLP. Both are trained with the same number of samples.

(a) It is highly likely that Deep MLP will have lower training error. (since deeper the powerful!)

(b) It is highly likely that the shallow MLP will have lower training error. (since Occam's Razor says so)

(c) If the number of training samples is small, Deep MLP is going to overfit.

(d) If the number of training samples is small, Shallow MLP is going to overfit.

(e) None of the above

AC

4 Simple Numerical/Analytical (?) FIB

1:Kernels

1. 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.

   $\kappa_1(\cdot, \cdot)$ is a *valid kernel*; and $\kappa_2(\cdot, \cdot)$ is a *invalid kernel*;

   FIB Both K1, K2 are valid kernels

2. 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.

   FIB No change

3. 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.

   FIB No change 1 + 2*p.T.q is a valid kernel

4. 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.

   Both $\kappa_1(\cdot, \cdot)$ **and** $\kappa_2(\cdot, \cdot)$ *have identical feature maps* $\phi()$.

   FIB distinct feature maps

5. 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.

   Both $\kappa_1(\cdot, \cdot)$ **and** $\kappa_2(\cdot, \cdot)$ *have distinct feature maps* $\phi()$.

   FIB No change

2: SVMs

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.

Ans: AD

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. 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) $[0,2]^T$

Ans: B Support vectors: $[1,0], [0,1], [2,0], [1,1]]$

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) If we remove any one of the support vectors from the training data and retrain the SVM, we will get a different solution.
(b) For this problem, there exists at least one sample, removal of it will lead to a different solution for the SVM.
(c) 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.
(d) 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.
(e) None of the above.

Ans: B On removing the points $[0,1]$ or $[2,0]$ the line remains the same but the support vector size decreases by 1. This results in same margin. On removing the points $[1,0]$ the line changes to $x = .5$ and the support vector size decreases by 1. On removing the points $[1,1]$ the line changes to $x = 1.5$ and the support vector size decreases by 1. Both of these lead to increased margin of 0.5 from $\frac{1}{\sqrt{8}}$. Removal of non support vector should not affect anything. Adding new support vector does not change status of existing support vector if it is on of the boundaries, for instance $X + Y = 2$ in this example

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.

(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 $[1,0]^T$ from $\mathcal{D}$, the margin increases.
(d) If we remove $[1,1]^T$ from $\mathcal{D}$, the margin increases.
(e) If we remove $[2,2]^T$ from $\mathcal{D}$, the margin increases.

Ans: CD On removing the points $[1,0]$ the line changes to $x = .5$. On removing the points $[1,1]$ the line changes to $x = 1.5$. Both of these lead to increased margin of 0.5 from $\frac{1}{\sqrt{8}}$.

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) Addition of $([0,2]^T, +)$ will change the support vector set, but not the margin.
(b) Addition of $([0, \frac{3}{2}]^T, +)$ will change the support vector set, and the margin.
(c) Addition of no sample can increase the margin.
(d) Addition of $([1,2]^T, +)$ does not change the support vector set and the margin.
(e) Addition of $([0, \frac{3}{2}]^T, +)$ will change the support vector set, but the number of support vectors will not change.

Ans: ABCD $[0,2]$ is on of the boundaries as described above; will be part of the new support vector set but the line and the margin remains same. (A) $[0, 1.5]$ will decrease margin and it will be part of the new support vector set, removing 2 points from earlier set. (B,E) $[1,2]$ is too far on the + side to affect anything.(D) C is true in general in SVMs.

3: MLP

1. Consider an MLP with 3 inputs, two hidden layers of 5 neurons each and two output neurons. All neurons have sigmoid activation. no bias.

   How many learnable parameters are there in this network? (3)*5 + (5)*5 + (5)*2 50

2. 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? (3+1)*5 + (5+1)*5 + (5)*2 60

3. Consider an MLP with 3 inputs, two hidden layers of 5 neurons each and two output neurons. All neurons have sigmoid activation. All neurons have bias.

   How many learnable parameters are there in this network? (3+1)*5 + (5+1)*5 + (5+1)*2 62

4. Consider an MLP with 4 inputs, two hidden layers of 5 neurons each and two output neurons. All neurons have sigmoid activation. All neurons have bias.

   How many learnable parameters are there in this network? (4+1)*5 + (5+1)*5 + (5+1)*2 67

5. Consider an MLP with 4 inputs, two hidden layers of 5 neurons each and one output neuron. All neurons have sigmoid activation. No bias.

   How many learnable parameters are there in this network? (4)*5 + (5)*5 + (5)*2 50

4 DDAG

For N classes DDAG, we require (N choose 2) binary classifiers and need to evaluate N-1 binary classifiers during inferenceIf there are 5 classes, a DDAG based multi class classifier will require —— binary classifiers to build the. DDAG. FIB 10 If there are 5 classes, a DDAG based multi class classifier will require evaluation of —— binary classifiers to make a decision. FIB 4 If there are 10 classes, a DDAG based multi class classifier will require —— binary classifiers to build the. DDAG. FIB 45 If there are 10 classes, a DDAG based multi class classifier will require evaluation of —— binary classifiers to make a decision. FIB 9 "Since for a K class problem, DDAG uses ${}^{K}C_{2}$ classifiers, the final decision can be ambigous". (Write Tue or False) FIB False

5 SV

1. 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 large (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 There will be most likely only two support vectors even after changing mean/variance

2. 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 equal and variances are always equal for both the classes:

   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 large (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.

   B/BC almost all of them will be support vectors no matter the mean and variance

3. 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).

   Bayesian Optimal Classifier gives 90% as the optimal accuracy.

   We use a linear SVM.

   (a) number of Support Vectors will be closer to 0.9 N.

   (b) number of Support Vectors will be closer to 0.9 d.

(c) number of Support Vectors will be closer to 0.1 N.

(d) number of Support Vectors will be closer to 0.1 d.

(e) Bayesian optimal rate has no influence on the number of Support Vectors.

C

4. Consider a two class classification problem in d dimensions. We know that both the classes can be modelled as multivariate Gaussians. We have $\frac{N}{2}$ samples each from both the classes (i.e., total of N).

   If $N << d$

   We use a linear SVM.

   (a) number of support vectors will be closer to d than closer to N

   (b) number of support vectors will be 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.

   B Changes in mean and variance may cause overlap

5. Consider a two class classification problem in d dimensions. We know that both the classes can be modelled as multivariate Gaussians. We have $\frac{N}{2}$ samples each from both the classes (i.e., total of N).

   If $d << N$

   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 large (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.

   I think it should depend on both mean and variance because overlap will be decided by both We cant say about A or B (depends on position of means)