Time for another one

175292

177863

175838

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
176165

176288

175252, 175151

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.

MLP three hidden neurons

175601, 175584, 175591

Single layer perceptron two input and or exor nand nor

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)

179755

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.

Consider Construct hard margin svm

176167

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?

Obj. fn. $J_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \sum_{i}^{N} \sum_{j}^{N} \alpha_i \alpha_j y_i y_j \, k(x_i, x_j)$

sign$\left(\sum_{i=1}^{N} \alpha_i y_i \, k(x_i, x) + b\right.$

$k(x_i, x) = \phi(x_i)^T \phi(x)$

Kernel svm positive or negative class
175714, 175875

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=——

20181057 2018101

svm geometrically solved optimal primal
176063, 176172

180287

If there are 5 classes, a DDAG based multi class classifier will require evaluation of —— binary classifiers to make a decision.

Answer: 10? (tiw)
178750,

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.

176305, 176337, 176264

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 $[1, -2]^T$

mlp two input no bias relu
176708, 176483

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*.

The Loss function that Logistic Regression uses is logarithmic loss

177011

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*.

Deep decision trees are prone to overfitting - labels not answers

176484

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*.

Linear Regression is a popular algorithm for regression problem (Kamble)
176760

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*

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
less than K (Kamble)
177278, 177279, 177264

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.*

The optimal solution to PCA and LDA can be orthogonal (Kamble)
177515

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.*

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.**

Consider an MLP with 5 layers with all linear activations and MSE loss. The problem of training this MLP is non-convex
177775

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.**

Anirudh: ReLu is a rectified linear activation function (?) (Non linear: Kamble)
176920

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

177379

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.*

177567

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.**

176631

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.*

178046

Quiz 3, Question 11

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

D (chan)is

178241

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.

Anirudh: A, D, E (pls confirm) (C is for Kernel SVM)

178546

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.

A ?(Sumba) D too? (Kamble)

178859, 178831, 178759

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.

179072, 179146, 179252

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 - risubh

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

179354, 179327, 179583

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*;

179873

180102

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?

Answer: 69

(4 * 5 + 5 * 5 + 5 * 2) (weights) + (5 + 5 + 4) (biases)

(tiw)

180048, 179931

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?

180422

"Since for a K class problem, DDAG uses ${}^{K}C_2$ classifiers, the final decision can be ambigous". (Write Tue or False)

Anirudh: False

180585, 180533

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 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.

180592

176716

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

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* $\phi()$.

176826

177046

177351

177863

180741

180203

If there are 5 classes, a DDAG based multi class classifier wi
classifiers to build the. DDAG.

10

179883

179519

179253

178771

178509

178151/178379/178332

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.

Anwer: A, B, C, D, E (tiwari)

Proof: http://huisaddison.com/blog/cute-proof-about-kernels.html

https://stats.stackexchange.com/questions/177100/linear-combination-of-two-kernel-functions

177994

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

D(SUMBA) BCD(agoo) not d (Kamble) inc != non dec

175150

175366

175793
176133
176230
176633
176944
177245
177383
177786
177959
178312
178710
178739

178827

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.

178805
179161
179387

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.

179560
179659
179938
180427
180602

180506
180399
180134

Quiz 3, Question 18

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?

179650

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 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.

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.

179469 A, D

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.

007 811

179311

Quiz 3, Question 15

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

178795
178695
178424
178117
177739
177444
177272
176774
176695
176416
175963
175883
175591
175147
175501

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)

NAND - risubh
175584

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)

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)

180693

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.

2018114013 2018114013 2018114013 2018114013

(C) - risubh , sai

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 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.

180433

"Since for a K class problem, DDAG uses ${}^{K}C_2$ classifiers, the final decision can be ambigous". (Write Tue or False)

False(Anshul)

175875

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 (risubh)

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=——

W = -2 , b = -1 (risubh)

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=——

W = -1 , b = -1 (risubh)
175617

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=——

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?

50(Anshul)
55 (4 inputs)(Anirudh)
62(with bias manvith)

179649

179604

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* $\phi()$.

True - risubh

178264

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.

a,e (SUMBA)

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.**

2018111025 2018111025 2018111025 2018111025