a
B

C

Considering a Singular Value Decomposition $X = U S V^T$, what are the special properties of matrices U, S and V ?

Select one:

- a. U, V are unitary and S is diagonal.
- b. U is orthogonal, V is upper triangular and S is symmetric.
- c. U, V are upper triangular, and S is diagonal.
- d. U, V are symmetric and S is orthogonal.

选A

Consider these two probability distributions on the same space Ω = {A, B, C, D, E}

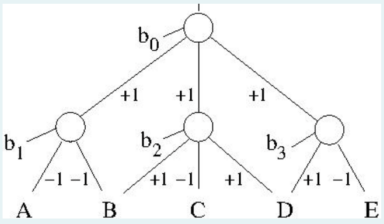
$p = \langle 1/16, 1/16, 1/8, 1/4, 1/2 \rangle$
$q = \langle 1/4, 1/16, 1/16, 1/2, 1/8 \rangle$

Compute (correct to at least two decimal places):

* The Entropy H(*p*):

* The KL-Divergence $D_{KL}(p \,||\, q)$:

Consider the following multi-layer perceptron, using the threshold activation function, and assume that TRUE is represented by 1; FALSE by 0.



For which values of the biases b_0, b_1, b_2 and b_3 would this network compute the logical function

(¬A ∨ ¬B) ∧ (B ∨ ¬C ∨ D) ∧ (D ∨ ¬E)

* b_0=
* b_1=
* b_2=
* b_3=

-2.5, 1.5 , 0.5, 0.5

Consider a Perceptron whose output is given by $h(w_0+w_1x_1+w_2x_2)$, where x_1, x_2 are inputs and $h()$ is the Heaviside (step) function.

Assume this Perceptron is being trained on the data in the following table, and that the current values of the weights are w_0= -0.5, w_1= -1 and w_2= -2.

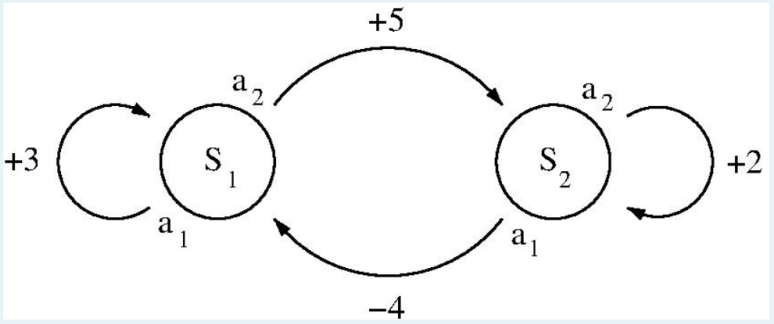
Training Example	x_1	x_2	Class
(a)	−2	2	Pos
(b)	2	1	Neg
(c)	−1	−1	Neg

If the Perceptron Learning Rule is applied to the current weights, using training item (a) and a learning rate of η=1.0, the new values for w_0, w_1 and w_2 at the end of this training step will be:

* w_0:
* w_1:
* w_2:

1.875 1.4375

Consider an environment with two states S = {S_1, S_2} and two actions A = {a_1, a_2}, where the (deterministic) transitions δ and reward R for each state and action are as follows:



Assuming a discount factor of γ = 0.7, determine:

* $\pi^*(S_1)$= ⬍⬍

* $\pi^*(S_2)$= ⬍⬍

Again assuming γ = 0.7, compute these values (correct to two decimal places):

* $Q^*(S_1,a_1)$

* $Q^*(S_1,a_2)$

* $Q^*(S_2,a_1)$

* $Q^*(S_2,a_2)$

If γ is allowed to vary between 0 and 1, for which range of values of γ is this policy optimal (correct to two decimal places)?

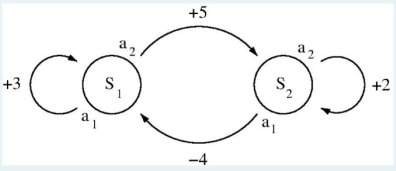
* Minimum value of γ:

* Maximum value of γ:

(1) a1 a2 (2)10 9.33 3 6.67 (3) 0.67 1

Time left 1:15:00

Consider an environment with two states S = {S_1, S_2} and two actions A = {a_1, a_2}, where the (deterministic) transitions δ and reward R for each state and action are as follows:



Assuming a discount factor of γ = 0.8, determine:

* $\pi^*(S_1)$= a1 ⬍⬍

* $\pi^*(S_2)$= a2 ⬍⬍

Again assuming γ = 0.8, compute these values (correct to two decimal places):

* $Q^*(S_1,a_1)$ 15

* $Q^*(S_1,a_2)$ 13

* $Q^*(S_2,a_1)$ 8

* $Q^*(S_2,a_2)$ 10

If γ is allowed to vary between 0 and 1, for which range of values of γ is this policy optimal (correct to two decimal places)?

* Minimum value of γ: 0.14

* Maximum value of γ: 0.75

A1 a2 15 13 8 10

D

For the Generative Adversarial Networks discussed in this course, the game between the Generator and Discriminator:

Select one:

- a. can be either zero-sum or not, but the zero-sum version produces better images
- b. is never zero-sum
- c. can be either zero-sum or not, but the non-zero-sum version produces better images
- d. is always zero-sum

E

The Actor-Critic algorithm combines:

Select one:

- a. Q-Learning and Policy Gradients
- b. two different Q-Learners (one for action selection, the other for value estimation)
- c. adversarially trained Generator and Discriminator networks
- d. Q-Learning and TD-Learning

选a

F

Which of these is NOT a method for dealing with the problem of vanishing or exploding gradients? Time le

Select one:

- a. Batch Normalization
- b. Conjugate Gradients
- c. Weight Initialization
- d. Rectified Linear Unit

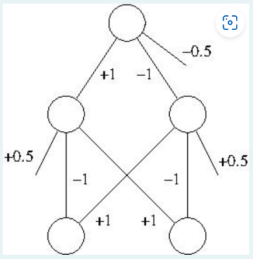
g
h
I

Time left 1:58:15

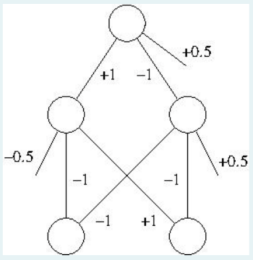
If 0=FALSE and 1=TRUE, which of these networks (with threshold activations at both the hidden and output layer) correctly computes the XOR function of two inputs?

Select one:

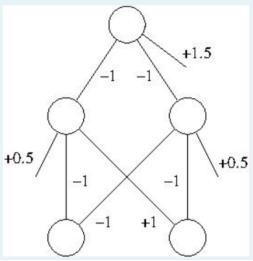
a.



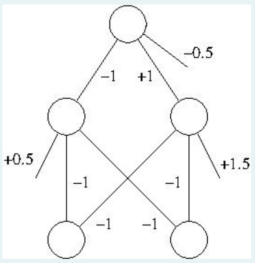
b.



c.



d.



Consider a neural network trained using softmax for a classification task with three classes 1, 2, 3. Suppose a particular input is presented, producing outputs

z_1= 1.3, z_2= 2.4, z_3= 3.1

Assuming the correct class for this input is Class 2, and that Prob(2) is the softmax probability of the network choosing Class 2, compute the following, to two decimal places:

* d(log Prob(2))/dz_1=
* d(log Prob(2))/dz_2=
* d(log Prob(2))/dz_3=

-0.10 0.70 -0.60
-0.10 0.70 -0.60

Consider a convolutional neural network which takes as input a 65-by-77 color image (i.e. with three channels R, G, B). The first convolutional layer has 18 filters that are 5-by-5, with stride 4 and no zero-padding.

Compute the number of:

* weights per neuron in this layer (including bias):
* neurons in this layer:
* connections into the neurons in this layer:
* independent parameters in this layer:

76 5472 415872 1368

Consider a convolutional neural network which takes as input a 65-by-77 color image (i.e. with three channels R, G, B). The first convolutional layer has 18 filters that are 3-by-3, with stride 2 and no zero-padding.

Compute the number of:

* weights per neuron in this layer (including bias): 28
* neurons in this layer: 21888
* connections into the neurons in this layer: 612864
* independent parameters in this layer: 504

Consider a Hopfield Network with the following weight matrix W:

| 0 0 -1 0 0 |
| 0 0 0 0 -1 |
| -1 0 0 +1 0 |
| 0 0 +1 0 +1 |
| 0 -1 0 +1 0 |

For each of the following vectors, state whether it is Stable or Not Stable for this network:

* [−1 +1 +1 +1 +1]:
* [+1 +1 −1 −1 −1]:
* [+1 +1 −1 +1 +1]:

Unstable stable unstable

Consider a Hopfield Network with the following weight matrix W:

| 0 0 -1 0 0 |
| 0 0 0 0 +1 |
| -1 0 0 -1 0 |
| 0 0 -1 0 +1 |
| 0 +1 0 +1 0 |

For each of the following vectors, state whether it is Stable or Not Stable for this network:

* [−1 +1 +1 +1 +1]: Stable
* [+1 +1 −1 −1 −1]: Not Stable
* [+1 +1 −1 +1 +1]: Stable

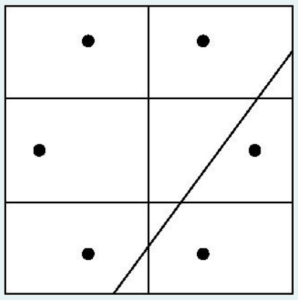
Consider a fully connected feedforward neural network with 6 inputs, 2 hidden units and 4 outputs, using tanh activation at the hidden units and sigmoid at the outputs. Suppose this network is trained on the following data, and that the training is successful.

Item	Inputs	Outputs
	123456	1234
1.	100000	0001
2.	010000	0011
3.	001000	0101
4.	000100	1000
5.	000010	1011
6.	000001	1100

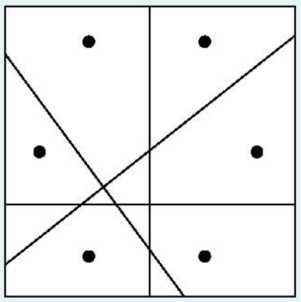
Which of these diagrams correctly shows a point in hidden unit space corresponding to each input, and, for each output, a line dividing the hidden unit space into regions for which the value of that output is greater/less than one half ?

Select one:

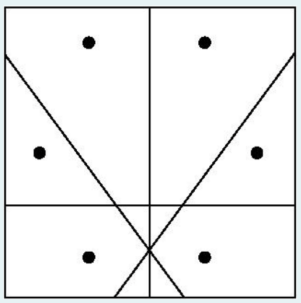
a.



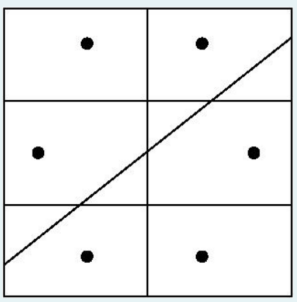
b.



c.



d.



J

K

l
m
n
O

Only 40% of the population have been vaccinated against a certain disease. Among those who **are** vaccinated, only 1% of them have the disease. But, among those who are **not** vaccinated, 2% of them have the disease.

If a random person is found to have the disease, what is the probability that they have been vaccinated?

(You can give your answer either as a percentage, or as a number between 0 and 1)

Answer:

0.125
0.25

p
q
R

Reinforcement Learning is when an agent is:

Select one:

- a. not presented with target outputs, but instead given a reward signal that it aims to maximize
- b. presented once with examples of inputs and their target outputs
- c. presented multiple times (over time) with the same examples of inputs and their target outputs
- d. only presented with the inputs and not target outputs, so it aims to find structure in these inputs

选A

s
T

The principle “The most likely hypothesis is the simplest one consistent with the data.” is called:

Select one:

- a. Ockham's Razor
- b. Bayes' Rule
- c. Perceptron Learning
- d. Maximum Likelihood

选a

The Actor-Critic algorithm combines:

Select one:

- a. two different Q-Learners (one for action selection, the other for value estimation)
- b. Q-Learning and Policy Gradients
- c. adversarially trained Generator and Discriminator networks
- d. Q-Learning and TD-Learning

选A

Two common methods for unsupervised pre-training of neural networks are:

Select one:

a. Autoencoder and Deep Boltzmann Machine

b. Bayesian Inference and Weight Initialization

c. Weight Initialization and Autoencoder

d. Deep Boltzmann Machine and Bayesian Inference

选A

The Context Layer in a Simple Recurrent Network:

Select one:

a. is comprised of the inputs in a sliding window around the current timestep

b. is computed from the current input and the previous hidden layer

c. is a copy of the hidden layer from the previous timestep

d. is computed from the current input and the previous output

选C

u

v

W

Question 1

Not yet answered

Marked out of 1.00

Flag question

Which of these is NOT a method for dealing with the problem of vanishing or exploding gradients?

Select one:

a. Conjugate Gradients

b. Rectified Linear Unit

c. Batch Normalization

d. Weight Initialization

Clear my choice

When training on linearly separable data using the Perceptron Learning Rule, what will happen if both the learning rate and the initial weights are scaled up by a large factor?

Select one:

a. The data will be learned successfully, in a smaller number of epochs

b. The data will be learned successfully, in about the same number of epochs

c. Learning may become unstable and fail to converge

d. The data will be learned successfully, but in a larger number of epochs

Which statement about word2vec is FALSE?

Select one:

a. It aims to maximise the log probability of a word, based on the surrounding words

b. Representations for the same word at the input and output layers are different

c. The tanh activation function is used at the hidden nodes

d. Performance improves if frequent words are sampled less often

Which architecture would have the best chance of learning an Embedded Reber Grammar?

Select one:

a. Gated Recurrent Unit

b. NetTalk system

c. Jordan Network

d. Elman Network

When using Batch Normalization, in the Testing phase, the Mean and Variance of the activations at each node are typically:

Select one:

- a. pre-computed from the training set
- b. estimated using running averages
- c. either of the above
- d. none of the above

选C

When comparing a Hopfield Network with a Boltzmann Machine, which statement is FALSE?

Select one:

- a. The updates are deterministic for one model, and stochastic for the other
- b. The formula for the energy function is different for the two models
- c. The range of activations is {-1,1} for one model and {0,1} for the other
- d. One model is used for retrieval, the other for generation

Which of these statements about Dropout is FALSE:

Select one:

- a. Dropout encourages redundancy
- b. Dropout encourages the weight values to be small
- c. Dropout helps to prevent overfitting
- d. Dropout simulates an ensemble of network architectures

选B

Whit type of Autoencoder explicitly forces the hidden features not to change much when the inputs are slightly altered?

Select one:

- a. Variational Autoencoder
- b. Contractive Autoencoder
- c. Denoising Autoencoder
- d. Sparse Autoencoder

选B

When using Batch Normalization, in the Testing phase, the Mean and Variance of the activations at each node are typically:

Select one:

- a. pre-computed from the training set
- b. estimated using running averages
- c. either of the above
- d. none of the above

Clear my choice

Which architecture would have the best chance of learning an Embedded Reber Grammar?

Select one:

- a. Jordan Network
- b. NetTalk system
- c. Elman Network
- d. Gated Recurrent Unit

Which statement about word2vec is FALSE?

Select one:

- a. The tanh activation function is used at the hidden nodes
- b. Performance improves if frequent words are sampled less often
- c. Representations for the same word at the input and output layers are different
- d. It aims to maximise the log probability of a word, based on the surrounding words

When comparing a Hopfield Network with a Boltzmann Machine, which statement is FALSE?

Select one:

- a. The range of activations is {-1,1} for one model and {0,1} for the other
- b. One model is used for retrieval, the other for generation
- c. The formula for the energy function is different for the two models
- d. The updates are deterministic for one model, and stochastic for the other

x
y
Z