20233_csci_544_30249: Applied Natural Language Processing

Announcements
Syllabus
Content
Assignments

Tools
USC Code of Ethics

Review Test Submission: Midterm

User	Kayvan Shah
Course	20233_csci_544_30249: Applied Natural Language Processing
Test	Midterm
Started	10/17/23 5:30 PM
Submitted	10/17/23 6:40 PM
Due Date	10/17/23 6:41 PM
Status	Completed
Attempt Score	459.99664 out of 600 points
Time Elapsed	1 hour, 9 minutes out of 1 hour and 10 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Incorrectly Answered Questions

Question 1

10 out of 10 points

When using an HMM for POS tagging, the emission probability is defined as:

Selected Answer: ✓ The probability of a word given a specific POS tag.

Answers:
The probability of transitioning from one POS tag to another.
✓ The probability of a word given a specific POS tag.
The likelihood of observing a particular word sequence in a corpus.
The likelihood of a POS tag sequence appearing in a random sentence.

Question 2

0 out of 10 points

47: Word2Vec use ____ word, w_a, to predict ____ word, w_b, such that p(__|__) is maximized.

Selected Answer: ✗ context, center, w_a, w_b

Answers:
✓ context, center, w_b, w_a
center, context, w_a, w_b
center, context, w_b, w_a
context, center, w_a, w_b

Question 3

10 out of 10 points

In PyTorch, which module is primarily utilized for defining a model's architecture (layers and operations)?

Selected Answer: ✓ torch.nn

Answers:
✓ torch.nn
torch.data
torch.optim
torch.utils

Question 4

10 out of 10 points

Document level representation is more suitable than word-level representation to do semantic inference task.

Selected Answer: ✓ False

Answers:
True
✓ False

Question 5

10 out of 10 points

Why are N-gram models often insufficient for understanding the meaning of a sentence in NLP?

Selected Answers:
✗ They do not consider word order
✓ They consider only local context

Answers:
They do not consider word order
They consider only global context

They consider only local context

They are computationally intensive

Question 6 6.66666 out of 10 points

Select all that are **incorrect** on the Bag of Word (BoW) method:

Selected Answers:
- Preserves order of words
- Considers word dependency

Answers:
- BoW leads to acceptable performance in some applications
- Preserves order of words
- Considers word dependency
- It is a simple method to implement
- Results in dense vectors

Question 7 10 out of 10 points

Dot product similarity is not a meaningful similarity metric on the one-hot word representations.

Selected Answer: True

Answers:
- True
- False

Question 8 5 out of 10 points

In the context of training neural networks, which of the following are purposes of using mini-batches? (Select all that apply)

Selected Answers:
- To reduce computational load
- To ensure all examples are used in every training step
- To utilize a mix of global and sample-specific information for updating parameters

Answers:
- To reduce computational load
- To ensure all examples are used in every training step
- To utilize a mix of global and sample-specific information for updating parameters
- To ensure deterministic convergence to a global minimum

Question 9 10 out of 10 points

Regarding the distinction between local and global ambiguity, please identify the correct instances of ambiguity in the following three sentences:

1. "I watched her dog as I painted the fence."
2. "She has a bright mind, and he's her sunshine."
3. "The sun rises in the east."

Select the sentence(s) that exhibit(s) ambiguity:

Selected Answer: Sentences 1 and 2

Answers:
- Sentence 2 only
- Sentences 3 and 2
- Sentences 1 and 2
- Sentence 1 only

Question 10 10 out of 10 points

Imagine that we have mapped every word in English to a normalized vector in R^n. What is the squared Euclidean distance between two vectors a and b in this space, given that θ denotes the angle between the two vectors.

Selected Answer: $2(1-\cos\theta)$

Answers:
- $2(1-\cos\theta)$
- $2(1+\sin\theta)$
- $2\sin^2\left(\frac{\theta}{2}\right)$
- $2(1-\sin\theta)$

Question 11 3.33333 out of 10 points

In the context of perceptron optimization, when might the perceptron algorithm still converge?

Selected Answers: When the training data is perfectly linearly separable.
When the training data is not linearly separable.

Answers: When the training data is perfectly linearly separable.
When the input data is extremely high-dimensional.
When the training data is not linearly separable.
When the model is regularized too heavily.

Question 12

10 out of 10 points

Consider a dataset with n documents and a vocabulary of size m. Let $A \in \mathbb{R}^{m \times n}$ denotes the term-document matrix. We decompose this matrix using SVD into $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$ and $V \in \mathbb{R}^{n \times n}$ and keep the results in three tensors in PyTorch. Which of the following formulas in PyTorch, can be used as the new representation of the i-th word after performing dimensionality reduction by keeping the top k singular values?

Selected Answer: U[i,:k]

Answers: U[i,:k]
V[i,k:]
V[:k,i]
U[k:,i]

Question 13

10 out of 10 points

You are working on a sentiment analysis project using the Naive Bayes method to classify movie reviews as either "Positive" or "Negative." You have collected the following data for a set of reviews:

Word	Positive Class Count	Negative Class Count
Great	7	1
Terrible	1	6
Acting	5	4
Plot	3	5
Absolutely	5	4

Given that we have overall 20 reviews (10 positives and 10 negatives). What is the probability that a review be **positive** if it has the word "**Terrible**"?

Selected Answer: 1/7

Answers: 6/10
6/7
7/10
1/7

Question 14

10 out of 10 points

In Natural Language Processing, what does the bigram probability P(A|B) represent?

Selected Answer: The probability of A given B has occurred

Answers: The probability of B followed by A
The probability of A and B occurring together
The probability of A given B has occurred
The probability of B given A has occurred

Question 15

6.66 out of 10 points

In the context of language models, what does "perplexity" measure?

Selected Answers: The amount of surprise
The accuracy of the model

Answers: The amount of surprise
The percentage of correctly predicted words
The length of the sentence
The accuracy of the model

Question 16

0 out of 10 points

Sub-word tokenization is a tokenization method used in NLP to split text into smaller units, often sub-word or character-level tokens instead of full

Sub-word tokenization is a tokenization method used in NLP to split text into smaller units, often sub-word or character-level tokens instead of full words. This approach is particularly helpful for handling languages with complex morphology or for working with languages that don't have clear spaces between words.

Assume you are given the following list of sub-word tokens:

Sub-word Tokens: self, less, good, bad, in, credi, ultra, ble, ity, ist, nation, al, ness

Now, how many sub-word tokens are in the sentence **"credibility"**?

Selected Answer: 2

Answers: 3
1
2
4

Question 17

10 out of 10 points

Consider the SVD decomposition for the square matrix A as $A = U_A \Sigma_A V_A^T$. We define a new matrix as $B = A^T A$. Considering its SVD decomposition as $B = U_B \Sigma_B V_B^T$, which answer is correct?

Selected Answer: $U_B = V_A$, $\Sigma_B = \Sigma_A^2$, $V_B = V_A$

Answers: $U_B = U_A$, $\Sigma_B = \Sigma_A^2$, $V_B = U_A$

$U_B = V_A^T$, $\Sigma_B = \Sigma_A$, $V_B = U_A^T$

$U_B = V_A$, $\Sigma_B = \Sigma_A^T$, $V_B = U_A$

$U_B = V_A$, $\Sigma_B = \Sigma_A^2$, $V_B = V_A$

Question 18

10 out of 10 points

Consider the common loss function used in SVM. If a data point is correctly classified and lies outside the margin, what is the loss for that data point?

Selected Answer: Zero.

Answers: Equal to the distance from the decision boundary.

Zero.

1.

Dependent on the number of support vectors.

Question 19

5 out of 10 points

Select all that is correct regarding the time complexity of decoding algorithms. (Consider decoding of a sentence with length n and k possible tags.

Selected Answers: Viterbi: $O(nk^2)$

Brute-force: $O(k^n)$

Answers: Greedy: $O(n+k)$

Viterbi: $O(nk^2)$

Brute-force: $O(k^n)$

Viterbi: $O(kn^2)$

Greedy: $O(n+k)$

Question 20

10 out of 10 points

When applying SVMs to text classification in NLP, a common preprocessing step is to:

Selected Answer: Represent text as feature vectors, e.g., using TF-IDF weighting.

Answers: Convert text into phonemes.

Embed text into a pre-trained neural network model.

Represent text as feature vectors, e.g., using TF-IDF weighting.

Assign syntactic roles to words in sentences.

Question 21 10 out of 10 points

Select all that are correct regarding the Generative and Discriminative models

Selected Answers:
- It is generally easier to design and train discriminative models
- Discriminative models are generally more effective for classification tasks

Answers:
- It is generally easier to design and train discriminative models
- Discriminative models are generally more effective for classification tasks
- Discriminative models can better distinguish outliers
- Generative models have an overfitting problem on a large dataset
- Generative models build a model of the posterior probability

Question 22 10 out of 10 points

	DT	NN	VB
DT	0	0.9	0.1
NN	0	0.5	0.5
VB	0.6	0.4	0

	The	Man	Saw	Rat
DT	1	0	0	0
NN	0	0.2	0.3	0.1
VB	0	0.2	0.5	0.3

Consider the POS tagging and the decoding task using HMM where the emission and transition probabilities are given in the tables. Based on these tables, calculate $p(x_3 = Saw \mid s_2 = NN)$?

Selected Answer: 0.4
Correct Answer: 0.4
Answer range +/- 0 (0.4 - 0.4)

Question 23 10 out of 10 points

By the universal approximation theorem, we can theoretically approximate almost any function with neural networks of sufficient width and depth.

Selected Answer: True
Answers:
- True
- False

Question 24 10 out of 10 points

In logistic regression for NLP tasks, regularization such as L1 or L2 is sometimes applied. What is the primary purpose of regularization?

Selected Answer: To prevent overfitting by adding a penalty to the magnitude of the coefficients.

Answers:
- To increase the model's training speed.
- To prevent overfitting by adding a penalty to the magnitude of the coefficients.
- To handle non-linear relationships in the data.
- To reduce the dimensionality of the text data.

Question 25 10 out of 10 points

In Natural Language Processing, when it comes to text classification, what is the role of Bayes' Rule in the Naive Bayes algorithm?

Selected Answer: Bayes' Rule is used to estimate the conditional probability of a document belonging to a particular class based on its feature vector.

Answers:
- Bayes' Rule is used to calculate the prior probability of a feature occurring in a class.
- Bayes' Rule is not used in the Naive Bayes algorithm.
- Bayes' Rule is used to estimate the likelihood of a feature occurring in a document for a given class.
- Bayes' Rule is used to estimate the conditional probability of a document belonging to a particular class based on its feature vector.

Question 26 10 out of 10 points

Select all that is correct when comparing Glove to Word2Vec.

Selected Answers:
- Similar to Word2Vec, Glove considers the occurrences of terms at the context level.
- In contrast to Glove, Word2Vec is a prediction-based model.

Answers:
- Similar to Word2Vec, Glove considers the occurrences of terms at the context level.
- In contrast to Glove, Word2Vec is a prediction-based model.
- In contrast to Word2Vec, Glove considers the occurrences of terms at the document level.

Similar to Glove, Word2Vec is a count-based model.

Question 27 10 out of 10 points

Which statement(s) is(are) true regarding overfitting in machine learning models?

Selected Answers: It occurs when a model captures noise in the training data as if it were a real pattern

Answers:
- It happens when a model performs poorly on the training data
- It occurs when a model captures noise in the training data as if it were a real pattern
- It is always preferable for improving model accuracy
- It is not related to model complexity

Question 28 10 out of 10 points

Skip-thought vectors are created using encoder-decoder model that the encoder maps a sentence into a vector and decoder conditions on this vector to generate other sentences.

Selected Answer: True

Answers:
- True
- False

Question 29 0 out of 10 points

Latent semantic analysis considers the occurrence of terms at document level, while Word2Vec considers the occurrence of the terms at the context level.

Selected Answer: False

Answers:
- True
- False

Question 30 10 out of 10 points

There are two approaches to the POS tagging. Modeling the $p(s_1,s_2,\ldots,s_n|x_1,x_2,\ldots,x_n)$ distribution corresponds to the discriminative approach while the $p(x_1,x_2,\ldots,x_n|s_1,s_2,\ldots,s_n)p(s_1,s_2,\ldots,s_n)$ distribution corresponds to the generative approach.

Selected Answer: True

Answers:
- True
- False

Question 31 10 out of 10 points

Which of the following statements are true about Probabilistic Language Models (PLMs)?

Selected Answers:
- PLMs assign a probability distribution over a sequence of words in a language.
- PLMs can be utilized to generate new text that is similar to the text they were trained on.

Answers:
- PLMs assign a probability distribution over a sequence of words in a language.
- They cannot be used for tasks like speech recognition or machine translation.
- They always require huge amounts of data for training.
- PLMs can be utilized to generate new text that is similar to the text they were trained on.

Question 32 10 out of 10 points

One application of the TF-IDF method is in open-domain question-answering. Here we use TF-IDF to find a list of relevant documents to an input text query. You are given a collection of six documents and an input query. Let's assume the term counts (TC) for the term "apple" have been calculated for each document. For simplicity, assume each document has only 10 words. Here is the TC for the term "apple" in the documents:

- Document 1: TC("apple") = 5
- Document 2: TC("apple") = 3
- Document 3: TC("apple") = 0 (Does not mention "apple")
- Document 4: TC("apple") = 0 (Does not mention "apple")
- Document 5: TC("apple") = 0 (Does not mention "apple")
- Document 6: TC("apple") = 2

Assume the input query contains the word “apple”, now, which is the TF-IDF score for the term "apple" in the **second** document.

Selected Answer: 3/10 log(2)

Answers:
- 3/10 log(1/2)
- 10/3 log(2)
- 10/3 log(1/2)
- 3/10 log(2)

Question 33 6.66666 out of 10 points

In NLP, which of the following statements are not the purposes for using the Viterbi algorithm in conjunction with HMMs?

Selected Answers:
- To calculate the most likely sequence of states given a sequence of observations.
- To assign POS tags to words in a supervised learning context.
- To maximize the likelihood of the HMM parameters in an unsupervised manner.

Answers:
- To calculate the most likely sequence of states given a sequence of observations.
- To assign POS tags to words in a supervised learning context.
- To maximize the likelihood of the HMM parameters in an unsupervised manner.
- To represent words as dense vectors in a high-dimensional space.

Question 34 10 out of 10 points

Logistic Regression is a generative approach to classification.

Selected Answer: False

Answers:
- True
- False

Question 35 5 out of 10 points

Which of the following formulas is correct in HMMs? (x_i denotes the word at index i and s_i is its tag)

Selected Answers:

$$p(x_1,x_2,\ldots,x_n,s_1,s_2,\ldots,s_n) = \prod_{i=1}^{n} p(x_i|s_i)p(s_1)\prod_{i=2}^{n} p(s_i|s_{i-1})$$

Answers:

$$p(x_1,x_2,\ldots,x_n,s_1,s_2,\ldots,s_n) = \prod_{i=1}^{n} p(x_i|s_i)p(s_1)\prod_{i=2}^{n} p(s_i|s_{i-1})$$

$$\log p(x_1,x_2,\ldots,x_n|s_1,s_2,\ldots,s_n) = \prod_{i=1}^{n} \log p(x_i|s_i)$$

$$p(s_1,s_2,\ldots,s_n) = p(s_1) + \sum_{i=2}^{n} p(s_i|s_{i-1})$$

$$p(x_2|s_1) = \sum_{s_2} p(x_2 \mid s_2)p(s_2 \mid s_1)$$

Question 36 10 out of 10 points

For a sentence W consisting of words W_1, W_2, ..., W_n, which of the following represents the chain rule of probability?

Selected Answer: P(W) = P(W_1) * P(W_2|W_1) * P(W_3|W_1,W_2) * ... * P(W_n|W_1,...,W_{n-1})

Answers:
- P(W) = P(W_1) * P(W_2) * ... * P(W_n)
- P(W) = P(W_1) * P(W_2|W_1) * P(W_3|W_1,W_2) * ... * P(W_n|W_1,...,W_{n-1})
- P(W) = P(W_1|W_2) * P(W_2|W_3) * ... * P(W_{n-1}|W_n)
- P(W) = P(W_1) * P(W_2) * ...P(W_n) * P(W_2|W_1) * (W_3|W_2) * ... * P(W_n|W_{n-1})

Question 37 10 out of 10 points

Select all that are **incorrect** about "natural language":

Selected Answers:
- Braille (a writing system used by people who are visually impaired) is a natural language
- Emoji are a natural language
- Body language is a natural language

Answers:
- Braille (a writing system used by people who are visually impaired) is a natural language
- Emoji are a natural language
- Body language is a natural language
- Upland Yuman (Native American language spoken in the southwestern United States) is a natural language
- Chinese is a natural language

Question 38 0 out of 10 points

You are tasked to write an application that gives the definition of a word in a given sentence in English (e.g. "the store is closed" → "store (n): a retail establishment selling items to the public."). Select all the levels of linguistic structures you think you'd need in your application.

Selected Answers:
- Morphology
- Lexical Semantics
- Compositional Semantics

Answers:
- Phonetics
- Morphology
- Syntax
- Lexical Semantics
- Compositional Semantics
- Production

Question 39 0 out of 10 points

Using the bigram probabilities below:

cat sleeps 0.40

cat caught 0.35

caught a 0.20

caught the 0.40

sleeps a 0.01

sleeps the 0.50

Which of the following word pairs is most likely to follow "The cat" based on the bigram

Selected Answer: caught the

Answers:
- caught a
- sleeps the
- caught the
- sleeps a

Question 40 10 out of 10 points

You have a vocabulary size of V=10,000 words. The word "apple" appears 3 times in a corpus. Using add-one smoothing, what is the smoothed probability of "apple"? Assume the total count of words in the corpus is N=40,000.

Selected Answer: 0.00008

Correct Answer: 0.00008

Answer range +/- 0.00001 (0.000070 - 0.000090)

Question 41 10 out of 10 points

Which of the following statements are true about Hidden Markov Models (HMMs) when used for POS tagging?

Selected Answers:
- HMMs model the joint probability of a word sequence and its corresponding POS tag sequence.
- HMMs require manually labeled data to train transition and emission probabilities.
- In HMM-based POS tagging, states correspond to POS tags, and observations correspond to words in sentences.

Answers:
- HMMs model the joint probability of a word sequence and its corresponding POS tag sequence.
- HMMs require manually labeled data to train transition and emission probabilities.
- HMMs can only be applied to languages with fixed word orders.
- In HMM-based POS tagging, states correspond to POS tags, and observations correspond to words in sentences.

Question 42 0 out of 10 points

Select all the options that **are not usually part** of the tokenization step:

Selected Answers:
- Removing stop words
- Removing external URL links
- Removing repeated words
- Removing high-frequency words such as “on”, “off” that do not have specific semantical information

Answers:
- Removing stop words
- Removing external URL links
- Removing repeated words
- Removing punctuation
- Removing extra spaces
- Removing high-frequency words such as “on”, “off” that do not have specific semantical information

Question 43 5 out of 10 points

Select all that are **correct** about **Rule-Based NLP Models:**

Selected Answers:
- rely on predefined linguistic and grammatical rules.

They have high performance in specific use cases
They can keep up with changes in the language (e.g. addition of new words)
requires domain knowledge of the language

Answers: rely on predefined linguistic and grammatical rules.
They have high performance in specific use cases
They can keep up with changes in the language (e.g. addition of new words)
Require large amounts of labeled training data.
requires domain knowledge of the language

Question 44 6.66666 out of 10 points

Which of the following are challenges or limitations faced by HMMs in the context of POS tagging?

Selected Answers: Handling unseen words that were not present in the training data.
Capturing long-distance dependencies between words.

Answers: Handling unseen words that were not present in the training data.
Capturing long-distance dependencies between words.
The need for a substantial amount of annotated training data.
Being computationally efficient due to the Viterbi algorithm.

Question 45 0 out of 10 points

46: Which of the followings are true about training a simple feedforward neural network?

Selected Answers: The reason we use stochastic gradient descent rather than original gradient descent is that stochastic gradient descent is more computationally efficient and requires less memory.
The training dataset can be reused as validation or test dataset after the training is completed.

Answers: normally, we perform gradient descent one time per batch.
The reason we use stochastic gradient descent rather than original gradient descent is that stochastic gradient descent is more computationally efficient and requires less memory.
The training procedure will always lead to convergence as long as we train the model with enough epochs.
The training dataset can be reused as validation or test dataset after the training is completed.

Question 46 10 out of 10 points

Which of the following statements are correct regarding N-gram models in Natural Language Processing?

Selected Answers: They can capture the probability of each word given its previous words in a sentence.
They cannot model the syntactic structure of a sentence.
They assume that the next word only depends on the previous two words for trigram

Answers: They can capture the probability of each word given its previous words in a sentence.
They cannot model the syntactic structure of a sentence.
They assume that the next word only depends on the previous two words for trigram
They always provide accurate predictions for next-word prediction tasks.

Question 47 10 out of 10 points

Imagine a 2-layer perceptron in the default setting, the first layer has 2 nodes and the second layer has 2 nodes as well. The weight of the first layer's nodes is [2,1], with the bias of -2, the weight of the second layer's node is [2,1], with the bias of -1. The activation function of ReLU is applied for both layers, what is the final output given the input as [1,2]?

Selected Answer: 7,3
Correct Answer: 5
Answer range +/- 0 (5 - 5)

Question 48 10 out of 10 points

45: Which of the following statements are true about RNN and FFN?

Selected Answers: Training an RNN is more likely to encounter vanishing/exploding gradient issues.
FFN requires fixed input size, while RNN does not.

Answers: RNN updates the weight based on the result of the last output node.
FFN is more computationally expensive when trained on data that has a long-term dependency.
Training an RNN is more likely to encounter vanishing/exploding gradient issues.
FFN requires fixed input size, while RNN does not.

Question 49 10 out of 10 points

In the context of SVMs for NLP tasks, a high value of the regularization parameter C will:

Selected Answer: Put more emphasis on correct classification over margin maximization.
Result in a wider margin and possibly more margin violations

Answers: Result in a wider margin and possibly more margin violations.

Put more emphasis on correct classification over margin maximization.

Prioritize maximizing the margin over correct classification.

Reduce the influence of support vectors on the decision boundary.

Question 50 10 out of 10 points

Given a PyTorch dataset with N=10,000 samples, if you use a batch size of 200 and train for 5 epochs, how many forward-backward passes (iterations) will the model undergo?

Selected Answer: 250

Correct Answer: 250

Answer range +/- 1 (249 - 251)

Question 51 10 out of 10 points

What are the training samples of Skip-Gram given a sentence "natural language processing is pretty useful", and the word "processing" is the center word?

Selected Answer: (processing, natural), (processing, language), (processing, is), (processing, pretty), (processing, useful)

Answers: (processing, is), (processing, pretty), (processing, useful)

(processing, natural), (processing, language)

(processing, natural), (processing, language), (processing, is), (processing, pretty), (processing, useful)

(processing, natural), (processing, language), (processing, processing), (processing, is), (processing, pretty), (processing, useful)

Question 52 10 out of 10 points

When using logistic regression for text classification, why is the sigmoid function commonly used as the activation function?

Selected Answer: It maps any input into a value between 0 and 1, making it suitable for binary classification as probability values.

Answers: It maps any input into a value between 0 and 1, making it suitable for binary classification as probability values.

It helps in converting non-linearly separable data to a higher dimension where it might be linearly separable.

It clusters similar words together in vector space.

It enhances the semantic meaning of words in a sentence.

Question 53 0 out of 10 points

Consider the SVD decomposition of matrix $A=\begin{bmatrix}0&1&1\\1&2&1\\1&1&0\end{bmatrix}$ as $U=\begin{bmatrix}\frac{\sqrt{6}}{6}&\frac{-\sqrt{2}}{2}&\frac{\sqrt{3}}{3}\\\frac{\sqrt{6}}{3}&0&\frac{-\sqrt{3}}{3}\\\frac{\sqrt{6}}{6}&\frac{\sqrt{2}}{2}&\frac{\sqrt{3}}{3}\end{bmatrix}$, $\Sigma=\begin{bmatrix}3&0&0\\0&\sigma_2&0\\0&0&\sigma_3\end{bmatrix}$ and $V=\begin{bmatrix}\frac{\sqrt{6}}{6}&\frac{\sqrt{2}}{2}&\frac{\sqrt{3}}{3}\\\frac{\sqrt{6}}{3}&0&\frac{-\sqrt{3}}{3}\\\frac{\sqrt{6}}{6}&\frac{-\sqrt{2}}{2}&\frac{\sqrt{3}}{3}\end{bmatrix}$. What is the value of σ_2?

Selected Answer: 0

Correct Answer: 1

Answer range +/- 0 (1 - 1)

Question 54 3.33333 out of 10 points

In SVM, the decision boundary may not be determined by:

Selected Answers: A random subset of data points to speed up computation.

Answers: All the data points in the dataset that are very far from the boundary, called support vectors.

Data points that are closest to the decision boundary, called support vectors

A random subset of data points to speed up computation.

Only the misclassified data points in each iteration.

Question 55

10 out of 10 points

For an RNN, how many dependencies in terms of time-steps are there if the weight W is computed through loss L_T of timestep T?

Selected Answer: T

Answers: T

T-1

T+1

1

Question 56

10 out of 10 points

In the context of SVMs, what can change if the value for the C parameter (or regularization parameter) changes:

Selected Answer: The width of the margin around the separating hyperplane.

Answers: The width of the margin around the separating hyperplane.

The number of support vectors used in the model.

The degree of the polynomial that models the separating boundary.

The transformation of data into higher-dimensional space.

Question 57

6.67 out of 10 points

48: which of the following are the design parameters of the Feedforward Neural Network?

Selected Answers: layer number

value in each node of the layers

activation functions

number of nodes in each layer

Answers: layer number

value in each node of the layers

activation functions

number of nodes in each layer

Question 58

0 out of 10 points

Select all that are **correct** about **deep learning:**

Selected Answers: Requires substantial computational resource

Requires substantial domain knowledge

Requires substantial pre-processing

Answers: Requires substantial computational resource

Requires substantial domain knowledge

Requires substantial pre-processing

Is a type of NLP model relying on statistical learning

Requires substantial feature engineering

More generalizable than rule-based NLP

Question 59

10 out of 10 points

Consider a log-linear model as $p(y|x,v) = \frac{e^{v.f(x,y)}}{\sum_{y'} e^{v.f(x,y')}}$. To maximize this probability for a given dataset, we have to use gradient descent over parameter v. What is the gradient of $\log p(y|x,v)$ w.r.t the v?

Selected Answer: $f(x,y) - \sum_{y'} f(x,y')p(y'|x,v)$

Answers: $f(x,y) - \sum_{y'} f(x,y')p(y'|x,v)$

$e^{f(x,y)} - \sum_{y'} e^{f(x,y')}$

$\frac{f(x,y)\sum_{y} e^{v.f(x,y')} - e^{v.f(x,y)}\sum_{y} f(x,y')e^{v.f(x,y')}}{(\sum_{y'} e^{v.f(x,y')})^2}$

$\frac{f(x,y)}{\sum_{y'} f(x,y')}$

Question 60

10 out of 10 points

For multi-class classification task which has only 1 true label for each instance, what is the best function to be applied for the last layer?

Selected Answer: Softmax

Answers:

Sigmoid

Softmax

ReLU

Tanh

Saturday, January 20, 2024 11:43:35 PM PST

← OK