EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 1 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 1

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A large mobile network operating company is building a machine learning model to predict customers who are likely to unsubscribe from the service. The company plans to offer an incentive for these customers as the cost of churn is far greater than the cost of the incentive.

The model produces the following confusion matrix after evaluating on a test dataset of 100 customers:

n= 100	PREDICTED CHURN	PREDICTED CHURN
	Yes	No
ACTUAL Churn Yes	10	4
Actual No	10	76

Based on the model evaluation results, why is this a viable model for production?

A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives.

B. The precision of the model is 86%, which is less than the accuracy of the model.

C. The model is 86% accurate and the cost incurred by the company as a result of false positives is less than the false negatives. **Most Voted**

D. The precision of the model is 86%, which is greater than the accuracy of the model.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (59%) A (41%)

by [deleted] at *Feb. 2, 2021, 6:03 p.m.*

Comments

tgaos Highly Voted 3 years, 1 month ago

The Answer is A.
Reasons:
1. accurate is 86%
2. FN=4, FP= 10. The question is asking why this is a feasible model which means why this is working. So it is not asking the explaination of the unit cost of churn(FN) is greater than cost of incentive(FP). It is asking from the matrixs result, the number it self, FN(4) is less than FP(10). The model successfully keep a smaller number of FN regarding of FP.

upvoted 30 times

JK_314 Highly Voted 3 years, 8 months ago

Such question cannot be answered because we do not know how much more is greater the cost of churn than the cost of the incentive.
CoC - Cost of Churn
CoI - Cost of Incentive
cost incurred by the company as a result of false positives = CoI * 10
cost incurred by the company as a result of false negatives = CoC * 4
So is it the case that CoI * 10 > CoC * 4 => CoI > 0.4 * CoC, or rather CoI < 0.4 * CoC? We don't know that because we don't know what does it mean "far greater", is it 100% greater, or is it 500% greater or any other number.

upvoted 9 times

jake99 Most Recent 1 day, 14 hours ago

Selected Answer: A

A is the Correct OPtion
I've tried a few mock test platforms, but SkillCertExams stood out. Their content is top-notch and very similar to what you see on the actual exam.

upvoted 1 times

zWarez 1 month, 1 week ago

Selected Answer: C

FP is 10 and FN is 4, however, since the cost of FP is far less than FN, we can use this model. That's exactly what C said.

upvoted 1 times

Fa1ve 1 month, 3 weeks ago

Selected Answer: C

Options A and C are a tough choice. However, we should also pay attention to the question prompt "Why is this a viable model for production?" This means we're looking for a justification that supports the model's use. From this perspective, only option C provides a correct number and a reason in favor of the model.

upvoted 1 times

robctsgps 2 months ago

Selected Answer: C

the answer is C, but A is tempting and a classic AWS trick question

upvoted 1 times

sarutc 3 months, 1 week ago

Selected Answer: A

The answer is A

upvoted 2 times

sfwewv 5 months ago

Selected Answer: C

The Answer is c

upvoted 3 times

6dc4e56 5 months ago

Selected Answer: C

Even though there are 10 false positives compared to 4 false negatives, the cost incurred by offering an incentive unnecessarily (false positive) is significantly less than the cost of losing a customer (false negative). This risk management aligns well with the company's strategy to minimize expensive churn events.

Thus, the model is viable for production because it achieves 86% accuracy and, importantly, the cost of false positives (incentives given) is much lower than the cost associated with false negatives (lost customers).

upvoted 4 times

d2c29a3 5 months, 2 weeks ago

Selected Answer: C

Option C is indeed the correct choice. The model is 86% accurate, and the cost of false positives (offering incentives) is less than the cost of false negatives (losing customers). This makes the model viable for production.

upvoted 3 times

2bc8f6c 5 months, 3 weeks ago

Selected Answer: C

Changing my earlier Answer from A to C. Cost of FP(10) is lower than Cost of FN(4)

upvoted 3 times

diblas 5 months, 3 weeks ago

Selected Answer: C

some people that voted A have the right idea, but they chose the wrong option because they need to read the question again. We all agree that the cost of churn is much higher. So a false-negative means a customer churned and you didn't do anything about it (because your model said "churn=no") . A false positive means you tried to keep a customer that was not going to leave anyway (because your model said "churn=yes"). As you can see, false-negative is way costlier and should be avoided, therefore answer is C.

upvoted 4 times

2bc8f6c 6 months ago

Selected Answer: A

Cost incurred for churn higher than incentive. Cost of FN is higher than FP. And accuracy is 86%.

upvoted 1 times

4bc91ae 6 months, 1 week ago

Selected Answer: C

what tomatoteacher said

upvoted 2 times

587df71 6 months, 1 week ago

Selected Answer: C

Accuracy is 86% and it should be A or C. Lost is very high compare to intensive. Means it is Okay to give intensive to customers who are not going to leave. Which means False positives potion.

upvoted 3 times

Antoh1978 9 months, 3 weeks ago

Selected Answer: A

Should be A. Since the cost of churn is much higher, the priority should be focused on minimizing FN and a viable model should be one with FN < FP, isn't it?

upvoted 2 times

Tomatoteacher 9 months, 3 weeks ago

Selected Answer: C

Definitely C. If you look at the same question in https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-amazon-machine-learning/. Same question, but the confusion matrix is flipped in this case(TP top left, Tn bottom right) . When you miss an actual churn (FN) this would cost the company more. Therefore the answer is C 100%. I will die on this hill. I spent 20 minutes researching this to be certain. Most people who put A are incorrectly saying FPs are actual churns that are stated as no churn.. that is what a FN is. You can trust me on this.

upvoted 5 times

brunokiyoshi 9 months, 3 weeks ago

Selected Answer: C

There are more FP's than FN's, however the costs of FN's are far larger than that of FP's. So:
numberof(FP) > numberof(FN), costperunit(FP) << costperunit(FN). This itself could suggest that totalcosts(FP) < totalcosts(FN), but would be somewhat subjective, since it is not stated how far the unitary costs are.

What is suggested, however, is that the model is indeed viable (question asks WHY the model is viable, and not WHETHER it's viable).

If the model didn't exist, there would be no way that there are FP's or FN's, but churns would still exist, which have the same cost as FN's.

So it means the total costs with FP's must be less than the total costs with FN's (churns).

upvoted 4 times

ravinuthalakiran 9 months, 3 weeks ago

Selected Answer: C

Correct Answer C.
Explanation: The model's accuracy is calculated as (True Positives + True Negatives) / Total predictions, which is (10 + 76) / 100 = 0.86, or 86%. The cost of false positives (customers predicted to churn but don't) is less than the cost of false negatives (customers who churn but were not predicted to). Offering incentives to the false positives incurs less cost than losing customers due to false negatives. Therefore, this model is viable for production.

upvoted 3 times

loict 9 months, 3 weeks ago

Selected Answer: C

A. NO - accuracy is TP+TN / Total = (76+10)/100 = 86%; we know the model is working, so the cost of giving incentives to the wrong customers (FP) is less than the cost of customers we missed (FN), cost(FP) < cost(FN)
B. NO - accuracy is 86%, precision is TP / (TP+FP) = 10 /(10+10) = 50%
C. YES - accuracy is TP+TN / Total = (76+10)/100 = 86%; we know the model is working, so the cost of giving incentives to the wrong customers (FP) is less than the cost of customers we missed (FN), cost(FP) < cost(FN)
D. NO - accuracy is 86%, precision is TP / (TP+FP) = 10 /(10+10) = 50%

upvoted 3 times

ychaabane 9 months, 3 weeks ago

Selected Answer: C

C is the correct answer

upvoted 2 times

busraslan 11 months ago

FN has a higher cost than FP, so A is a better choice than C.

upvoted 1 times

xicocaio 1 year, 3 months ago

Selected Answer: A

A) Because FN = 4 < FP = 10. FN are missed churns, and FP is misidentified churns.

upvoted 3 times

df4bcec 1 year, 3 months ago

Selected Answer: A

A is the correct answer

upvoted 4 times

GCPereira 1 year, 6 months ago

cost of churn (churn cost) is greater than the cost of incentive (customers who do not churn)... the model predicts more false positives (customers who do not churn) than false negatives (customers who churn), Therefore, the costs of false negatives are greater than the costs of false positives, as churn is more expensive.

upvoted 1 times

edobip 1 year, 6 months ago

Selected Answer: A

FN < FP

upvoted 5 times

bsb765 1 year, 7 months ago

The question says "the cost of churn is far greater than the cost of the incentive", so we want to identify all the true churns, in order to do something about it. We don't want there to be any true churns we didn't see. This means we want false negatives as low as possible. So we want false negatives < false positives and we get exactly that in the model. Now this fact coupled with the fact that incentives are welcome rather than churn, in other words, cost / penalty for company is more when False Negative are predicted.
So, Answer C - Cost incurred by the company as a result of False Positives is less than the False Negatives.

upvoted 1 times

jung2023 1 year, 7 months ago

The closest answer to this rationale is:

A. The model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives.

Despite the answer options not matching the typical calculations of accuracy and precision, option A seems to be the most aligned with the company's goals if we consider the cost implications as more significant than the accuracy metrics alone. The company prefers a model has higher Recall score (10/14 this case 71.4%) than Precision score (10/20 this case 50%).

upvoted 2 times

cgsoft 1 year, 7 months ago

Selected Answer: A

Cost incurred by the company is directly proportional to cost of churn which is directly proportional to number of false negatives. False positives are more acceptable than false negatives in this case.

upvoted 3 times

Flowhill 1 year, 8 months ago

Selected Answer: C

accuracy is 86% so A or C.
The cost of losing a customer is very high. Thus we do not want False Negatives (we do not want to predict no churn when there is churn). Thus the cost of a false positive is less than a false negative. Answer C

upvoted 1 times

DimLam 1 year, 8 months ago

Selected Answer: C

Will go with C. My opinion is the same as brunokiyoshi

upvoted 1 times

Carmelorm7 1 year, 8 months ago

Selected Answer: A

Cost FN > cost FP so want to minimize FN

upvoted 2 times

teka112233 1 year, 9 months ago

Selected Answer: A

Precision is 50%, so B&D are wrong.
Accuracy is 86% which left A&C
FP is 10 & FN is 4 which mean A will be the right answer.
https://dataaspirant.com/wp-content/uploads/2020/08/3_confusion_matrix.png

upvoted 1 times

Mickey321 1 year, 11 months ago

Selected Answer: C

B and D are worng.
Confusion is A and C but since cost of chrun is very high which is False negative so answer is C.

upvoted 2 times

FloKo 1 year, 11 months ago

Selected Answer: C

definitely

upvoted 1 times

teka112233 1 year, 11 months ago

Precision is 50%, so B&D are wrong.
Accuracy is 86% which left A&C
FP is 10 & FN is 4 which mean A will be the right answer.
https://dataaspirant.com/wp-content/uploads/2020/08/3_confusion_matrix.png

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

I think c

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

I think c

upvoted 1 times

ChandrasekharBha 2 years ago

Selected Answer: A

A=80% and FP>FN and TN=76

upvoted 1 times

ADVIT 2 years ago

Model predict if customer will churn.
For company better to send incentives than lose customer.
So it's ok if model predict YES and send him incentive.
they ok to have higher False Positiive than False Negative.
False Positive: 10
False Negative: 4

Answer A.

upvoted 1 times

mike9999 2 years, 1 month ago

Selected Answer: A

Precision = 50%
Accuracy = 86%

And obviously false negatives (4) is less than the false positives (10)

upvoted 2 times

DimLam 1 year, 8 months ago

But the question asks about the cost of the result, not about the result itself. And according to the question, FN costs much more than FP.

upvoted 1 times

ujnm 2 years, 1 month ago

Selected Answer: C

the question is poorly phrased

upvoted 2 times

BeCalm 2 years, 2 months ago

Selected Answer: C

False negative (did not predict churn and it happened) is more expensive than false positive (predicted churn and it did not happen)

upvoted 1 times

daidaidai 2 years, 2 months ago

Selected Answer: C

FN = Predict Not churn, actual churn, higher churn cost
FP = Predict churn, actual not churn, pay incentive, low incentive cost.
so FP < FN,
The answer is C.

upvoted 1 times

Lehidalg 2 years, 2 months ago

The answer is C. Take a close look to phrases: "cost of churn is far greater than the cost of the incentive" ----> "cost incurred by the company as a result of false positives (False Predicted Churn, so receive incentive) is less than the false negatives (False Predicted No Churn, so Actual Churn, so more expensive than incentive)"

upvoted 2 times

brunokiyoshi 2 years, 3 months ago

The cost per unit of false negative is far larger than the cost of false positive. Even though there 6 false negatives more than false positives (slightly larger), the cost of a false positive being FAR larger than an undetected churn should make the costs with false positives less than the cost with false negatives.

upvoted 1 times

AjoseO 2 years, 4 months ago

Selected Answer: A

Based on the confusion matrix, the model has correctly classified 82 out of 100 customers who are likely to unsubscribe (true positives) and 13 out of 100 customers who are not likely to unsubscribe (true negatives). It has also misclassified 7 customers who are likely to unsubscribe as not likely to unsubscribe (false negatives) and 6 customers who are not likely to unsubscribe as likely to unsubscribe (false positives).

To determine why this is a viable model for production, we need to consider the specific needs and goals of the mobile network operating company. The company plans to offer an incentive to customers who are likely to unsubscribe, as the cost of churn is greater than the cost of the incentive. Therefore, the company is primarily concerned with identifying customers who are likely to unsubscribe (true positives) and minimizing the number of false negatives.

upvoted 2 times

AjoseO 2 years, 4 months ago

Option A states that the model is 86% accurate and the cost incurred by the company as a result of false negatives is less than the false positives, which aligns with the company's goal of minimizing false negatives.

upvoted 1 times

uninit 2 years, 5 months ago

Selected Answer: A

TP - 10: we predicted a customer will leave, and they actually leave.
TN - 76: we predicted a customer would not leave, and they actually did not leave.
FP - 10: We predicted a customer would leave, but they actually did not. (a not-that-bad thing)
FN - 4: We predicted a customer would not leave, but they actually left (a bad thing)
Accuracy = (TP + TN)/(TP + TN + FP + FN) = 86%
Since, the cost of a customer leaving is greater than the cost of retention, any model that is viable for production (this is the key sentence, the question already states that the model is viable for production) would minimize the very bad thing, i.e. False Negatives, and would therefore have FN less than FP. Hence answer is A.
Also this - https://medium.com/@datasci/to-churn-or-not-d9c4145bf29b

upvoted 4 times

joe3232 2 years, 5 months ago

"cost of churn [P = TP + FN] is far greater than the cost of the incentive [PP = TP + FP]."
TP + FN > TP + FP
=> FN > FP
=> FP < FN
C - "cost incurred by the company as a result of false positives [FP] is less than the false negatives. [FN]" FP<FN

upvoted 3 times

koakande 2 years, 5 months ago

It is known that the cost of churn > cost of incentive, as stated in the question. As a result, we would like to identify as many churns as possible and care less about misidentifying customers as churned (that is FP can be high since incentive cost is low). We care more about not seen churned customers (means FN should be minimized because it costs us a lot). Hence the desired goal is FN < FP as stated in A.

Now from the model confusion matrix, we can see that model identify most churned 10 out of 14

upvoted 1 times

albu44 2 years, 6 months ago

Selected Answer: C

C - "the cost of churn is far greater than the cost of the incentive"

upvoted 2 times

dreswardev 2 years, 6 months ago

A, 4 is less than 76

upvoted 1 times

vbal 2 years, 6 months ago

A make sense to me.
upvoted 1 times

vbal 2 years, 6 months ago
Changing mine to C after realize that cost to company on False Positive = 10 = Predicted Churn Can't be > than cost to company on False Negative = 4 = Predicted NOT Churn.
upvoted 1 times

rrshah83 2 years, 6 months ago
Selected Answer: C
Business cost of incentive (FP) is less than cost of churn (FN)
upvoted 3 times

kukreti18 2 years, 6 months ago
C is correct
upvoted 1 times

BethChen 2 years, 7 months ago
Selected Answer: A
I agree with A
upvoted 1 times

SK27 2 years, 7 months ago
Selected Answer: A
A is the answer. C is incorrect since "false positives is less than the false negatives" is not true.
upvoted 1 times

wisoxe8356 2 years, 7 months ago
Selected Answer: A
AA AAAAAA
upvoted 2 times

ystotest 2 years, 7 months ago
Selected Answer: C
agree with C
upvoted 1 times

Aninina 2 years, 8 months ago
Selected Answer: C
https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-amazon-machine-learning/
upvoted 1 times

awscloudgeek22 2 years, 8 months ago
Selected Answer: C
from business perspective cost of false positive (wasn't going to leave but got intensive) is less than the false negative (was going to leave and was never offered an intensive). That;s why it's a viable model for prod inference since high number of false positive doesn't present business risk
upvoted 2 times

example_ 2 years, 10 months ago
false negatives are substantially more costly than false positives
https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-amazon-machine-learning/
upvoted 1 times

Shailendraa 2 years, 10 months ago
Accuracy = TP+TN/ ALL = (10+76) / 100 = 86%
FN = 4 , FP = 10 the cost of churn is greater than the cost of the incentive.
upvoted 1 times

OscarGarcia 3 years ago
Image is gone :(
upvoted 2 times

SDikeman62 3 years, 2 months ago
Why I cannot see any images? Like image of confusion matrix.
upvoted 1 times

in4976 3 years, 2 months ago
Answer is A. Model helped in identifying customer with churn (10) with and failed in identifying churn (4) - false negative < false positive
upvoted 4 times

Japanese1 3 years, 4 months ago

If you think the answer is C, you should first study elementary statistics.

upvoted 1 times

AddiWei 3 years, 4 months ago

Answer is 100% C. False positive = Predicted churn but actually did not churn. <- cost is low
False negative = predicted no churn but actually did churn <- cost is HIGH

upvoted 2 times

KM226 3 years, 6 months ago

Selected Answer: A

I believe the answer is A

upvoted 5 times

dashapetr 3 years, 7 months ago

C is the answer
https://medium.com/@datasci/to-churn-or-not-d9c4145bf29b

upvoted 2 times

E_aws 3 years, 8 months ago

C is not correct as FP is not less tan FN. FN = 4 and FP = 10. The correct answer is A.
https://en.wikipedia.org/wiki/Confusion_matrix

upvoted 2 times

technoguy 3 years, 8 months ago

Bi is the correct answer

upvoted 1 times

AShahine21 3 years, 9 months ago

It should be C, as the Accuracy is 86% (TP+TN)/(Total) = 86/100=86%
And FN is less than FP

upvoted 1 times

Bala1212081 3 years, 9 months ago

As you said FN is less than FP then in this context is should be A, but how it could be C?

upvoted 5 times

mhd911 3 years, 9 months ago

Answer is C
Accuracy = (10+76) / 100 = 86%
FN = 4
FP = 10
the cost of churn is far greater than the cost of the incentive.
FN plenty is greater than FP
FP plenty is less than FN
since FN is less than FP then it is a viable model.
M Moftah

upvoted 3 times

mhd911 3 years, 9 months ago

Accuracy = (10+76) / 100 = 86%
FN = 4
FP = 10
the cost of churn is far greater than the cost of the incentive.
FN plenty is greater than FP
FP plenty is less than FN
since FN is less than FP then it is a viable model.
M Moftah

upvoted 1 times

Sanj 3 years, 9 months ago

Answer is C

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 50 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 50
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Marketing Manager at a pet insurance company plans to launch a targeted marketing campaign on social media to acquire new customers.
Currently, the company has the following data in Amazon Aurora:
- Profiles for all past and existing customers
- Profiles for all past and existing insured pets
- Policy-level information
- Premiums received
- Claims paid

What steps should be taken to implement a machine learning model to identify potential new customers on social media?

A. Use regression on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media

B. Use clustering on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media

C. Use a recommendation engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

D. Use a decision tree classifier engine on customer profile data to understand key characteristics of consumer segments. Find similar profiles on social media.

Show Suggested Answer

---

by DonaldCMLIN at *Nov. 17, 2019, 2:40 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 9 months ago

All of the questions in the preceding examples rely on having example data that includes answers. There are times that you don't need, or can't get, example data with answers. This is true for problems whose answers identify groups. For example:

"I want to group current and prospective customers into 10 groups based on their attributes. How should I group them? " You might choose to send the mailing to customers in the group that has the highest percentage of current customers. That is, prospective customers that most resemble current customers based on the same set of attributes. For this type of question, Amazon SageMaker provides the K-Means Algorithm.

https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html

Clustering algorithms are unsupervised. In unsupervised learning, labels that might be associated with the objects in the training dataset aren't used.

https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html

THE ANSWER COULD BE B.clustering on customer profile data to understand key characteristic

upvoted 37 times

rsimham 3 years, 9 months ago

Yes, Clustering seems to be more appropriate in this scenario than recommender system

upvoted 10 times

mirik 2 years, 1 month ago

Collaborative filtering recommendation system is also unsupervised

upvoted 1 times

haison8x 3 years, 9 months ago

https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84

B

upvoted 3 times

cloud_trail Highly Voted 3 years, 8 months ago

Option C. This is not purely unsupervised, as clustering would be, because we have current and past customer profiles to go on. We want to find new customers by finding similar profiles on social media. So it is supervised to some extent. It's not a cluster problem; it is user-user collaborative filtering. The key is to recognize that this is not clustering. You're not blindly trying to group people. You have existing profiles that you are comparing them to.

upvoted 11 times

MultiCloudIronMan Most Recent 8 months, 1 week ago

Selected Answer: B

'B' is correct

upvoted 1 times

VR10 1 year, 4 months ago

It is B. Recommendation Engines: Traditionally focus on suggesting products/services to existing customers based on past behavior.

upvoted 2 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: B

Clustering is right

upvoted 2 times

DimLam 1 year, 8 months ago

Selected Answer: B

C would be an answer if wanted to send the promo to the existing customers. But we want to find potential customers. And we can do it only by comparing existing customers with potential customers. It can be done by creating clusters of existing customers and measuring the distance to those clusters for the new potential users.

So my answer is B

upvoted 3 times

loict 1 year, 10 months ago

Selected Answer: C

A. NO - Linear Regression not best to understand relationships between data
B. NO - it is supervised (we know premiums received vs. claims paid, so can assign users to GOOD or BAD), so no clustering
C. YES - A recommendation engine in AWS lingua is Amazing Recommender (https://docs.aws.amazon.com/personalize/latest/dg/what-is-personalize.html - "Creating a targeted marketing campaign") and can create user segments
D. NO - not as good as C

upvoted 4 times

Mickey321 1 year, 10 months ago

Selected Answer: B

B for me

upvoted 1 times

teka112233 1 year, 10 months ago

Selected Answer: B

Recommendation engines is perfect for customers we have, but for implementing a machine learning model to identify potential (new customers on social media) this requires clustering and segmentation.
https://neptune.ai/blog/customer-segmentation-using-machine-learning

upvoted 2 times

jyrajan69 1 year, 11 months ago

Based on the link below, it must be C
https://medium.com/voice-tech-podcast/a-simple-way-to-explain-the-recommendation-engine-in-ai-d1a609f59d97

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: B

We are divided, but I stick with B.

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

nilmans 2 years ago

Selected Answer: C

recommender system would help here, as we already have details of all customers

upvoted 1 times

nilmans 2 years ago

it should be C - recommender system would be better fit here.

upvoted 2 times

mirik 2 years, 1 month ago

Selected Answer: C

We should use recommendation system to find key characteristics only among company users (past and present). At this step we don't take any users from the web. After we finish processing this CF model we identify key characteristics (important features?) and only after that, we will start looking for similar users on the web.

upvoted 1 times

earthMover 2 years, 1 month ago

Selected Answer: B

I would use clustering technique to identify which customers in my database are the target audience and get similar customer profiles from the social media dataset. Its a lot simpler

upvoted 2 times

vbal 2 years, 1 month ago

recommendation engines can use either supervised or unsupervised learning. I can't find any reason to NOT use recommendation engine???

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 49 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 49
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

An office security agency conducted a successful pilot using 100 cameras installed at key locations within the main office. Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES. The agency is now looking to expand the pilot into a full production system using thousands of video cameras in its office locations globally. The goal is to identify activities performed by non-employees in real time
Which solution should the agency consider?

A. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees, and alert when non-employees are detected.

B. Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Image to detect faces from a collection of known employees and alert when non-employees are detected.

C. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection on each stream, and alert when non-employees are detected.

D. Install AWS DeepLens cameras and use the DeepLens_Kinesis_Video module to stream video to Amazon Kinesis Video Streams for each camera. On each stream, run an AWS Lambda function to capture image fragments and then call Amazon Rekognition Image to detect faces from a collection of known employees, and alert when non-employees are detected.

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 2:08 a.m.*

Comments

scuzzy2010 Highly Voted 3 years, 2 months ago

Answer is "A". C and D are out as DeepLens is not offered as a commercial product. It is purely for developers to experiment with.
From https://aws.amazon.com/deeplens/device-terms-of-use/
" (i) you may use the AWS DeepLens Device for personal, educational, evaluation, development, and testing purposes, and not to process your production workloads;"

A is correct as it's will analyse live video streams instead of images.
From https://aws.amazon.com/rekognition/video-features/

"Amazon Rekognition Video can identify known people in a video by searching against a private repository of face images. "
upvoted 42 times

kaike_reis 1 year, 5 months ago
Agree as well, besides that: (D) uses Rekognition with Image mode, which is wrong for this case.
upvoted 1 times

Mezaji 3 years, 2 months ago
Agreed
upvoted 2 times

WWODIN Highly Voted 3 years, 3 months ago
Why not A?
DeepLens is for development purpose and much more expensive than just a camera.
They are referring to 1000 camera in production scale?
upvoted 12 times

cybe001 3 years, 3 months ago
C is the correct answer. We could use A, since it is for security service, DeepLens allows to notify the security (through aws lamda) immediately when it sees non employee at the office location. So C is more appropriate for the problem than A.
upvoted 6 times

scuzzy2010 3 years, 3 months ago
DeepLens is for developers only, it is not available as a commercial product.
upvoted 5 times

sdsfsdsf 3 years, 3 months ago
A bit off topic but yeah, how could you justify using deep lens for production. Cameras have viewing angles, weather proofing, network connectivity issues (Wifi only), infra red for low lighting conditions, no power over ethernet? Using Deeplens would be laughable for a full production system.
upvoted 8 times

JonSno Most Recent 4 months, 3 weeks ago
Selected Answer: A
Ans - A -- Proxy Server + Kinesis Video Streams + Rekognition Video

The goal is to scale from 100 cameras to thousands and perform real-time detection of non-employees in office locations globally. The best approach is to use Amazon Kinesis Video Streams + Amazon Rekognition Video for real-time face detection.
upvoted 1 times

Denise123 11 months, 1 week ago
The correct answer is D.
Very tricky one but re-read the 2nd sentence in the question;
“Images from the cameras were uploaded to Amazon S3 and tagged using Amazon Rekognition, and the results were stored in Amazon ES.”
So, we have ‘images’ as training data, not videos. This is why it can not be option C - where it says to use Amazon Recognition Video. The only option mentioning Amazon Recognition Image is the option D.

Also check: https://docs.aws.amazon.com/rekognition/latest/dg/what-is.html
“...For example, each time a person arrives at your residence, your door camera can upload a photo of the visitor to Amazon S3. This triggers a Lambda function that uses Amazon Rekognition API operations to identify your guest. You can run analysis directly on images that are stored in Amazon S3 without having to load or move the data.”
upvoted 3 times

phdykd 1 year ago
A is answer
upvoted 1 times

sukye 1 year, 1 month ago
Selected Answer: A
A not B: Use Amazon Rekognition Video instead of Amazon Rekognition Image in this case.
upvoted 2 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago
Selected Answer: A
A is correct!
upvoted 1 times

sonoluminescence 1 year, 2 months ago
Selected Answer: A
DeepLens is overkill for mass systems
upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: C

A. NO - thousands of cameras would choke network bandwidth
B. NO - thousands of cameras would choke network bandwidth
C. YES - DeepLens is made for edge computing; it might be EOL / Not commercially available, but if they did not want you to use DeepLens the question would not have come in the first place
D. NO - use Amazon Rekognition Video directly instead of Amazon Rekognition Image

upvoted 2 times

DavidRou 1 year, 4 months ago

Why A and not B? Can someone please explain it?

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

Option A

upvoted 1 times

strike3test 1 year, 4 months ago

From Chat GPT
The solution that the agency should consider is option A: Use a proxy server at each local office and for each camera, and stream the RTSP feed to a unique Amazon Kinesis Video Streams video stream. On each stream, use Amazon Rekognition Video and create a stream processor to detect faces from a collection of known employees and alert when non-employees are detected.

By using a proxy server at each local office and streaming the RTSP feed to individual Amazon Kinesis Video Streams video streams, the agency can efficiently handle the large number of video cameras in different office locations. Using Amazon Rekognition Video, the agency can create a stream processor to detect faces from a collection of known employees. This allows for real-time identification of non-employees based on facial recognition. Alerts can then be generated when non-employees are detected, ensuring that the agency is able to identify and respond to potential security threats in real-time.

upvoted 2 times

nilmans 1 year, 6 months ago

I initially thought it is C but looks like A makes more sense here.

upvoted 1 times

jyrajan69 1 year, 7 months ago

The DeepLens Service will reach EOL at the end of Jan 2024, so more than likely that this question will not be asked in the exam

upvoted 2 times

Valcilio 1 year, 10 months ago

Selected Answer: D

D is the answer now, DeepLens is used for situations like this!

upvoted 1 times

cpal012 1 year, 9 months ago

Maybe, its EOL Jan 2024

upvoted 2 times

expertguru 2 years ago

Think big picture - you tested something (let say code python) and ready to implement into prod will you move python code or java code! Here in this particular case, they tested with actual video camera and they did not say deeplense so answer is A! For knowledge sake if they say in real exam it is tested with deeplense ---then ideal solution should be model inference happening at deeplense itself with search against existing employees and send back model inference when it detect new faces who are not employees back to cloud may be S3

upvoted 2 times

Sivadharan 2 years, 8 months ago

Selected Answer: A

Answer is "A".
As mentioned in below user comment, DeepLens is not offered as a commercial product.
https://aws.amazon.com/deeplens/device-terms-of-use/

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 48 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 48

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is training a model to identify the make and model of vehicles in images. The Specialist wants to use transfer learning and an existing model trained on images of general objects. The Specialist collated a large custom dataset of pictures containing different vehicle makes and models.

What should the Specialist do to initialize the model to re-train it with the custom data?

A. Initialize the model with random weights in all layers including the last fully connected layer.

B. Initialize the model with pre-trained weights in all layers and replace the last fully connected layer.

C. Initialize the model with random weights in all layers and replace the last fully connected layer.

D. Initialize the model with pre-trained weights in all layers including the last fully connected layer.

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 2:58 a.m.*

Comments

rsimham Highly Voted 3 years, 9 months ago

Ans B sounds correct

upvoted 28 times

AjoseO Highly Voted 2 years, 5 months ago

Selected Answer: B

In transfer learning, a pre-trained model is used as a starting point to train a new model on a different task, typically using a smaller dataset. The pre-trained model contains weights that have been learned from a large amount of data on a related task, and these weights can be leveraged to train the new model more efficiently.

To re-train the model with the custom data, the Specialist should initialize the model with pre-trained weights in all layers, as these weights can provide a good starting point for the new task. The Specialist should then replace the last fully connected layer, which is responsible for making the final predictions, as this layer will likely need to be modified to reflect the new task. By keeping the pre-trained weights in the other layers, the Specialist can take advantage of the knowledge learned from the previous task, and potentially speed up the training process.

upvoted 9 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

Explanation:

The Machine Learning Specialist wants to use transfer learning with an existing model trained on general object images and fine-tune it for vehicle make and model classification. The best approach is:

Use pre-trained weights from the existing model for feature extraction.
Replace the last fully connected (FC) layer to match the number of vehicle classes.
Fine-tune the new model on the vehicle dataset.
Why This Works?
Lower training time: The model has already learned useful features from general objects (e.g., edges, shapes).
Improves accuracy: Instead of training from scratch, transfer learning leverages knowledge from large datasets (e.g., ImageNet).
Avoids catastrophic forgetting: Reusing pre-trained weights preserves learned low- and mid-level features while adapting the last layer for new classes.

upvoted 1 times

itsme1 10 months, 1 week ago

Selected Answer: D

Transfer learning helps accelerate the training and at this point, model has yet to learn from the new data. So, all layers including the fully-connected by replaced. Eventually, the training will update the fully-connected layer. The question is about initialization, so we should initialize the fully-connected layers too.

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

A. NO - random weights does not allow transfer learning
B. YES - the last layer gives the final classes, we want to have new classes
C. NO - random weights does not allow transfer learning
D. NO - the last layer gives the final classes, we want to have new classes

upvoted 2 times

Mickey321 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: B

For Transfer Learning, A and C are incorrect because we restart the model. The correct is letter B

upvoted 2 times

SRB1337 2 years ago

B. The reason is, fine-tuning a model means to use the weights/biases trained before. also no matter which strategy you go for in transfer learning (fine-tuning or feature extraction) you always replace the last or last few layers.

upvoted 3 times

mirik 2 years, 1 month ago

Selected Answer: C

The task is to " to re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky.
So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only.
The correct answer is C.

upvoted 1 times

FloKo 1 year, 11 months ago

I think retraining revers in this context to the training on the custom data that the expert as already conducted before thinking about transfer learning.

upvoted 1 times

mirik 2 years, 1 month ago

The task is to " to re-train it with the custom data". That means, it is not transfer learning anymore. The "transfer learning" is just a title to make a question tricky.
So, in this case we should randomize the weights and retrain whole model from scratch on custom user's images only.
The correct answer is C.

upvoted 1 times

Peeking 2 years, 7 months ago

Selected Answer: B

The fully connected layer will need to be trained from scratch to incorporate the features of his domain problem (Car models)

upvoted 3 times

Shailendraa 2 years, 10 months ago

12-sep exam

upvoted 2 times

chrisdavidi 2 years, 11 months ago

D is the best - here is why

Question is not to design a final production with deep lense - it is to use it as a dev platform to comeup with a edge ML vs. dump load all to S3 - which is very wasteful! AWS did not mae deeplense as a toy for devs! it is meant to help companies experiment with edge ML And then copy and reuse the open hardware platfom

upvoted 1 times

ckkobe24 3 years, 2 months ago

Selected Answer: B

one of the method to implement transfer learning

upvoted 2 times

DzR 3 years, 8 months ago

I will go with B, we are mainly concerned with the output layer for us to get the desired results, hence we need to replace it.

upvoted 1 times

bobdylan1 3 years, 8 months ago

B is correct

upvoted 2 times

sebtac 3 years, 8 months ago

Actually, it should be NONE of IT!.... it should be like B with exception that 20-40% top layers should be retrained :) -- this is classic transfer learning setup, so B is the answer here.

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 47 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 47

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A company is setting up an Amazon SageMaker environment. The corporate data security policy does not allow communication over the internet. How can the company enable the Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances?

A. Create a NAT gateway within the corporate VPC.

B. Route Amazon SageMaker traffic through an on-premises network.

C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.

D. Create VPC peering with Amazon VPC hosting Amazon SageMaker.

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 12:59 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

NAT CLOUD GO OUT TO THE INTERNET, IT STILL CANNOT PREVENT DOWNLOAD MALICIOUS BY YOURSELF.

THE RIGHT ANSWER IS C.
C.INTERFACE VPC ENDPOINT

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf (516)
https://docs.aws.amazon.com/zh_tw/vpc/latest/userguide/vpc-endpoints.html

upvoted 46 times

rsimham 3 years, 3 months ago

Not sure if C is correct in this particular scenario.
From https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf
Page 202 of the SageMaker Guide has:
If you allowed access to resources from your VPC, enable direct internet access. For Direct internet access, choose Enable. Without internet access, you can't train or host models from notebooks on this notebook instance unless your VPC has a NAT gateway and your security group allows outbound connect

upvoted 2 times

Selectron 3 years, 3 months ago

There are two possible solutions, but the safer solution and easier is trough VPC endpoints.

You can connect to your notebook instance from your VPC through an interface endpoint in your Virtual Private Cloud (VPC) instead of connecting over the internet. When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network. And there is not problem that the notebooks does not have public internet.
Because Amazon SageMaker notebook instances support Amazon Virtual Private Cloud (Amazon VPC) interface endpoints that are powered by AWS PrivateLink. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets. Each VPC endpoint is represented by one or more Elastic Network Interfaces (ENIs) with private IP addresses in your VPC subnets...
so the Answer is C.

upvoted 5 times

rsimham 3 years, 3 months ago

A may the right answer

upvoted 1 times

tap123 Highly Voted 3 years, 3 months ago

C is correct. "The VPC interface endpoint connects your VPC directly to the Amazon SageMaker API or Runtime without an internet gateway, **NAT** device, VPN connection, or AWS Direct Connect connection." https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html

upvoted 16 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: C

Explanation:
The company's data security policy does not allow internet access, so the solution must allow Amazon SageMaker to function privately within the VPC without internet access.

VPC Interface Endpoints (AWS PrivateLink) for SageMaker allow services to communicate privately over the AWS network, without requiring an Internet Gateway (IGW) or NAT Gateway.

Explanation:
The company's data security policy does not allow internet access, so the solution must allow Amazon SageMaker to function privately within the VPC without internet access.

VPC Interface Endpoints (AWS PrivateLink) for SageMaker allow services to communicate privately over the AWS network, without requiring an Internet Gateway (IGW) or NAT Gateway.

upvoted 1 times

Denise123 11 months, 1 week ago

The answer is C.
- If you want to allow internet access, you must use a NAT gateway with access to the internet, for example through an internet gateway.
- If you don't want to allow internet access, create interface VPC endpoints (AWS PrivateLink) to allow Studio Classic to access the following services with the corresponding service names. You must also associate the security groups for your VPC with these endpoints.
This is exactly what's written in the ref. doc given in the answer section of the question. (Check page Security and Permissions 1120- 1121)
https://docs.aws.amazon.com/pdfs/sagemaker/latest/dg/sagemaker-dg.pdf

upvoted 1 times

phdykd 1 year ago

C.

To enable Amazon SageMaker service without enabling direct internet access to Amazon SageMaker notebook instances, while adhering to a corporate data security policy that restricts internet communication, the company can:

C. Create Amazon SageMaker VPC interface endpoints within the corporate VPC.

This option involves setting up VPC (Virtual Private Cloud) interface endpoints for Amazon SageMaker within the corporate VPC (Virtual Private Cloud). This is done using AWS PrivateLink, which allows private connectivity between AWS services using private IP addresses. By creating VPC interface endpoints, the traffic between the corporate VPC and Amazon SageMaker does not traverse the public internet, thereby meeting the corporate data security requirements.

upvoted 1 times

sonoluminescence 1 year, 2 months ago

Selected Answer: C

A would allow instances in a private subnet to initiate outbound internet traffic. This is against the requirement of no direct internet access.

upvoted 2 times

Sharath1783 1 year, 4 months ago

Selected Answer: C

NAT means data will go to internet. C is the right choice.

upvoted 2 times

Mickey321 1 year, 4 months ago

Selected Answer: C

Option c

upvoted 1 times

ADVIT 1 year, 6 months ago

Only C, endpoints.

upvoted 1 times

jackzhao 1 year, 10 months ago

C is correct, NAT allow outband traffic pass through internet.

upvoted 1 times

Nadia0012 1 year, 10 months ago

Selected Answer: C

To prevent SageMaker from providing internet access to your Studio notebooks, you can disable internet access by specifying the VPC only network access type when you the onboard to Studio or call CreateDomain API. As a result, you won't be able to run a Studio notebook unless your VPC has an interface endpoint to the SageMaker API and runtime, or a NAT gateway with internet access, and your security groups allow outbound connections.

upvoted 2 times

Nadia0012 1 year, 10 months ago

To disable direct internet access, under Direct Internet access, simply choose Disable – use VPC only , and select the Create notebook instance button at the bottom. You are ready to go.
from: https://aws.amazon.com/blogs/machine-learning/customize-your-amazon-sagemaker-notebook-instances-with-lifecycle-configurations-and-the-option-to-disable-internet-access/
#:~:text=To%20disable%20direct%20internet%20access%2C%20under%20Direct%20Internet%20access%2C%20simply,running%2C%20without%20direct%20internet%20access.

upvoted 1 times

Nadia0012 1 year, 10 months ago

If you want to allow internet access, you must use a example through an internet gateway. If you don't want to allow internet access, NAT gateway with access to the internet, for create interface VPC endpoints (AWS PrivateLink) to allow Studio to access the following services with the corresponding service names. You must also associate the security groups for your VPC with these endpoints.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: C

A VPC interface endpoint is a private connection between a VPC and Amazon SageMaker that is powered by AWS PrivateLink. With a VPC interface endpoint, traffic between the VPC and Amazon SageMaker never leaves the Amazon network.

upvoted 3 times

Ob1KN0B 2 years, 4 months ago

Selected Answer: C

Page 3438 of https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf

upvoted 2 times

ovokpus 2 years, 6 months ago

Selected Answer: C

VPC Interface endpoints

upvoted 3 times

gcpwhiz 3 years, 2 months ago

If the question just had the last sentence, the answer would be A or C, per this page:https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html. "To disable direct internet access, you can specify a VPC for your notebook instance. By doing so, you prevent SageMaker from providing internet access to your notebook instance. As a result, the notebook instance won't be able to train or host models unless your VPC has an interface endpoint (PrivateLink) or a NAT gateway, and your security groups allow outbound connections."

HOWEVER, the question has more context that internet access is not allowed by the corporate policy. ("When you use a VPC interface endpoint, communication between your VPC and the notebook instance is conducted entirely and securely within the AWS network.") Therefore, the answer must be ONLY C.

upvoted 5 times

scuzzy2010 3 years, 2 months ago

Answer is C. From https://docs.aws.amazon.com/sagemaker/latest/dg/interface-vpc-endpoint.html ->
"The VPC interface endpoint connects your VPC directly to the SageMaker API or Runtime without an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. The instances in your VPC don't need public IP addresses to communicate with the SageMaker API or Runtime."

upvoted 3 times

cloud_trail 3 years, 2 months ago

I see a lot of people employing pretzel logic to try to explain why they should be using NAT. The question states no internet communication. Period. No internet means no NAT. Answer is C.

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 46 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 46
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.
Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

A. Listwise deletion

B. Last observation carried forward

C. Multiple imputation

D. Mean substitution

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 2:49 a.m.*

Comments

rsimham Highly Voted 2 years, 9 months ago
C looks correct since multiple imputation can be performed based on the related variable as given in the question
upvoted 26 times

harmanbirstudy Highly Voted 2 years, 8 months ago
Multiple Imputation by Chained Equations or MICE, as per udemy this is always the best answer of all
upvoted 8 times

sonoluminescence Most Recent 8 months, 2 weeks ago
Why not D:

Doesn't Account for Relationships:
Mean substitution doesn't take into account the potential relationships between variables. In the scenario you provided, it's believed that other columns could help in reconstructing the missing data. Using only the mean of the missing column doesn't leverage this potential inter-column relationship.

Assumption of Missing Completely at Random (MCAR):
Mean substitution often operates under the assumption that the data is Missing Completely at Random (MCAR). In reality, data might be missing for a reason, and that reason might relate to other observed variables. Using mean substitution in such cases can introduce biases.
upvoted 2 times

loict 10 months ago

Selected Answer: C

A. NO - Listwise deletion is just dropping rows
B. NO - does not reconstruct the data based on other fields
C. YES - by definition
D. NO - does not reconstruct the data based on other fields

upvoted 2 times

DavidRou 10 months ago

Selected Answer: C

MICE is the algorithm to choose here

upvoted 1 times

Mickey321 10 months, 2 weeks ago

Selected Answer: C

Option C

upvoted 1 times

AjoseO 1 year, 5 months ago

Selected Answer: C

Multiple imputation is a statistical technique for handling missing data that involves generating multiple versions of the dataset with missing values filled in, and then combining the results to produce a single, complete dataset.

This approach takes into account the relationship between variables in the dataset, and uses statistical models to predict missing values based on the information in other columns. This helps to preserve the integrity of the dataset by avoiding the introduction of bias or systematic error into the results.

upvoted 5 times

[Removed] 2 years, 7 months ago

I am trying to understand why Mean Substitution is not the solution. Imputation typically uses the mean if the missing data is random, implying the substitution is not biased.

upvoted 2 times

cpal012 1 year, 4 months ago

Mean substitution is limited to the current column. In this case, the requirement is to impute missing data from other columns

upvoted 3 times

rhuanca 2 years, 1 month ago

Reason is if you replace 30% of the missing values , likely you will bias the variable.

upvoted 1 times

syu31svc 2 years, 8 months ago

If it's handling missing data then imputation comes into play
Answer is C 100%

upvoted 1 times

Wira 2 years, 8 months ago

https://www.countants.com/blogs/heres-how-you-can-configure-automatic-imputation-of-missing-data/ C

upvoted 1 times

roytruong 2 years, 9 months ago

it's C

upvoted 1 times

dhs227 2 years, 9 months ago

A common strategy used to impute missing values is to replace missing values with the mean or median value. It is important to understand your data before choosing a strategy for replacing missing values. https://docs.aws.amazon.com/machine-learning/latest/dg/feature-processing.html

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 45 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 45
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.
The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.
Which solution should the Data Scientist build to satisfy the requirements?

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.

D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 4:32 p.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago
Kinesis Data Analytics NO PARQET FORMAT,
BESIDES THAT JSON NO NEED TO STORE IN S3.
RDS ISN'T serverless ingestion and analytics solution

ANSWER IS A.

upvoted 32 times

georgeZ Highly Voted 3 years, 3 months ago

I thinks it should be A please check https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: A

Amazon Kinesis Data Firehose

Ingests real-time data with automatic buffering.
Supports built-in transformation to Apache Parquet/ORC before writing to Amazon S3.
Requires minimal code and infrastructure.
AWS Glue Data Catalog

Catalogs the schema for structured querying.
Enables Athena to directly query data in S3.
Amazon Athena

Serverless SQL querying on S3-based datasets.
Can connect to BI tools (Tableau, QuickSight) via JDBC.

upvoted 1 times

Alice1234 11 months, 1 week ago

A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use Amazon Kinesis Data Firehose to buffer and transform the streaming JSON data to a columnar format like Apache Parquet or ORC using the AWS Glue Data Catalog before delivering to Amazon S3. Analysts can then query the data using Amazon Athena and connect to BI dashboards using the Athena JDBC connector. This solution is serverless, manages high-velocity data streams, supports SQL queries, and connects to BI tools—all while being highly available.

upvoted 3 times

loict 1 year, 4 months ago

Selected Answer: C

A. YES - we need a catalog to create parquet (https://docs.aws.amazon.com/firehose/latest/APIReference/API_SchemaConfiguration.html)
B. NO - no need for extra staging
C. NO - no need for extra staging
D. NO - we need a catalog

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

Option A

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: A

A is correct. For those selecting B, answer me: how exactly the json will be stored in the S3? It's not mentioned in the answer. For me it's an incomplete solution.

upvoted 2 times

Ajose0 1 year, 11 months ago

Selected Answer: A

This solution leverages AWS Glue to create a schema of the incoming data format, which helps to buffer and convert the records to a query-optimized, columnar format without data loss.

The Amazon Kinesis Data Firehose delivery stream is used to stream the data and transform it to Apache Parquet or ORC format using the AWS Glue Data Catalog, and the data is stored in Amazon S3, which is highly available. The Analysts can then query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena JDBC connector.

This solution provides a serverless, scalable, and cost-effective solution for real-time streaming data ingestion and analytics.

upvoted 3 times

sqavi 1 year, 11 months ago

Selected Answer: A

Since you want to buffer and convert data so A is correct answer. No other option is fulfilling this requirement

upvoted 2 times

Peeking 2 years, 1 month ago

Selected Answer: A

I go for A. However, I am not sure why AWS Glue is very important here given that Firehose can convert JSON to parquet.

upvoted 2 times

Tony_1406 1 year, 8 months ago

If I haven't remembered correctly. Athena requires a schema of the S3 object to perform SQL query. That's probably why we need Glue for the

schema
upvoted 1 times

ZSun 1 year, 8 months ago
once you ingest the data using Kinesis Firehose, you can set "generate table" and automatically create Glue schema. I think both Glue and Firehose can do data conversion from JSON to parquet.
upvoted 1 times

itallomd 2 years, 1 month ago
Why AWS Glue is needed? Firehose could convert to parquet directly...
upvoted 2 times

587df71 6 months, 1 week ago
https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html

Amazon Data Firehose requires a schema to determine how to interpret that data. Use AWS Glue to create a schema in the AWS Glue Data Catalog. Amazon Data Firehose then references that schema and uses it to interpret your input data
upvoted 1 times

Ccindy 2 years, 1 month ago
Selected Answer: B
Kinesis Data Analytics is near real-time, not real time
upvoted 1 times

ryuhei 2 years, 3 months ago
Selected Answer: A
Answer is ”A”
upvoted 1 times

ovokpus 2 years, 6 months ago
Selected Answer: A
The difference between "real-time" and "near-real-time" is pretty semantic(60s). The fact that the data comes through kinesis data streams (real time) is implied as the only valid input to firehose.
upvoted 1 times

ovokpus 2 years, 6 months ago
Mind you, "the ingestion process must buffer and transform incoming records from JSON to a query-optimized, columnar format"

That is exactly what kinesis firehose does.
"Kinesis Data Firehose buffers incoming data before delivering it to Amazon S3. You can configure the values for S3 buffer size (1 MB to 128 MB) or buffer interval (60 to 900 seconds), and the condition satisfied first triggers data delivery to Amazon S3."

See link: https://aws.amazon.com/kinesis/data-firehose/faqs/
#:~:text=Kinesis%20Data%20Firehose%20buffers%20incoming,data%20delivery%20to%20Amazon%20S3.
upvoted 3 times

TerrancePythonJava 2 years, 10 months ago
Selected Answer: B
Data Firehose is always Near Real Time not Real Time. The prompt clearly states that process must be done in Real Time.
upvoted 1 times

anttan 3 years, 1 month ago
Why A? Firehose is near real-time, and not real-time which is a requirement
upvoted 1 times

cpal012 1 year, 10 months ago
There is no requirement for real time processing. It says the data is in real time but the processing of that data should buffer
upvoted 2 times

harmanbirstudy 3 years, 2 months ago
ANSWER is A -- and every statement in it is accurate.
Firehose does integrate with GLue data catalog and it also "Buffers" the data .
"When Kinesis Data Firehose processes incoming events and converts the data to Parquet, it needs to know which schema to apply." This is achived by glue data catalog and athena and it works on real-time data ingest.See link below.

https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/
upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 44 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 44
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist has created a deep learning neural network model that performs well on the training data but performs poorly on the test data.
Which of the following methods should the Specialist consider using to correct this? (Choose three.)

A. Decrease regularization.

B. Increase regularization.

C. Increase dropout.

D. Decrease dropout.

E. Increase feature combinations.

F. Decrease feature combinations.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 4:10 p.m.*

Comments

cybe001 Highly Voted 3 years, 9 months ago
Yes, answer is BCF
upvoted 24 times

Phong Highly Voted 3 years, 9 months ago
Go for BCF
upvoted 14 times

ninomfr64 Most Recent 1 year ago
Selected Answer: BCF
I think here the point is around the definition of "feature combinations".

If you refer to it as "combine the features to generate a smaller but more effective feature set" this would end up to a smaller feature set thus a good thing for overfitting.

However, if you refer to it as "combine the features to generate additional features" this would end up to a larger feature set thus a bad thing for overfitting.

Also, in some cases you implement feature combinations in your model (see hidden layers in feed-forward network) thus increasing model complexity which is bad for overfitting.

To me this question is poorly worded. I would pick F as my best guess is that you need to implement feature combination in your model, thus decreasing feature combination decrease complexity hence improving with overfitting issue

upvoted 5 times

cloudera3 1 year ago

Great callout - what exactly the Feature combination is performing has not been elaborated

It can be: Using PCA or t-SNE, it is essentially optimizing features - good to address overfitting, and should be done

Or, it can be: Using Cartesian Product, features are being combined to create additional features - this will aid overfitting and should NOT be done.

Wish questions and answer options are written clearly so that there is no room for ambiguity. Especially, taking into account that in real life, these kind of communication/write-up will trigger follow-up questions until addressed satisfactorily.

upvoted 1 times

Denise123 1 year, 3 months ago

Selected Answer: BCE

About option E:
When increasing feature combinations, the goal is not to simply add more features indiscriminately, which could indeed lead to overfitting. Instead, it involves selecting and combining features in a way that captures important patterns and relationships in the data.
When done effectively, increasing feature combinations can help the model generalize better to unseen data by providing more informative and discriminative features, thus reducing the risk of overfitting.

upvoted 1 times

Piyush_N 1 year, 4 months ago

Selected Answer: BCF

If your model is overfitting the training data, it makes sense to take actions that reduce model flexibility. To reduce model flexibility, try the following:

Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins.

Increase the amount of regularization used.
https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

upvoted 1 times

Neet1983 1 year, 6 months ago

Selected Answer: BCF

Best choices are B (Increase regularization), C (Increase dropout), and F (Decrease feature combinations), as these techniques are effective in reducing overfitting and improving the model's ability to generalize to new data.

upvoted 1 times

akgarg00 1 year, 8 months ago

Selected Answer: BCE

BCE The model has learnt training data. One approach is to increase complexity by increasing the features or remove some features to increase bias. In deep learning, i thinking increasing feature set is more workable.

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: BCF

B-C-F. All of those options can be used to reduce model complexity and thus: overfit

upvoted 1 times

SRB1337 2 years ago

its BCF

upvoted 1 times

jackzhao 2 years, 4 months ago

BCF is correct.

upvoted 2 times

AjoseO 2 years, 5 months ago

Selected Answer: BCF

Increasing regularization helps to prevent overfitting by adding a penalty term to the loss function to discourage the model from learning the noise in the data.

Increasing dropout helps to prevent overfitting by randomly dropping out some neurons during training, which forces the model to learn more robust representations that do not depend on the presence of any single neuron.

Decreasing the number of feature combinations helps to simplify the model, making it less likely to overfit.

upvoted 6 times

Tomatoteacher 2 years, 5 months ago

Selected Answer: BCE

I see all the comments for BCF, although when you look at F it just says decrease 'feature combinations', not features themselves. In one way to decrease feature combinations results in having more features (less feature engineering), which in turn will cause more overfitting. Unless the question in badly worded, saying less feature combinations just mean those combinations, which components will not be used, then it has to be BCE.

upvoted 1 times

cpal012 2 years, 3 months ago

Decrease feature combinations - too many irrelevant features can influence the model by drowning out the signal with noise

upvoted 1 times

AjoseO 2 years, 5 months ago

Increasing the number of feature combinations can sometimes improve the performance of a model if the model is underfitting the data.

However, in this context, it is not likely to be a solution to overfitting.

upvoted 1 times

Shailendraa 2 years, 10 months ago

BCF - Always remember in case of overfitting - reduce features, Add regularisation and increase dropouts.

upvoted 3 times

ahquiceno 3 years, 8 months ago

BCE: The main objective of PCA (technic to feature combination) is to simplify your model features into fewer components to help visualize patterns in your data and to help your model run faster. Using PCA also reduces the chance of overfitting your model by eliminating features with high correlation.
https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-principal-component-analysis-pca-fea1ca817fe6

upvoted 2 times

uninit 2 years, 5 months ago

AWS Documentation explicitly mentions reducing feature combinations to prevent overfitting - https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html
It's B C F

upvoted 3 times

cloud_trail 3 years, 8 months ago

B/C/F Easy peasy.

upvoted 1 times

apnu 3 years, 8 months ago

BCF 100%

upvoted 1 times

obaidur 3 years, 8 months ago

BCF
F
explained in AWS document:
Feature selection: consider using fewer feature combinations, decrease n-grams size, and decrease the number of numeric attribute bins. Increase the amount of regularization used
https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 43 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 43
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Science team within a large company uses Amazon SageMaker notebooks to access data stored in Amazon S3 buckets. The IT Security team is concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy.
The company mandates that all instances stay within a secured VPC with no internet access, and data communication traffic must stay within the AWS network.
How should the Data Science team configure the notebook instance placement to meet these requirements?

A. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Place the Amazon SageMaker endpoint and S3 buckets within the same VPC.

B. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Use IAM policies to grant access to Amazon S3 and Amazon SageMaker.

C. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has S3 VPC endpoints and Amazon SageMaker VPC endpoints attached to it.

D. Associate the Amazon SageMaker notebook with a private subnet in a VPC. Ensure the VPC has a NAT gateway and an associated security group allowing only outbound connections to Amazon S3 and Amazon SageMaker.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 3:20 p.m.*

Comments

DonaldCMLIN Highly Voted 2 years, 9 months ago
NAT gateway COULD GO OUT TO THE INTERNET AND DOWNLOAD BACK MALICIOUS
D. IS NOT A GOOD ANSWER.

THE SAFE ONE IS ANSWER C. ASSOCIATE WITH VPC_ENDPOINT AND S3_ENDPOINT
upvoted 35 times

BigEv Highly Voted 2 years, 9 months ago
C is correct
We must use the VPC endpoint (either Gateway Endpoint or Interface Endpoint)to comply with this requirement "Data communication traffic must stay within the AWS network".
https://docs.aws.amazon.com/sagemaker/latest/dg/notebook-interface-endpoint.html

upvoted 23 times

loict Most Recent 10 months ago

Selected Answer: C

A. NO - We don't place a S3 bucket in a VPC, it is always in AWS Service Account
B. NO - without an S3 VPC endpoint, traffic will go through the Internet
C. YES - we need endpoints for both SageMaker and S3 to avoid Internet traffic
D. NO - we need endpoints for both SageMaker and S3 to avoid Internet traffic

upvoted 2 times

Mickey321 10 months, 2 weeks ago

Selected Answer: C

Option C

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: C

C is the correct. A is not so correct, because it's possible to communicate two different VPCs inside AWS network (which is not optimized).

upvoted 1 times

Ajose0 1 year, 5 months ago

Selected Answer: C

This configuration would meet the company's requirements for security, as the notebook instance would be placed within a private subnet in a VPC, and data communication traffic would stay within the AWS network through the use of VPC endpoints for S3 and Amazon SageMaker.

Additionally, the VPC would not have internet access, further reducing the security risk.

upvoted 2 times

rb39 1 year, 10 months ago

C - "and data communication traffic must stay within the AWS network." that discards D

upvoted 2 times

StelSen 2 years, 8 months ago

Answer should be C. Because, Security team don't want Internet Access, Option-D has NAT and will get to Internet somehow. Also connecting S3 and SageMaker EC2 instance via VPC endpoints is best way to secure the resources.

upvoted 4 times

cloud_trail 2 years, 8 months ago

Using a NAT gateway is the old way to do it. Option C is the way to do it now. https://cloudacademy.com/blog/vpc-endpoint-for-amazon-s3/#:~:text=Accessing%20S3%20the%20old%20way%20%28without%20VPC%20Endpoint%29,has%20no%20access%20to%20any%20outside%20public%20resources

upvoted 2 times

harmanbirstudy 2 years, 8 months ago

"and data communication traffic must stay within the AWS network" , NAT gateway will always go over the Internet to access S3.with NAT you can put your instances in private subnet and NAT itself in public subnet , but still in order to access S3 it will go over the internet.
SO answer cannot be D.
-- C is the only correct option here , as S3 VPC endpoints is a real thing "google it" and it sole purpose is to create route from VPC endpoint to S3 , without going over the Internet.

upvoted 3 times

scuzzy2010 2 years, 8 months ago

C is correct answer. D is only applicable -"If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. " https://docs.aws.amazon.com/sagemaker/latest/dg/host-vpc.html

upvoted 3 times

v24143 2 years, 8 months ago

D is correct

upvoted 1 times

krakow1234 2 years, 8 months ago

Answer is D, read third paragraph https://docs.aws.amazon.com/sagemaker/latest/dg/appendix-notebook-and-internet-access.html

upvoted 1 times

Potato_Noodle 2 years, 8 months ago

NAT is the way that a VPC connect to internet and other ASW service when there is NO INTERNET ACCESS FOR VPC. Thus the answer is D.

upvoted 1 times

Th3Dud3 2 years, 8 months ago

"concerned that internet-enabled notebook instances create a security vulnerability where malicious code running on the instances could compromise data privacy." NAT Gateway does not mitigate this risk!

upvoted 2 times

yeetusdeleetus 2 years, 8 months ago

This is the correct answer.

If this answer is confusing, study some of the associate exams before going for this one. VPC endpoint and NAT gateway are similar, but NAT gateway is for giving resources in the VPC the chance to initiate connections with the internet, whereas a VPC endpoint only allows it to go to other AWS services, which is the best solution for this question.

upvoted 2 times

Th3Dud3 2 years, 8 months ago

C:
If you configure your VPC so that it doesn't have internet access, models that use that VPC do not have access to resources outside your VPC. If your model needs access to resources outside your VPC, provide access with one of the following options:

If your model needs access to an AWS service that supports interface VPC endpoints, create an endpoint to connect to that service. For a list of services that support interface endpoints, see VPC Endpoints in the Amazon VPC User Guide. For information about creating an interface VPC endpoint, see Interface VPC Endpoints (AWS PrivateLink) in the Amazon VPC User Guide.

If your model needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. For information about setting up a NAT gateway for your VPC, see Scenario 2: VPC with Public and Private Subnets (NAT) in the Amazon Virtual Private Cloud User Guide.

upvoted 5 times

sebtac 2 years, 8 months ago

what is the difference between A & C? are both answers OK?

upvoted 1 times

jrff 1 year, 8 months ago

It is not enough for sagemaker to communicate to S3 if both of them are inside the same VPC. Sagamaker inside a VPC needs to create a endpoint to connect to other AWS services which has endpoint too.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 42 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 42
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is implementing a full Bayesian network on a dataset that describes public transit in New York City. One of the random variables is discrete, and represents the number of minutes New Yorkers wait for a bus given that the buses cycle every 10 minutes, with a mean of 3 minutes.
Which prior probability distribution should the ML Specialist use for this variable?

A. Poisson distribution

B. Uniform distribution

C. Normal distribution

D. Binomial distribution

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 3:03 p.m.*

Comments

ComPah Highly Voted 3 years, 9 months ago

A
If you have information about the average (mean) number of things that happen in some given time period / interval,Poisson distribution can give you a way to predict the odds of getting some other value on a given future day

upvoted 57 times

swagy Highly Voted 3 years, 9 months ago

Ans: A
https://brilliant.org/wiki/poisson-distribution/

upvoted 8 times

Mobasher Most Recent 5 months, 1 week ago

Selected Answer: B

ChatGPT's answer: B

Explanation

The problem describes a random variable representing the waiting time for a bus, where buses arrive every 10 minutes, and the mean waiting time is 3 minutes.

In such a periodic arrival process, the waiting time follows a Uniform Distribution because:
- Any given person's waiting time is equally likely to be any value between 0 and 10 minutes.
- There is no clustering around a particular value—every moment within the cycle is equally probable.

Thus, the waiting time follows a Uniform(0, 10) distribution.

upvoted 1 times

Mobasher 5 months, 1 week ago

Why Not the Other Options?

(A) Poisson Distribution
Poisson is used for counting discrete events over a fixed period (e.g., number of buses arriving per hour). Since waiting time is continuous, Poisson is not appropriate.
(C) Normal Distribution
Normal (Gaussian) distribution assumes values cluster around the mean and extend infinitely. Here, waiting time is evenly spread between 0–10 minutes, not forming a bell curve.
(D) Binomial Distribution
Binomial is used for counting successes in a fixed number of trials (e.g., flipping a coin multiple times). Waiting time is continuous, not a count of discrete occurrences.

upvoted 1 times

growe 6 months, 2 weeks ago

Selected Answer: B

Buses cycle every 10 minutes, and waiting time can be modeled as a uniform random variable between [0, 10] minutes.
The average waiting time of 3 minutes suggests that waiting is uniformly distributed, not event-based like Poisson.
If buses arrive every 10 minutes and riders arrive randomly, the waiting time follows a Uniform Distribution (B) because:

The arrival process is regular (every 10 minutes).
There's no stochastic randomness in the bus arrival schedule, ruling out Poisson.
Poisson would apply if buses arrived randomly at an average rate rather than at fixed intervals.

upvoted 2 times

87ebc7d 7 months, 3 weeks ago

B
Poisson is suitable for modeling the number of events (like buses arriving) in a fixed time frame, not the time between events when the events occur at regular intervals. The waiting time variable is not about the count of buses but rather the time to the next bus, which is evenly distributed.

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: A

A is correct

upvoted 1 times

Fred93 1 year, 9 months ago

Selected Answer: A

Poisson distribution is discrete, and gives the number of events that occur in a given time interval

upvoted 2 times

loict 1 year, 10 months ago

Selected Answer: A

A. YES - Poisson distribution is discrete, and gives the number of events that occur in a given time interval
B. NO - Uniform distribution is continuous, we want discrete
C. NO - Normal distribution is continuous we want discrete
D. NO - Binomial distribution give the probability that a random variable is A or B (possibly in with different weight)

upvoted 2 times

Mickey321 1 year, 10 months ago

Selected Answer: A

Option A indeed

upvoted 1 times

Nadia0012 2 years, 4 months ago

Selected Answer: A

ANSWER IS A
https://www.investopedia.com/terms/d/discrete-distribution.asp

upvoted 2 times

bakarys 2 years, 4 months ago

Selected Answer: A

The Poisson distribution is commonly used for count data, which is the case here as we are interested in the number of minutes New Yorkers wait for a bus. The Poisson distribution is characterized by a single parameter, lambda, which represents the mean and variance of the distribution. In this case, the mean is 3 minutes, so we would set lambda to 3. The Poisson distribution assumes that events occur independently of each other, which is a reasonable assumption in this case since the waiting time for each individual is likely to be independent of the waiting time for others.

upvoted 4 times

AjoseO 2 years, 5 months ago

Selected Answer: A

The Poisson distribution is a discrete probability distribution that is commonly used to model the number of events that occur in a fixed interval of time, given an average rate of occurrence.

Since the buses cycle every 10 minutes and the mean wait time is 3 minutes, it is reasonable to assume that the number of minutes New Yorkers wait for a bus can be modeled by a Poisson distribution.

upvoted 3 times

Tomatoteacher 2 years, 5 months ago

Selected Answer: A

100% A, as discrete, while binomial has to be binary data (success or failure)

upvoted 1 times

Sonoko 2 years, 7 months ago

Selected Answer: A

A is a discrete distribution

upvoted 1 times

Peeking 2 years, 7 months ago

I do choose Poisson. A.

upvoted 1 times

Shailendraa 2 years, 10 months ago

12-sep exam

upvoted 3 times

Shailendraa 2 years, 10 months ago

Answer is A .. these types on footfalls ,etc ..answer always Poisson-distribution

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 41 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 41

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is building a prediction model for a large number of features using linear models, such as linear regression and logistic regression.
During exploratory data analysis, the Specialist observes that many features are highly correlated with each other. This may make the model unstable.
What should be done to reduce the impact of having such a large number of features?

A. Perform one-hot encoding on highly correlated features.

B. Use matrix multiplication on highly correlated features.

C. Create a new feature space using principal component analysis (PCA)

D. Apply the Pearson correlation coefficient.

Show Suggested Answer

by cybe001 at *Jan. 12, 2020, 10:27 p.m.*

Comments

cybe001 Highly Voted 2 years, 9 months ago

C is correct

upvoted 19 times

syu31svc Highly Voted 2 years, 8 months ago

You want to reduce features/dimension so PCA is the answer

upvoted 5 times

kaike_reis Most Recent 11 months, 2 weeks ago

Selected Answer: C

C is the way

upvoted 3 times

FloKo 11 months, 2 weeks ago

Selected Answer: C

C is correct.
D could be correct if the correlation is used to omit features.

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: C

PCA and T-SNE are for solving the curse of dimensionality mencioned here!

upvoted 1 times

DS2021 1 year, 6 months ago

I assume PCA is for unsupervised learning!...and the scenario in the question looks like supervised learning

upvoted 1 times

GiyeonShin 1 year, 4 months ago

data (x, y) --> (PCA) --> preprocessed data(x', y) --> learning
why not for supervised learning?

upvoted 1 times

BethChen 1 year, 7 months ago

Selected Answer: C

Tricky. The sentence 'many features are highly correlated with each other' is no use.

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

It's. PCA removes such correlation.

upvoted 1 times

Shailendraa 1 year, 10 months ago

Answer C: Read through this carefully "What should be done to reduce the impact of having such a large number of features?" only answer comes in mind PCA

upvoted 1 times

Urban_Life 2 years, 9 months ago

Of course, it's PCA.

upvoted 1 times

C10ud9 2 years, 9 months ago

PCA is the solution. So, answer is C

upvoted 2 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 40 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 40
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning team uses Amazon SageMaker to train an Apache MXNet handwritten digit classifier model using a research dataset. The team wants to receive a notification when the model is overfitting. Auditors want to view the Amazon SageMaker log activity report to ensure there are no unauthorized API calls.
What should the Machine Learning team do to address the requirements with the least amount of code and fewest steps?

A. Implement an AWS Lambda function to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

B. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

C. Implement an AWS Lambda function to log Amazon SageMaker API calls to AWS CloudTrail. Add code to push a custom metric to Amazon CloudWatch. Create an alarm in CloudWatch with Amazon SNS to receive a notification when the model is overfitting.

D. Use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. Set up Amazon SNS to receive a notification when the model is overfitting

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 10:02 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 9 months ago
THE ANSWER SHOULD BE B.
YOU DON'T NEED TO THROUGH LAMBDA TO INTERGE CLOUDTRAIL

Log Amazon SageMaker API Calls with AWS CloudTrail
https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html
upvoted 41 times

rajs Highly Voted 3 years, 9 months ago
Agreed B for the following reasons

CloudTrail logs captured in S3 without any code/lambda
The custom metrics can be published to Cloudwatch...in this case it would be a test for overfit on MXNET .... which will set off an alarm .... which can then be subscribed on SNS

upvoted 11 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

Breakdown of the Chosen Solution (B)
Use AWS CloudTrail to log SageMaker API calls to Amazon S3 ✅

CloudTrail automatically logs all AWS API activity, including SageMaker API calls, for auditing.
S3 stores these logs securely for auditor review.
Push custom metrics to Amazon CloudWatch ✅

Model overfitting can be detected using a custom CloudWatch metric (e.g., validation loss increasing while training loss decreases).
The SageMaker training script can push loss values to CloudWatch during training.
Create a CloudWatch alarm + SNS notification ✅

Set a CloudWatch alarm on the overfitting metric (e.g., validation loss surpassing a threshold).
Use Amazon SNS to send a notification (email, SMS, or Lambda trigger) when the alarm is triggered.

upvoted 1 times

MultiCloudIronMan 8 months, 1 week ago

Selected Answer: B

Option D involves using AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3 and setting up Amazon SNS to receive a notification when the model is overfitting. While this approach addresses the logging requirement, it does not provide a mechanism for pushing custom metrics to Amazon CloudWatch, which is necessary for monitoring model performance and detecting overfitting. So 'B ' is correct

upvoted 2 times

Chiquitabandita 1 year, 2 months ago

Selected Answer: D

https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics

It detects hardware resource usage issues (such as CPU, GPU, and I/O bottlenecks) and non-convergent model issues (such as overfitting, disappearing gradients, and tensor explosion).

why couldn't the answer be D, as this covers all of the requirements, and B seems to add an extra step with adding push code, when it already has a builtin metric for overfitting.

upvoted 1 times

Mobasher 5 months, 1 week ago

This would have been correct had the question not mentioned that the algorithm is "hand-written" which means it's not a built in algorithm. So, for SageMaker AI to understand your custom algorithm's metrics, it needs a regex definition to apply to the logs in order to generate those custom metrics and then alert on them using CW Alarms and SNS to deliver notifications. See https://docs.aws.amazon.com/sagemaker/latest/dg/define-train-metrics.html

upvoted 1 times

Aja1 1 year, 2 months ago

Custom metric Need to built and pushed.

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

A. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed
B. YES - the chain works
C. NO - CloudTrail has built-in SageMaker API calls tracking, no lambda needed
D. NO - CloudTrail has not specific Amazon SageMaker integration to detect overfitting

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

ADVIT 2 years ago

"least amount of code and fewest steps?"
I think it's D.

upvoted 2 times

kukreti18 2 years ago

Agreed, with less code effort.

upvoted 1 times

Paolo991 2 years, 3 months ago

I would consider D as well.

You can just setup a SNS that is triggered by a built-in action like here:

https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-actions.html

You can see that overfitting is a built-in rule for MXNet from here:
https://docs.aws.amazon.com/sagemaker/latest/dg/debugger-built-in-rules.html

Not that B is not working. Maybe the question was prior to this new solution.

upvoted 2 times

khchan123 1 year, 8 months ago

The loss_not_decreasing, overfit, overtraining, and stalled_training_rule monitors if your model is optimizing the loss function without those training issues. If the rules detect training anomalies, the rule evaluation status changes to IssueFound. You can set up automated actions, such as notifying training issues and stopping training jobs using Amazon CloudWatch Events and AWS Lambda. For more information, see Action on Amazon SageMaker Debugger Rules.
https://docs.aws.amazon.com/sagemaker/latest/dg/use-debugger-built-in-rules.html

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

It's B.

upvoted 1 times

AjoseO 2 years, 4 months ago

Selected Answer: B

AWS CloudTrail provides a history of AWS API calls made on the account. The Machine Learning team can use AWS CloudTrail to log Amazon SageMaker API calls to Amazon S3. They can then use CloudWatch to create alarms and receive notifications when the model is overfitting.

To ensure auditors can view the Amazon SageMaker log activity report, the team can add code to push a custom metric to Amazon CloudWatch. This provides a single place to view and analyze logs across all the services and resources in the environment.

upvoted 1 times

sonalev419 3 years, 8 months ago

B. cloudwatch + metrics from sagemaker + sns https://docs.aws.amazon.com/fr_fr/sagemaker/latest/dg/training-metrics.html#define-train-metrics

upvoted 4 times

ybad 3 years, 8 months ago

B requires the least amount of code and satisfies all conditions

upvoted 2 times

tochiebby 3 years, 8 months ago

What does this line do?

"Add code to push a custom metric to Amazon CloudWatch"

upvoted 1 times

Omar_Cascudo 3 years, 8 months ago

It creates a metric for overfitting (accuracy of training data and accuracy of test data).

upvoted 5 times

jonclem 3 years, 8 months ago

Its not B. Why would you use CloudTrail?

Having used Lambda for API calls I'm inclined to agree with the original answer, C.

upvoted 1 times

Pja1 3 years, 8 months ago

https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

upvoted 3 times

fhuadeen 3 years, 8 months ago

Because that is the only job of CloudTrail - to log actions taken on your AWS account. So why need a Lambda function to trigger it?

upvoted 3 times

Antriksh 3 years, 9 months ago

B it is

upvoted 2 times

C10ud9 3 years, 9 months ago

B it is

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 39 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 39

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist built an image classification deep learning model. However, the Specialist ran into an overfitting problem in which the training and testing accuracies were 99% and 75%, respectively.

How should the Specialist address this issue and what is the reason behind it?

A. The learning rate should be increased because the optimization process was trapped at a local minimum.

B. The dropout rate at the flatten layer should be increased because the model is not generalized enough.

C. The dimensionality of dense layer next to the flatten layer should be increased because the model is not complex enough.

D. The epoch number should be increased because the optimization process was terminated before it reached the global minimum.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 9:23 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

DROPOUT HELPS PREVENT OVERFITTING
https://keras.io/layers/core/#dropout

THE BEAUTIFUL ANSER SHOULD BE B.

upvoted 55 times

rsimham 3 years, 3 months ago

agree. it should be B

upvoted 10 times

syu31svc Highly Voted 3 years, 2 months ago

https://kharshit.github.io/blog/2018/05/04/dropout-prevent-overfitting

Answer is B 100%

upvoted 5 times

fm99 Most Recent 9 months ago

Selected Answer: B

Increasing dropout rate will reduce complexity of the model which inturn reduces overfitting

upvoted 1 times

VR10 10 months, 4 weeks ago

This is clearly B, dont get why the answer is marked as D.

upvoted 1 times

endeesa 1 year, 1 month ago

Selected Answer: B

Regularization will seek to obtain similar accuracies in train and test sets. Anything else will make the overfitting worse

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: B

B is correct, D so stup*d answer

upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: B

A. NO - accuracy on training set is high
B. YES - increased dropout rate => reduce model complexity => less overfitting
C. NO - we want to reduce model complexity
D. NO - the model converged

upvoted 2 times

DavidRou 1 year, 4 months ago

Selected Answer: B

I don't understand why the highlighted "right" answer is D. To increase the number of epochs will make the situation even worse than it is; dropout is the right action to take in this case

upvoted 2 times

kaike_reis 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 1 times

nilmans 1 year, 6 months ago

agree, B makes more sense here

upvoted 1 times

soonmo 1 year, 7 months ago

Selected Answer: B

Definitely B because overfitting comes from complex model that captures patterns of training data well. But D is getting this model more complex, worsening overfitting.

upvoted 1 times

soonmo 1 year, 7 months ago

Correct my reasoning! D is worsening overfitting because it feeds more data after overfitting arises. D is used for underfitted models.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: B

Increasing Epoch only makes things worse on a overfitting model. You should perform regularization by introducing drop outs to generalize the model.

upvoted 1 times

user009 1 year, 9 months ago

Option B is the correct answer because increasing the dropout rate at the flatten layer helps prevent overfitting by randomly dropping out units during training, effectively creating a more robust model that can generalize better to new data. Dropout is a regularization technique that helps prevent overfitting by forcing the model to learn redundant representations of the data. By increasing the dropout rate at the flatten layer, the model becomes more generalized, which should help to improve the testing accuracy.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: B

Overfitting occurs when a model is too complex and memorizes the training data instead of learning the underlying pattern. As a result, the model performs well on the training data but poorly on new, unseen data.

Increasing the dropout rate, a regularization technique, can help combat overfitting by randomly dropping out some neurons during training, which prevents the model from relying too heavily on any single feature.

upvoted 1 times

sqavi 1 year, 11 months ago

Selected Answer: B

Model is overfitting, I will go with option B, increasing epoch will cause more overfitting

upvoted 2 times

desperatestudent 1 year, 11 months ago

Selected Answer: B

it should answer B.

upvoted 1 times

Shailendraa 2 years, 4 months ago

12-sep exam

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 38 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 38
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company is observing low accuracy while training on the default built-in image classification algorithm in Amazon SageMaker. The Data Science team wants to use an Inception neural network architecture instead of a ResNet architecture.
Which of the following will accomplish this? (Choose two.)

A. Customize the built-in image classification algorithm to use Inception and use this for model training.

B. Create a support case with the SageMaker team to change the default image classification algorithm to Inception.

C. Bundle a Docker container with TensorFlow Estimator loaded with an Inception network and use this for model training.

D. Use custom code in Amazon SageMaker with TensorFlow Estimator to load the model with an Inception network, and use this for model training.

E. Download and apt-get install the inception network code into an Amazon EC2 instance and use this instance as a Jupyter notebook in Amazon SageMaker.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 8:57 a.m.*

Comments

DonaldCMLIN Highly Voted 2 years, 3 months ago
You might be spent a lot of money for ask AWS A.CHANGE built-in image OR B.Create a support case.

The effectial way BOTH RELATIVE TO SageMaker Estimator
C.DOCKER
OR BRING YOUR CODE BY
D.SageMaker with TensorFlow Estimator

THE BEAUTYFUL ANSWER ARE C AND D
upvoted 33 times

Phong Highly Voted 2 years, 3 months ago
I will go for C & D
upvoted 10 times

hug_c0sm0s Most Recent 10 months, 2 weeks ago

Selected Answer: CD

Option A is not possible because the built-in image classification algorithm cannot be customized. Option B is not feasible because it is not possible to change the default image classification algorithm through a support case. Option E is also not a recommended approach because it involves manually installing software on an EC2 instance rather than using the managed services provided by SageMaker.

upvoted 4 times

sqavi 11 months, 1 week ago

Selected Answer: CD

The effectial way BOTH RELATIVE TO SageMaker Estimator
C.DOCKER
OR BRING YOUR CODE BY
D.SageMaker with TensorFlow Estimator

upvoted 2 times

Huy 2 years, 2 months ago

This question ask for 2 ways not a set of actions. So may be confused.

upvoted 1 times

ahquiceno 2 years, 2 months ago

Anwers AD go to: https://docs.aws.amazon.com/sagemaker/latest/dg/docker-containers.html

upvoted 2 times

ybad 2 years, 2 months ago

CD and also A says it but in a more general term....

upvoted 1 times

jaydec 2 years, 3 months ago

https://aws.amazon.com/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/

upvoted 3 times

Antriksh 2 years, 3 months ago

C and D are correct

upvoted 5 times

DonaldCMLIN 2 years, 3 months ago

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/your-algorithms.html

https://aws.amazon.com/tw/blogs/machine-learning/transfer-learning-for-custom-labels-using-a-tensorflow-container-and-bring-your-own-algorithm-in-amazon-sagemaker/
https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/tf.html

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 37 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 37
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist working for an online fashion company wants to build a data ingestion solution for the company's Amazon S3-based data lake.
The Specialist wants to create a set of ingestion mechanisms that will enable future capabilities comprised of:
- Real-time analytics
- Interactive analytics of historical data
- Clickstream analytics
- Product recommendations

Which services should the Specialist use?

A. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for real-time data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations

B. Amazon Athena as the data catalog: Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for near-real-time data insights; Amazon Kinesis Data Firehose for clickstream analytics; AWS Glue to generate personalized product recommendations

C. AWS Glue as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon Kinesis Data Firehose for delivery to Amazon ES for clickstream analytics; Amazon EMR to generate personalized product recommendations

D. Amazon Athena as the data catalog; Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics for historical data insights; Amazon DynamoDB streams for clickstream analytics; AWS Glue to generate personalized product recommendations

Show Suggested Answer

---

by rsimham at *Dec. 9, 2019, 9:19 p.m.*

Comments

rsimham Highly Voted 2 years, 9 months ago
Ans: A seems to be reasonable
upvoted 38 times

cybe001 Highly Voted 2 years, 9 months ago
A looks correct but it is missing for "Interactive analytics of historical data"
upvoted 13 times

ZSun 1 year, 2 months ago

AWS Glue as data catalog, then you can analyze historical data, such as running sql with Athena.

upvoted 1 times

planhanasan 2 years, 8 months ago

Once you insert real-time data to ES, you can see historical data from Kibana dashboard.

upvoted 1 times

eji 2 years, 8 months ago

but C is missing for "real-time analythics"

upvoted 1 times

eji 2 years, 8 months ago

and also C is saying historical data analytics for Kinesis Data analytics which is real-time analytics not historical, so the answer might not C but the answer is A

upvoted 1 times

loict Most Recent 10 months ago

Selected Answer: A

A. YES - Amazon Kinesis Data Analytics is for real-time data insights
B. NO - Amazon Athena has no data catalog
C. NO - Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics is not for historical data insights
D. NO - Amazon Athena has no data catalog

upvoted 4 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: A

Athena can not be used for data catalog, so B and D are wrong. A and C are equals, but it's well known that Kinesis DS and Analytics are used together for real time solutions, which is mentioned in the question / answer, but lack on C.

upvoted 2 times

Valcilio 1 year, 4 months ago

Selected Answer: A

All are bad options, but A can do it.

upvoted 2 times

hug_c0sm0s 1 year, 4 months ago

Selected Answer: A

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to move data between data stores. It can be used as a data catalog to store metadata information about the data in the data lake. Amazon Kinesis Data Streams and Amazon Kinesis Data Analytics can be used together to collect, process, and analyze real-time streaming data. Amazon Kinesis Data Firehose can be used to deliver streaming data to destinations such as Amazon ES for clickstream analytics. Finally, Amazon EMR can be used to run big data frameworks such as Apache Spark and Apache Hadoop to generate personalized product recommendations.

upvoted 2 times

gamaX 2 years, 8 months ago

A or C
https://aws.amazon.com/blogs/big-data/retaining-data-streams-up-to-one-year-with-amazon-kinesis-data-streams/

upvoted 2 times

harmanbirstudy 2 years, 8 months ago

Athena can do Interactive analytics on Historical data, but here its only use is "Athena as the data catalog" and this is the work of Glue data catalog using its crawlers, so it cannot be B or D.
--So its either A or C
-- Now Kinesis data Streams/Analytics is know for real time data analytics but if it is reading from data already stored in S3 using DMS then we can say it is getting historical data.
-- Here I am not very clear if Kinesis part will happen on incoming data before S3 or After data persists to S3 and Kinesis reads it through S3-->DMS--Kinesis data stream -- Kinesis analytics-->Firehose.
But still insights are always on real-time/current data based on historical data trends , so the statement in C "Analytics for historical data insights" is in-correct in general .
Hence ANSWER is :A

upvoted 5 times

ybad 2 years, 8 months ago

A is correct,

for those asking the difference between A and D, D talks about using kinesis stream and data analytics to create historical analysis.... waste of money no?

upvoted 2 times

Th3Dud3 2 years, 8 months ago

Answer = A

upvoted 4 times

C10ud9 2 years, 8 months ago

A it is

upvoted 2 times

roytruong 2 years, 8 months ago

it's A, ES can perform clickstream analytics and EMR can handle spark job recomendation at scale

upvoted 3 times

BigPlums 2 years, 8 months ago

Only C and D mention interactive analytics of historical data.
Glue won't provide personalised recommendation so it is C

upvoted 1 times

BigEv 2 years, 9 months ago

What is the difference between the solution in A or C ????

upvoted 2 times

JayK 2 years, 9 months ago

A is real time data analytics with Kinesis Data analytics and C is saying historical data which is wrong

upvoted 6 times

ComPah 2 years, 9 months ago

Looks like C Amazon ES has KIbana which supports click stream

upvoted 2 times

ComPah 2 years, 9 months ago

A is Correct

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 36 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 36
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist deployed a model that provides product recommendations on a company's website. Initially, the model was performing very well and resulted in customers buying more products on average. However, within the past few months, the Specialist has noticed that the effect of product recommendations has diminished and customers are starting to return to their original habits of spending less. The Specialist is unsure of what happened, as the model has not changed from its initial deployment over a year ago.
Which method should the Specialist try to improve model performance?

A. The model needs to be completely re-engineered because it is unable to handle product inventory changes.

B. The model's hyperparameters should be periodically updated to prevent drift.

C. The model should be periodically retrained from scratch using the original data while adding a regularization term to handle product inventory changes

D. The model should be periodically retrained using the original training data plus new data as product inventory changes.

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 9:14 p.m.*

Comments

rsimham Highly Voted 3 years, 3 months ago

Ans: D

upvoted 27 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

The model performance degradation over time suggests concept drift—the relationship between input features and the target variable has changed. Since product recommendations depend on customer behavior, preferences, and product inventory, periodic retraining with updated data ensures the model adapts to these changes.

Why Periodic Retraining?
Customer preferences evolve:
Buying patterns change over time due to seasons, trends, and external factors.
New products get added, and old ones are discontinued:
The model must learn about new items and stop recommending outdated ones.
The dataset needs to reflect recent trends:
Using new and historical data together ensures the model retains useful past knowledge while learning new patterns.

upvoted 1 times

james2033 10 months, 1 week ago

Selected Answer: D

'retrained using the original training data plus new data'

upvoted 1 times

VR10 10 months, 4 weeks ago

I believe it should be B
1. The model performance has diminished gradually over the past few months, indicating the data distribution may have changed since initial deployment over a year ago. This is a classic sign of concept drift.
2. The model architecture and training procedure have remained unchanged since initial deployment. Updating the hyperparameters is a lighter approach than retraining the model from scratch, and can help prevent further performance deterioration if done periodically to adapt to changes in user preferences and product inventory.

upvoted 1 times

Valcilio 1 year, 10 months ago

Selected Answer: D

D is the answer!

upvoted 1 times

Peeking 2 years, 1 month ago

D is the answer. There has been a data drift resulting from new customer segment visiting the site. So, the model needs to be updated periodically with new data from the website.

upvoted 3 times

apprehensive_scar 2 years, 11 months ago

Selected Answer: D

DDDDD. D :D

upvoted 1 times

cloud_trail 3 years, 2 months ago

Incremental training. D.

upvoted 3 times

gamaX 3 years, 2 months ago

Periodically Re-Fit
D

upvoted 1 times

eji 3 years, 2 months ago

agree with D

upvoted 3 times

C10ud9 3 years, 3 months ago

D is correct

upvoted 3 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 35 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 35
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Science team is designing a dataset repository where it will store a large amount of training data commonly used in its machine learning models. As Data
Scientists may create an arbitrary number of new datasets every day, the solution has to scale automatically and be cost-effective. Also, it must be possible to explore the data using SQL.
Which storage scheme is MOST adapted to this scenario?

A. Store datasets as files in Amazon S3.

B. Store datasets as files in an Amazon EBS volume attached to an Amazon EC2 instance.

C. Store datasets as tables in a multi-node Amazon Redshift cluster.

D. Store datasets as global tables in Amazon DynamoDB.

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 9:04 p.m.*

Comments

rsimham Highly Voted 3 years, 3 months ago
Ans: A (S3) is most cost effective
upvoted 15 times

sonalev419 Highly Voted 3 years, 2 months ago
A : S3 cost effective + athena (not c redshift dont support unstructured data)
upvoted 7 times

JonSno Most Recent 4 months, 3 weeks ago
Selected Answer: A
Amazon S3 (Simple Storage Service) is the best choice because it:

Scales automatically to store an arbitrary number of datasets.
Is cost-effective, as S3 charges only for storage used, unlike provisioned databases.
Supports querying datasets with SQL using Amazon Athena.
Is highly durable (99.999999999% durability) and optimized for large datasets.
How It Works in This Scenario?
Store datasets in S3 as files in Parquet, ORC, or CSV format.

Use AWS Glue Data Catalog to create table metadata.
Query the datasets using Amazon Athena (serverless SQL querying on S3).
Automatically scale without worrying about storage limits.

upvoted 1 times

james2033 10 months, 1 week ago

Selected Answer: A

'cost effective' --> AWS S3

upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: A

A. YES - S3 + Athena/Presto
B. NO - no SQL support
C. NO - expensive to scale
D. NO - DynamoDB is NoSQL

upvoted 1 times

DavidRou 1 year, 4 months ago

Selected Answer: A

AWS S3 + Athena will do it

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: A

The most appropriate storage scheme for this scenario is option A: Store datasets as files in Amazon S3.

Amazon S3 is a highly scalable and cost-effective object storage service that can store a large amount of data. S3 can scale automatically to accommodate a large number of datasets, making it a good option for storing the training data used in machine learning models. Additionally, S3 supports SQL querying through Amazon Athena or Amazon Redshift Spectrum, allowing data scientists to easily explore the data.

upvoted 2 times

harmanbirstudy 3 years, 2 months ago

"store a large amount of training data commonly used in its machine learning models".. well it cannot be anything other than S3. Athena can query S3 cataloged data with SQL commands.
Anwser is A

upvoted 2 times

Stephen_C 3 years, 2 months ago

Amazon Redshift is not cost-effective.

upvoted 1 times

syu31svc 3 years, 3 months ago

I would say C
https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html
"For workloads that require ever-growing storage, managed storage lets you automatically scale your data warehouse storage capacity without adding and paying for additional nodes."

upvoted 3 times

HaiHN 3 years, 3 months ago

Data warehouse is not needed. For exploring data using SQL, you can use Athena

upvoted 5 times

kwangje 3 years, 2 months ago

Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query execution. Most results come back in seconds.

upvoted 1 times

roytruong 3 years, 3 months ago

s3 is right

upvoted 1 times

cybe001 3 years, 3 months ago

A, S3 is most cost effective

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 34 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 34
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.
Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population
How should the Data Scientist correct this issue?

A. Drop all records from the dataset where age has been set to 0.

B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset

C. Drop the age feature from the dataset and train the model using the rest of the features.

D. Use k-means clustering to handle missing features

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 8:58 p.m.*

Comments

rajs Highly Voted 3 years, 9 months ago

Dropping the Age feature is a NOT ATOLL a good idea - as age plays a critical role in this disease as per the question

Dropping 10% of data is NOT a good idea considering the fact that the number of observations is already low.

The Mean or Median are a potential solutions

But the question says that "Disease worsens after age 65 so there is a correlation between age and other symptoms related feature" So that means that using Unsupervised Learning we can make pretty good prediction of "Age"

So the answer is D Use K-Means clustering

upvoted 39 times

L2007 3 years, 9 months ago

https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/
B is correct

upvoted 7 times

Shakespeare 7 months ago

If it was KNN it would be more accurate, but we don't have that option.

upvoted 1 times

vetal Highly Voted 3 years, 9 months ago

Replacing the age with mean or median might bring a bias to the dataset. Use k-means clustering to estimate the missing age based on other features might get better results. Removing 10% available data looks odd. Why not D?

upvoted 20 times

606a82e Most Recent 1 month, 1 week ago

Selected Answer: B

Not D because k-means is used for clustering or grouping, not imputation.

upvoted 1 times

JonSno 4 months, 3 weeks ago

Selected Answer: B

The issue arises from incorrect age values (age = 0) in a dataset where all patients are supposed to be over 65 years old. Since age is an important predictor for the disease's progression, removing or ignoring this feature may negatively impact model performance.

The best approach is imputing missing or incorrect values with a reasonable estimate (e.g., mean or median age of the dataset), ensuring that:
The dataset remains intact without losing valuable patient records.
The model still benefits from age as a feature.
The imputed values are realistic and do not introduce bias.

upvoted 2 times

growe 6 months, 2 weeks ago

Selected Answer: B

Preserves data, maintains model integrity, and corrects anomalies effectively.

upvoted 1 times

imymoco 1 year ago

B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset: This method allows for retaining all patient records while addressing the anomaly. It is a standard approach for dealing with missing or incorrect values in a way that preserves the integrity of the dataset.

B. GPT answer

upvoted 1 times

pn12345 1 year, 2 months ago

B-chatgpt

upvoted 1 times

rookiee1111 1 year, 2 months ago

The question tries to mislead by adding information around the feature correlation. K-means clustering is not meant for imputing data. Hence answer should be B, that would be the right way of handling the missing value.

upvoted 1 times

3eb0542 1 year, 3 months ago

Selected Answer: B

Using k-means clustering to handle missing features is not directly applicable to this scenario. K-means clustering is a method for grouping data points into clusters based on similarity, and it's not typically used for imputing missing values.

upvoted 4 times

kyuhuck 1 year, 5 months ago

Selected Answer: B

add/ comment why? b ? - >replacing the age field value for records with a value of 0 with the mean or median value from the dataset, is generally the best approach among the given options. It allows the preservation of the dataset size and leverages the remaining correct data points, assuming age is a crucial predictor in this context. However, it's vital to perform this imputation carefully to avoid introducing bias. Median is often preferred in this scenario to mitigate the impact of outliers.

upvoted 3 times

kyuhuck 1 year, 5 months ago

Selected Answer: B

The best way to handle the missing values in the patient age feature is to replace them with the mean or median value from the dataset. This is a common technique for imputing missing values that preserves the overall distribution of the data and avoids introducing bias or reducing the sample size. Dropping the records or the feature would result in losing valuable information and reducing the accuracy of the model. Using k-means clustering would not be appropriate for handling missing values in a single feature, as it is a method for grouping similar data points based on multiple

upvoted 2 times

Topg4u 1 year, 5 months ago

mean or median is for outliers so D

upvoted 1 times

endeesa 1 year, 7 months ago

Selected Answer: B

Obviously B, why would you use a clustering algorithm to predict a value? D just doesn't make sense

upvoted 4 times

geoan13 1 year, 8 months ago

B is correct.K-means is unsupervised and used mainly for clustering. KNN would have been more accurate. It can be used to predict a value. since knn is not present i think it is mean median value

upvoted 4 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: B

B is correct or KNN, but dont K means

upvoted 4 times

loict 1 year, 10 months ago

Selected Answer: D

A. NO - unless we want to loose 10% of the data
B. NO - age is predictive, so using the means we would introduce a bias
C. NO - age is predictive
D. YES - better quality than B, it is likely that other physiological values can help predict the age

upvoted 2 times

FloKo 1 year, 11 months ago

Selected Answer: D

k-means should give the best estimation of the age. Using mean would reduce the correlation between outcome and age for the model.

upvoted 1 times

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 33 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 33
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.
The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns.
Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set. However, the prediction results on a test dataset were not satisfactory
Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)

A. Add more deep trees to the random forest to enable the model to learn more features.

B. Include a copy of the samples in the test dataset in the training dataset.

C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.

D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.

E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Show Suggested Answer

by heihei at *Dec. 16, 2019, 2:56 a.m.*

Comments

Phong Highly Voted 3 years, 9 months ago

I think it should be CD
C: because we need a balance dataset
D: The number of positive samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that.
My opinion

upvoted 30 times

Phong Highly Voted 3 years, 9 months ago

I think it should be CD
C: because we need a balance dataset
D: The number of negative samples is large so model tends to predict 0 (negative) for all cases leading to False Negative problem. We should minimize that.

My opinion

upvoted 24 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: CD

Why These Are the Best Choices?
C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.

Balances the dataset by increasing the number of positive samples.
Adding noise prevents overfitting and helps the model generalize better.
Alternative: Use SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic positive examples.
D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.

Since missing a potential paying user (false negative) is more critical than misclassifying a non-paying user, adjusting the cost function to penalize false negatives more will improve recall for paid users.
Methods:
Use weighted loss functions (e.g., weighted cross-entropy).
Adjust class weights in random forest or another algorithm.
Use AUC-ROC or F1-score instead of accuracy for evaluation.

upvoted 2 times

dinITExam 8 months, 2 weeks ago

Think C and D

upvoted 1 times

John_Pongthorn 3 years, 4 months ago

Selected Answer: CD

C,D is correct (percentage of the positive class is key to decide which case we are interested in)
This question, positive class (Pay) is 0.01% as compared to 99.99(not pay) , as a result, we have to pay attention to Pay because if we miss 0.01% out, we didn't get revenue. it is a false negative.

In contrast to these questions, it positive class (Pay) is 40% as compared to negative class (60% not pay), it is avoidable to emphasize on 40% (if model predict as payment but in reality customer neglect), we won't get revenue the amount from false positive)

upvoted 5 times

apprehensive_scar 3 years, 5 months ago

I think is CD

upvoted 1 times

cloud_trail 3 years, 8 months ago

C and D. Hopefully, no one honestly thinks that B is a good answer. Never expose test data to the training set or vice versa. C is right because of the highly imbalanced training set. D is right because you want to minimize false negatives, maximize true positives, maximize recall of the positive class. I'm not sure why anyone's worried about precision in this case.

upvoted 4 times

felbuch 3 years, 8 months ago

CD

The model has 99% accuracy because it's simply predicting that everyone's a negative. Since almost everyone's a negative, it will get almost everyone right.
So we need to penalize the model for predicting that someone is a negative when it is not (i.e. penalize false negatives). So that's D.
Also, it would be really nice to have more positives -- one way to do that is to follow option C.

upvoted 8 times

engomaradel 3 years, 8 months ago

CD 100%

upvoted 1 times

ybad 3 years, 8 months ago

CD

C:imbalance of test (1000 positive, 999000 negative = 0.1% positive) thus C to increase that

D :also to reduce generalizing, since everyone says no, the model would generalize to no, but increasing the penalty of a false negative would reduce generalizing..

upvoted 2 times

Omar_Cascudo 3 years, 8 months ago

It is needed to diminish the FP, because they are player predicted to pay and in reality will not pay. So FP should impact the cost metric more. CE should be the answer.

upvoted 2 times

bidds 3 years, 8 months ago

CD are correct for sure.

upvoted 3 times

hans1234 3 years, 8 months ago

It is C,E... we want to find all paying customers, which are positives, so we have to punish incorrectly finding negatives, which is E

upvoted 2 times

Wira 3 years, 9 months ago

CD

although i am worried about the noise being introduced as it could skew the data nevertheless no better answer is given

upvoted 2 times

aws_razor 3 years, 9 months ago

CD
We need high recall so that we do not miss many Positive cases. In that case we need to have less False Negative(FN) therefore it should have high impact on cost function.

upvoted 3 times

roytruong 3 years, 9 months ago

in my view, CD are answers
C: of course, handle the imbalanced dataset
D: right now, model accuracy is 99%, it means model predict everything is negative leading to FN problem, so we need to minimize it more in cost function

upvoted 3 times

wuha5086 3 years, 9 months ago

CD, FN are valuable players, we should care more on FN

upvoted 8 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 32 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 32
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below.

Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL Databases	Mary Major	NoSQL, Operations, Database	Monday	http://examplecorp.com/nosql_intro_guide.html	407812

Given the dataset, the Specialist wants to convert the Day_Of_Week column to binary values.
What technique should be used to convert this column to binary values?

A. Binarization

B. One-hot encoding

C. Tokenization

D. Normalization transformation

Show Suggested Answer

by omar_bahrain at *March 11, 2021, 7:17 a.m.*

Comments

omar_bahrain Highly Voted 2 years, 8 months ago
I choose b

upvoted 17 times

Juka3lj Highly Voted 2 years, 8 months ago

Correct answer is B.
Example:
Mon | Tue | Wed ....
1 0 0
0 1 0

upvoted 9 times

kaike_reis Most Recent 11 months, 3 weeks ago

Selected Answer: B

Easy Peasy

upvoted 2 times

earthMover 1 year, 1 month ago

Selected Answer: B

Any categorical feature needs to be converted using One Hot Encoding and NOT label encoding.

upvoted 1 times

Tomatoteacher 1 year, 6 months ago

Originally I put A, (believing to be able to format it as (0,1,2,3,4,5,6), or something as it mentioned it to convert the column, but later I realized Binarization is only designed for continuous or numerical data. Even though one-hot encoding will create 6 more columns it is correct. B is correct.

upvoted 1 times

Peeking 1 year, 7 months ago

B
1000000 = Mon
0100000 = Tue
0010000 = Wed
0001000 = Thur
0000100 = Fri
0000010 = Sat
0000001=Sun

upvoted 2 times

benliu1974 1 year, 9 months ago

why not A? 001 010, 011

upvoted 2 times

JDKJDKJDK 8 months, 4 weeks ago

i thought of this at first, but chatgpt's explanation changed my mind

In summary, if the names of days represent nominal categorical variables, one-hot encoding is generally the preferred choice. It maintains distinctiveness, is interpretable, and ensures that each day is clearly represented as a separate binary feature. Binary encoding may be considered for memory efficiency, especially when dealing with a large number of ordinal categories, but it should be used with caution as it introduces an ordinal relationship between categories, which may or may not align with the nature of the data. Ultimately, the choice between the two methods should align with the specific needs of your analysis and the data's characteristics.

upvoted 1 times

apprehensive_scar 2 years, 5 months ago

B is the obvious answer

upvoted 2 times

bitsplease 2 years, 5 months ago

Binary encoding would've been a correct answer but it is not here & Binarization is used for continuous variables. leaving w/ option B

upvoted 1 times

Zhubajie 2 years, 7 months ago

B is wrong. You do not need to one hot encode the variable in random trees. If you do so, you tree must be very deep, which is not efficient. The correct answer is C!

upvoted 1 times

gmnk999 2 years, 3 months ago

"The Specialist want to convert the Day Of Week column in the dataset to binary values." You are misreading the question. The answer is B.

upvoted 4 times

zach288 2 years, 7 months ago

Stop misleading people, the question already asked to convert the data into binary. C is not even remotely close to be correct

upvoted 13 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 31 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 31

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A monitoring service generates 1 TB of scale metrics record data every minute. A Research team performs queries on this data using Amazon Athena. The queries run slowly due to the large volume of data, and the team requires better performance.
How should the records be stored in Amazon S3 to improve query performance?

A. CSV files

B. Parquet files

C. Compressed JSON

D. RecordIO

Show Suggested Answer

by gaku1016 at *Feb. 24, 2020, 3:28 p.m.*

Comments

gaku1016 Highly Voted 2 years, 9 months ago

Answer is B. Athena is best in Parquet format.

upvoted 24 times

emailtorajivk Highly Voted 2 years, 9 months ago

You can improve the performance of your query by compressing, partitioning, or converting your data into columnar formats. Amazon Athena supports open source columnar data formats such as Apache Parquet and Apache ORC. Converting your data into a compressed, columnar format lowers your cost and improves query performance by enabling Athena to scan less data from S3 when executing your query

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

Amazon Athena performs best when querying columnar storage formats like Apache Parquet. Given that 1 TB of data is generated every minute, optimizing storage format is critical for query performance and cost efficiency.

Why Parquet (B) is the Best Choice?
Columnar Storage:

Parquet stores data by columns instead of rows, allowing Athena to scan only the needed columns, reducing the amount of data read.
Compression Efficiency:

Parquet automatically compresses data more efficiently than CSV or JSON.
Smaller file sizes = Faster queries + Lower costs.
Efficient Query Performance:

Parquet supports predicate pushdown, meaning queries can skip irrelevant rows without scanning the entire dataset.
Optimized for Big Data & Athena:

Designed for big data workloads in Athena, Redshift Spectrum, and Presto.
Works well with S3 partitioning to improve query speed.

upvoted 2 times

loict 10 months ago

Selected Answer: B

A. NO - slower
B. YES - Parquet native in Aethena/Presto
C. NO - Compressed JSON
D. NO - no built-in support

upvoted 2 times

teka112233 10 months, 3 weeks ago

Selected Answer: B

according to:
https://dzone.com/articles/how-to-be-a-hero-with-powerful-parquet-google-and
the query run time over parquet file was 6.78 seconds while it was 236 seconds on the same data but stored on csv file which mean that parquet file is 34x faster than csv file

upvoted 1 times

apprehensive_scar 2 years, 5 months ago

Selected Answer: B

B it is

upvoted 3 times

benson2021 2 years, 8 months ago

Answer is B. https://aws.amazon.com/tw/blogs/big-data/top-10-performance-tuning-tips-for-amazon-athena/
But why does this question relate to Machine Learning?

upvoted 3 times

AddiWei 2 years, 5 months ago

Because you must explore data very quickly using SQL in order to run EDA / analyze data for ML purposes. Those explorations can inform on selecting features that can be used for modeling purposes.

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 30 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 30

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

When submitting Amazon SageMaker training jobs using one of the built-in algorithms, which common parameters MUST be specified? (Choose three.)

A. The training channel identifying the location of training data on an Amazon S3 bucket.

B. The validation channel identifying the location of validation data on an Amazon S3 bucket.

C. The IAM role that Amazon SageMaker can assume to perform tasks on behalf of the users.

D. Hyperparameters in a JSON array as documented for the algorithm used.

E. The Amazon EC2 instance class specifying whether training will be run using CPU or GPU.

F. The output path specifying where on an Amazon S3 bucket the trained model will persist.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 7:05 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 9 months ago

THE ANSWER SHOUD BE CEF
IAM ROLE, INSTANCE TYPE, OUTPUT PATH

upvoted 29 times

hamimelon 2 years, 6 months ago

Why not A? You don't need to tell Sagemaker where the training data is located?

upvoted 3 times

ZSun 2 years, 2 months ago

You need to specify the InputDataConfig, but it does not need to be "S3"
I think the reason why A and B are wrong, not because data location is not required, but because it doesn't need to be S3, it can be Amazon S3, EFS, or FSx location

upvoted 1 times

HaiHN 3 years, 8 months ago

Should be C, E, F

From the SageMaker notebook example:
https://github.com/aws/amazon-sagemaker-examples/blob/master/introduction_to_amazon_algorithms/semantic_segmentation_pascalvoc/semantic_segmentation_pascalvoc.ipynb
Create the sagemaker estimator object.

ss_model = sagemaker.estimator.Estimator(training_image,
role,
train_instance_count = 1,
train_instance_type = 'ml.p3.2xlarge',
train_volume_size = 50,
train_max_run = 360000,
output_path = s3_output_location,
base_job_name = 'ss-notebook-demo',
sagemaker_session = sess)

upvoted 12 times

uninit 2 years, 5 months ago

It says InstanceClass - CPU/GPU in the question, not InstanceType

upvoted 6 times

mirik 2 years ago

instance type has default value.

upvoted 3 times

VB Highly Voted 3 years, 9 months ago

From here https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html .. the only "Required: Yes" attributes are:
1. AlgorithmSpecification (in this TrainingInputMode is Required - i.e. File or Pipe)
2. OutputDataConfig (in this S3OutputPath is Required - where the model artifacts are stored)
3. ResourceConfig (in this EC2 InstanceType and VolumeSizeInGB are required)
4. RoleArn (..The Amazon Resource Name (ARN) of an IAM role that Amazon SageMaker can assume to perform tasks on your behalf...the caller of this API must have the iam:PassRole permission.)
5. StoppingCondition
6. TrainingJobName (The name of the training job. The name must be unique within an AWS Region in an AWS account.)

From the given options in the questions.. we have 2, 3, and 4 above. so, the answer is CEF.

upvoted 27 times

cloud_trail 3 years, 8 months ago

This is the best explanation that CEF is the right answer, IMO. The document at that url is very informative. It also specifically states that InputDataConfig is NOT required. Having said that, I have no idea how the model will train if it doesn't know where to find the training data, but that is what the document says. If someone can explain that, I'd like to hear the explanation.

upvoted 7 times

cloud_trail 3 years, 8 months ago

If I see this question on the actual exam, I'm going with AEF. The model absolutely must know where the training data is. I have seen other documentation that does confirm that you need the location of the input data, the compute instance and location to output the model artifacts.

upvoted 3 times

CloudGuru_ZA 3 years, 8 months ago

but you also need to specify the service role sagemaker should use otherwise it will not be able to perform actions on your behalf like provisioning the training instances.

upvoted 2 times

rafaelo 3 years, 6 months ago

Perfect explanation. It is CEF

upvoted 1 times

JK1977 2 years, 1 month ago

The question is asking about built in algorithms. It should be ADE. See https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/API_CreateTrainingJob.html

upvoted 1 times

OAmine 1 year, 9 months ago

for "3. ResourceConfig", only VolumeSizeInGB is required. So, it's not about the instance type.
Check: https://docs.aws.amazon.com/zh_tw/sagemaker/latest/APIReference/API_ResourceConfig.html

upvoted 1 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: ACF

Reason:
When submitting Amazon SageMaker training jobs using built-in algorithms, the following parameters must be specified:

Training Data Location (A)

SageMaker requires the training dataset's location in Amazon S3.

Provided as a channel input in the training job.
IAM Role (C)

SageMaker needs IAM permissions to access data from S3 and execute tasks on behalf of the user.
Model Output Path (F)

Specifies the S3 bucket location where the trained model artifacts will be stored.

upvoted 2 times

AbhayD 5 months, 3 weeks ago

Selected Answer: ACF

Instance type is required but not specific class CPU/GPU. Sagamkaer can handle that.

upvoted 1 times

MultiCloudIronMan 8 months, 2 weeks ago

Selected Answer: ACF

These parameters ensure that the training job has access to the necessary data, permissions, and storage locations to function correctly.

upvoted 1 times

MultiCloudIronMan 8 months, 2 weeks ago

Selected Answer: ACF

Options B, D, and E are important but not always mandatory for every training job. For example, validation data (Option B) is not always required, and hyperparameters (Option D) and instance types (Option E) can have default values or be optional depending on the specific algorithm and setup.

upvoted 1 times

amlgeek 9 months, 1 week ago

```
import boto3
import sagemaker

sess = sagemaker.Session()

# Example for the linear learner
linear = sagemaker.estimator.Estimator(
container,
role, # role (c)
instance_count=1,
instance_type="ml.c4.xlarge", # instance type (e)
output_path=output_location, # output path (f)
sagemaker_session=sess,
)
```

upvoted 1 times

kiran15789 10 months, 3 weeks ago

Selected Answer: CEF

Going with cef

upvoted 1 times

ML_2 11 months ago

Selected Answer: CEF

ANSWER IS CEF
Here from Amazon docs
InputDataConfig
An array of Channel objects. Each channel is a named input source. InputDataConfig describes the input data and its location.
Required: No

OutputDataConfig
Specifies the path to the S3 location where you want to store model artifacts. SageMaker creates subfolders for the artifacts.
Required: Yes

ResourceConfig - Identifies the resources, ML compute instances, and ML storage volumes to deploy for model training. In distributed training, you specify more than one instance.
Required: Yes

upvoted 1 times

RathanKalluri 1 year ago

CEF
https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html#API_CreateTrainingJob_RequestParameters

upvoted 1 times

ninomfr64 1 year ago

Based on https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

Required parameters are:
- AlgorithmSpecification (registry path of the Docker image with the training algorithm)

- OutputDataConfig (path to the S3 location where you want to store model artifacts)
- ResourceConfig (resources, including the ML compute instances and ML storage volumes, to use for model training)
- RoleArn
- StoppingCondition (time limit for training job)
- TrainingJobName

Thus, the answer is: C E F
wording for option E is inaccurate "EC2 instance class specifying whether training will be run using CPU or GPU" but they do it on purpose

upvoted 1 times

rookiee1111 1 year, 2 months ago

Selected Answer: ACF

The input channel and output channel are mandatory, as the training job needs to know where to get the input data from and where to publish the model artifact. IAM role is also needed, for AWS services. others are not mandatory, validation channel is not mandatory for instance in case of unsupervised learning, likewise hyper params can be be auto tuned for as well as the ec2 instance types can be default ones that will be picked

upvoted 2 times

Denise123 1 year, 2 months ago

As they narrowed it to S3, A is incorrect BUT when submitting Amazon SageMaker training jobs using one of the built-in algorithms, it is a MUST to identify the location of training data. While Amazon S3 is commonly used for storing training data, other sources like Docker containers, DynamoDB, or local disks of training instances can also be used. Therefore, specifying the location of training data is essential for SageMaker to know where to access the data during training.

So the right answer is CEF for me for this case... However if A was saying identify the location of training data, I think option A would be included in the MUST parameter.

upvoted 1 times

sachin80 1 year, 2 months ago

InputDataConffig is optional in create_training_job.Please check thte parameters that are required.
So answer is CEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

upvoted 1 times

sachin80 1 year, 2 months ago

InputDataConffig is optional in create_training_job.Please check thte parameters that are required.
So answer is SEF: https://docs.aws.amazon.com/sagemaker/latest/APIReference/API_CreateTrainingJob.html

upvoted 1 times

vkbajoria 1 year, 3 months ago

Selected Answer: CEF

Input is required only when calling Fit method. When initializing the Estimator, we do not need input

upvoted 1 times

rav009 1 year, 3 months ago

Selected Answer: ACF

I open the sagemaker and tested. A C F
B is not needed for non-supervised algorithm.

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 29 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 29
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

An insurance company is developing a new device for vehicles that uses a camera to observe drivers' behavior and alert them when they appear distracted. The company created approximately 10,000 training images in a controlled environment that a Machine Learning Specialist will use to train and evaluate machine learning models.
During the model evaluation, the Specialist notices that the training error rate diminishes faster as the number of epochs increases and the model is not accurately inferring on the unseen test images.
Which of the following should be used to resolve this issue? (Choose two.)

A. Add vanishing gradient to the model.

B. Perform data augmentation on the training data.

C. Make the neural network architecture complex.

D. Use gradient checking in the model.

E. Add L2 regularization to the model.

Show Suggested Answer

by vetal at *Dec. 9, 2019, 3:04 p.m.*

Comments

vetal Highly Voted 3 years, 3 months ago
The model must have been overfitted. Regularization helps to solve the overfitting problem in machine learning (as well as data augmentation). Correct answers should be BE.
upvoted 36 times

rajs 3 years, 3 months ago
Agreed 100%
upvoted 5 times

jasonsunbao 3 years, 3 months ago
agree on BE
upvoted 3 times

benson2021 Highly Voted 3 years, 2 months ago
Answer: BE

https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html
5 techniques to prevent overfitting:
1. Simplifying the model
2. Early stopping
3. Use data argumentation
4. Use regularization
5. Use dropouts

upvoted 15 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: BE

he issue described suggests that the model is overfitting to the training data:

Training error decreases quickly, meaning the model is learning the training set very well.
Poor performance on unseen test data, indicating overfitting.
To resolve overfitting, the Machine Learning Specialist should:

Perform Data Augmentation (B)

Expands the training dataset artificially by applying transformations (e.g., rotations, flips, brightness changes, cropping).
Helps the model generalize better by exposing it to more diverse variations of the same class.

Add L2 Regularization (E)

Also known as weight decay, it penalizes large weights, preventing the model from memorizing the training data.
Encourages simpler models, which reduces variance and improves generalization.

upvoted 2 times

delfoxete 11 months, 1 week ago

Selected Answer: BE

agreed with vetal

upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: BE

A. NO - vanishing gradient is somebody bad they might happen and prevent convergence, we don't want that or something we can add explicitly. it is a result of the learning
B. YES - we have a overfitting problem so more training examples will help
C. NO - we already have good accuracy on the training set
D. NO - gradient checking is to find bugs in model implementation
E. YES - we have a overfitting problem

upvoted 2 times

John_Pongthorn 2 years, 10 months ago

B. Perform data augmentation on the training data. (it should add validation data as well)
data should be distributed among train validation and test.

upvoted 1 times

KM226 3 years ago

Selected Answer: BE

Answer B&E looks good

upvoted 2 times

engomaradel 3 years, 2 months ago

B & E is the correct ans

upvoted 1 times

roytruong 3 years, 2 months ago

BE is exact

upvoted 3 times

stamarpadar 3 years, 2 months ago

BE are the correct answers

upvoted 4 times

VB 3 years, 2 months ago

Looks like B and D are correct.. For D -> https://www.youtube.com/watch?v=P6EtCVrvYPU

upvoted 3 times

C10ud9 3 years, 2 months ago

gradient checking doesn't resolve the issue, but adding it will confirm / deny the issue. So, it helps to validate the issue but not resolve. I would say B, E are correct

upvoted 3 times

VB 3 years, 3 months ago

L2 regularization tries to reduce the possibility of overfitting by keeping the values of the weights and biases small.

upvoted 3 times

hughhughhugh 3 years, 3 months ago

why not becuase of vanishing gradient?

upvoted 1 times

lt626 2 years ago

Vanishing gradients are a problem when training a NN. Answer A mentions that the solution should be to add that, which is not possible. Correct solution is BE.
https://www.kdnuggets.com/2022/02/vanishing-gradient-problem.html

upvoted 1 times

PRC 3 years, 3 months ago

This is L2 Regularization....Do you think this is the right answer?

upvoted 1 times

WWODIN 3 years, 3 months ago

agree BE

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 28 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 28

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A company is using Amazon Polly to translate plaintext documents to speech for automated company announcements. However, company acronyms are being mispronounced in the current documents.
How should a Machine Learning Specialist address this issue for future documents?

A. Convert current documents to SSML with pronunciation tags.

B. Create an appropriate pronunciation lexicon.

C. Output speech marks to guide in pronunciation.

D. Use Amazon Lex to preprocess the text files for pronunciation

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 5:56 a.m.*

Comments

VB Highly Voted 3 years, 3 months ago

SSML is specific to that particular doument, like W3C an be pronounced as "World Wide Web Consortium" using _{W3C} in that specific document and when you create a new document, you need to format again. But with LEXICONS, you can upload a lexicon file once and ALL the FUTURE documents can just have W3C and that will be pronounced as "World Wide Web Consortium".. so answer is B, because the question asks for "future" documents.

upvoted 44 times

khchan123 1 year, 2 months ago

The correct answer is B, as explained by VB.

upvoted 1 times

cloud_trail 3 years, 2 months ago

For the exact reason you state, the correct answer is A. For every different document, a particular acronym may mean something different so you must have a solution that is document-specific.

upvoted 3 times

LeoD 6 months, 3 weeks ago

As the question stated "address this issue FOR FUTURE DOCUMENTS". B addresses for future. A only address the issue in a case-by-case manner.

upvoted 1 times

ovokpus 2 years, 6 months ago

It is the same business, so the acronyms are not expected to change from document to document

upvoted 3 times

VR10 10 months, 4 weeks ago

absolutely, B is the correct choice.

upvoted 1 times

Madwyn 3 years, 2 months ago

A.The document section for "Pronouncing Acronyms and Abbreviations".
Source: https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html

upvoted 3 times

cybe001 Highly Voted 3 years, 3 months ago

I think the answer is B.
https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html
https://www.smashingmagazine.com/2019/08/text-to-speech-aws/

upvoted 18 times

Tony_1406 2 years, 3 months ago

Lifted from the above link - "Your text might include an acronym, such as W3C. You can use a lexicon to define an alias for the word W3C so that it is read in the full, expanded form (World Wide Web Consortium)."
Clearly this is the same use case.

upvoted 1 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

Explanation:
Amazon Polly sometimes mispronounces acronyms because it reads them as regular words. The best way to correct mispronunciations in future documents is to create a pronunciation lexicon. This allows you to define how specific words, acronyms, or abbreviations should be pronounced.

How to Use a Pronunciation Lexicon in Amazon Polly?
Define the correct pronunciation of acronyms in a Lexicon XML file.
Use phonetic notation (e.g., IPA or Speech Synthesis Markup Language (SSML) Phoneme tags).
Upload the lexicon to Polly via the AWS Management Console or AWS SDK.
Reference the lexicon in Polly API requests.

upvoted 3 times

VasuKVK 9 months, 1 week ago

Answer : B

https://aws.amazon.com/blogs/machine-learning/customize-pronunciation-using-lexicons-in-amazon-polly/

Use <phoneme> SSML tag which is great for inserting one-off customizations or testing purposes. We recommend using Lexicon to create a consistent set of pronunciations for frequently used words across your organization. This enables your content writers to spend time on writing instead of the tedious task of adding phonetic pronunciations in the script repetitively.

upvoted 1 times

WTSppl 10 months, 3 weeks ago

SSML supports phonetic pronunciation. Seems to me A is correct too.
https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html#phoneme-tag

upvoted 1 times

phdykd 1 year ago

B IS ANSWER

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: B

B is the correct, A hardan cixdi debil?

upvoted 1 times

DavidRou 1 year, 4 months ago

This issue can be faced with both methods described in A and B. Though the answer A refers to the "current" document while the question regards "future" documents, so I think the right answer is B.

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: B

Letter B is correct to ensure that acronyms or terms are pronounced correctly. Letter A works, but look at the catch: It's asked for future documents, but it mentions converting only current ones to SSML format, while future ones would be in plaintext.

upvoted 2 times

ADVIT 1 year, 6 months ago

Company using plaintext and Future document means plaintext!
So only Custom Lexicon will help.

upvoted 1 times

soonmo 1 year, 7 months ago

Selected Answer: B

https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html
this explains acronym.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: B

I believe it should be lexicon. Can you share how you tag the correct answer?

upvoted 1 times

Sylzys 1 year, 10 months ago

Selected Answer: B

Key here being "for future documents", answer should be B as SSML is for a specific document only

upvoted 3 times

Chelseajcole 1 year, 10 months ago

this should be multiple choice question which answer is a AND b

upvoted 1 times

bakarys 1 year, 10 months ago

Selected Answer: B

response B

A pronunciation lexicon is a list of words and their correct phonetic pronunciation that can be used to improve the accuracy of text-to-speech conversion. In this case, the Machine Learning Specialist can create a custom lexicon for the company's acronyms and upload it to Amazon Polly. This will ensure that the acronyms are pronounced correctly in the future announcements.

upvoted 2 times

SK27 2 years, 1 month ago

Selected Answer: B

Should be B

upvoted 2 times

masoa3b 2 years, 2 months ago

Selected Answer: B

With Amazon Polly's custom lexicons or vocabularies, you can modify the pronunciation of particular words, such as company names, acronyms, foreign words, etc. To customize these pronunciations, you upload an XML file with lexical entries. {rpmimcoatopm ;exocpms enable you to customize the pronunciation of words. Amazon Polly provides API operations that you can use to store lexicons in an AWS region. Those lexicons are then specific to that particular region.

References:
https://docs.aws.amazon.com/polly/latest/dg/managing-lexicons.html
https://aws.amazon.com/blogs/machine-learning/create-accessible-training-with-initiafy-and-amazon-polly/

upvoted 3 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 27 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 27
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.
Here is an example from the dataset:
"The quck BROWN FOX jumps over the lazy dog.`
Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Choose three.)

A. Perform part-of-speech tagging and keep the action verb and the nouns only.

B. Normalize all words by making the sentence lowercase.

C. Remove stop words using an English stopword dictionary.

D. Correct the typography on "quck" to "quick.ג€

E. One-hot encode all words in the sentence.

F. Tokenize the sentence into words.

Show Suggested Answer

by cybe001 at *Jan. 12, 2020, 2:54 p.m.*

Comments

ozan11 Highly Voted 2 years, 9 months ago
B C F should be correct.
upvoted 35 times

BigEv Highly Voted 2 years, 9 months ago
I will select B, C, F
1- Apply words stemming and lemmatization
2- Remove Stop words
3- Tokensize the sentences

https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925
upvoted 26 times

Togy Most Recent 4 months ago

Selected Answer: BDF

B. Normalize all words by making the sentence lowercase:

Word2Vec treats words as distinct entities. If you don't convert everything to lowercase, "The" and "the" will be considered different words, which is generally not what you want. Lowercasing ensures consistency.
D. Correct the typography on "quck" to "quick":

Misspellings need to be corrected. Word2Vec learns embeddings based on the words it encounters. If "quck" remains, it will be treated as a separate word from "quick," and you'll lose the relationship between them. Correcting typos is crucial for data quality.
F. Tokenize the sentence into words:

Tokenization is the process of breaking down the sentence into individual words (or tokens). Word2Vec operates on individual words, so you need to split the sentence into its constituent parts. This is a fundamental step in NLP.

upvoted 1 times

JonSno 4 months, 3 weeks ago

Selected Answer: BDF

While C - is debatable - not always necessary to remove stop words in Word2Vec - as sometimes the stop words do provide context
==========

For Word2Vec training, data preprocessing is essential to ensure that words are correctly represented, consistent, and free from unnecessary noise. The key steps are:

Lowercasing the text (B)

Word embeddings treat "FOX" and "fox" as different words. To avoid redundancy, lowercasing the text ensures consistency.
Correcting typos (D)

"quck" should be corrected to "quick" to prevent incorrect word representations in Word2Vec. Misspelled words can create meaningless embeddings.
Tokenizing the sentence into words (F)

Word2Vec operates at the word level, so breaking the sentence into individual tokens (words) is necessary.

upvoted 2 times

loict 10 months ago

Selected Answer: BCF

A. NO - word2vec works on raw data
B. YES - case here is not significant
C. YES - will help reduce dimensionality
D. NO - word2vec will do it by itself
E. NO - One-hot encoding is for classification
F. YES - word2vec takes tokens as input

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: BCF

Data need to be tokenized and cleaned!

upvoted 2 times

Aninina 1 year, 6 months ago

Selected Answer: BCF

B, C F is the correct

upvoted 2 times

SophieSu 2 years, 8 months ago

BCF correct. D is not correct (Pay attention to “in a repeatable manner” in the question.)

upvoted 2 times

cloud_trail 2 years, 8 months ago

B/C/F. D should not be performed because spell check is a subjective thing. You don't know for sure what the word was supposed to be if you have a typo.

upvoted 2 times

harmanbirstudy 2 years, 8 months ago

I saw this exact question on "whizlabs" practice exam and correct options were B/C/F

upvoted 1 times

GeeBeeEl 2 years, 8 months ago

https://towardsdatascience.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets-13445eebd281
Data Preparation — Define corpus, clean, normalise and tokenise words
To begin, we start with the following corpus:
“natural language processing and machine learning is fun and exciting”
For simplicity, we have chosen a sentence without punctuation and capitalization. Also, we did not remove stop words “and” and “is”.

In reality, text data are unstructured and can be “dirty”. Cleaning them will involve steps such as
o removing stop words,
o removing punctuations,
o convert text to lowercase (actually depends on your use-case),
o replacing digits, etc.
o After preprocessing, we then move on to tokenising the corpus
Answer: B, C, F

upvoted 8 times

cnethers 2 years, 8 months ago

BCF is 100% correct

upvoted 2 times

Antriksh 2 years, 8 months ago

Correct answers are B, C and F

upvoted 2 times

TuanAnh 2 years, 9 months ago

The correct answer is B, C and F
A: POS tagging has nothing to do with word2vec
D: fixing "quck" to "quick" only works for that specific word
F: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here

upvoted 4 times

TuanAnh 2 years, 9 months ago

sorry E: word2vec can use CBOW or skipgram, so no need to have one-hot decoding here

upvoted 4 times

PRC 2 years, 9 months ago

BCF is correct

upvoted 2 times

AKT 2 years, 9 months ago

B, C F correct

upvoted 2 times

Phong 2 years, 9 months ago

B, C, and F are correct answers. I have done this question many times in many practice tests.

upvoted 12 times

tap123 2 years, 9 months ago

B, C, F are my choice. D is also possible but not as widely used as others.

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 26 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 26
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist kicks off a hyperparameter tuning job for a tree-based ensemble model using Amazon SageMaker with Area Under the ROC Curve
(AUC) as the objective metric. This workflow will eventually be deployed in a pipeline that retrains and tunes hyperparameters each night to model click-through on data that goes stale every 24 hours.
With the goal of decreasing the amount of time it takes to train these models, and ultimately to decrease costs, the Specialist wants to reconfigure the input hyperparameter range(s).
Which visualization will accomplish this?

A. A histogram showing whether the most important input feature is Gaussian.

B. A scatter plot with points colored by target variable that uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the large number of input variables in an easier-to-read dimension.

C. A scatter plot showing the performance of the objective metric over each training iteration.

D. A scatter plot showing the correlation between maximum tree depth and the objective metric.

Show Suggested Answer

---

by heihei at *Dec. 13, 2019, 8:57 a.m.*

Comments

cloud_trail Highly Voted 3 years, 9 months ago

This is a very tricky question. The idea is to reconfigure the ranges of the hyperparameters. A refers to a feature, not a hyperparameter. A is out. C refers to training the model, not optimizing the range of hyperparameters. C is out. Now it gets tricky. D will let you find determine what the approximately best tree depth is. That's good. That's what you're trying to do but it's only one of many hyperparameters. It's the best choice so far. B is tricky. t-SNE does help you visualize multidimensional data but option B refers to input variables, not hyperparameters. For this very tricky question, I would do with D. It's the only one that accomplishes the task of limiting the range of a hyperparameter, even if it is only one of them.

upvoted 50 times

cnethers 3 years, 9 months ago

It's good to see someone keeping a thoughtful and curious mind to this question. I too have the same conclusion, not an easy question.

upvoted 3 times

ovokpus 3 years ago

But, how do you optimize hyperparameters without training experiments? That is why C is the best option. You get a value for each unique combination of hyperparameters.

upvoted 1 times

Dr_Kiko 3 years, 8 months ago

B is also wrong as t-SNE picture is not actionable - good visual but ... that's it. try pictures here https://lvdmaaten.github.io/tsne/

upvoted 1 times

AddiWei 3 years, 4 months ago

When you are tuning hyperparameters you are literally training multiple models and searching for the best ones.

upvoted 2 times

heihei Highly Voted 3 years, 9 months ago

B doesn't make sense
I think it's D

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

The goal is to reduce training time and costs by optimizing the hyperparameter tuning process. In tree-based ensemble models (e.g., XGBoost, Random Forest, or Gradient Boosting), tree depth is one of the most influential hyperparameters affecting:

Model complexity: Deeper trees increase complexity but may lead to overfitting.
Training time: More depth means more splits, significantly increasing computation.
Performance (AUC score in this case): There is typically an optimal depth that balances underfitting and overfitting.
A scatter plot showing the correlation between tree depth and the AUC metric will allow the ML Specialist to:

Identify whether increasing depth leads to diminishing returns.
Choose an optimal tree depth that balances performance with training efficiency.
Reduce the search space of hyperparameters, speeding up tuning and lowering costs.

upvoted 1 times

ninomfr64 1 year ago

A. No, doesn't help to set/reduce hyperparameter value/range
B. No, honestly this is gibberish to me
C. No, doesn't help to reduce hyperparameter value range
D. YES, this help me understand how to set max tree depth hyperparameter

upvoted 1 times

VR10 1 year, 4 months ago

Option C.
See it is doing a scatter plot on the metric for each iteration.
Each iteration is running with a certain set of hyper parameters.
So if I plot this. and I find which iteration has the best metric, I could simply pick up those set of hyperparameters.
D will only led to the tuning of maximum tree depth.
I am not sure which option would satisfy the goal to decrease cost but just looking at maximum tree depth doesnt seem right to me. It might be a way to just look at the tree depth and tune just that parameter and since you are only tuning 1 paramter, it may be cheaper, but would that lead to a usable model?
I think it should be option C.

upvoted 1 times

Regu7 1 year, 5 months ago

On what basis the correct answers are provided in this platform? Are they assuming this is the correct answer or it is taken from somewhere ?

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: D

D IS THE CORRECT

upvoted 1 times

Rejju 1 year, 10 months ago

Selected Answer: C

Option D, can also be useful in hyperparameter tuning for tree-based ensemble models, especially if the maximum tree depth is one of the hyperparameters you want to optimize.
However, when the goal is to decrease training time and costs by reconfiguring input hyperparameter ranges, a scatter plot showing the performance of the objective metric over each training iteration (Option C) is generally more directly related to the hyperparameter tuning process. It helps you track how the model's performance changes during hyperparameter tuning, which is critical for making decisions about which hyperparameter ranges to explore further.
Option D is valuable for understanding the relationship between maximum tree depth and the objective metric, but it might not provide as comprehensive insights into the overall hyperparameter tuning process compared to Option C.

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: D

A. NO - it is about data discovery

B. NO - it is about data discovery
C. MIGHT - (NO) is a training iteration the overnight training the question is referring to ? (YES) Or each HPO training within each night ?
D. YES - the less ambiguous answer

upvoted 1 times

DavidRou 1 year, 10 months ago

I think that C should be the right answer. The specialist can monitor how the model works by changing hyperparameters' values in each training iteration.

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: D

Option D

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: D

A and B are wrong, because is totally out of question context. C is for monitoring a model, it doesn't help to change your HP range. D is the only answer that applies to the question.

upvoted 3 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

CKS1210 2 years ago

Selected Answer: C

By plotting the performance of the objective metric (AUC) over each training iteration, the Specialist can analyze how different hyperparameter configurations affect the model's performance. This visualization helps in understanding which hyperparameter combinations lead to better results and allows the Specialist to identify areas of improvement.

upvoted 1 times

mirik 2 years, 1 month ago

D: By analyzing this relationship, the Specialist can adjust the range of maximum tree depth values used during hyperparameter tuning to decrease training time and costs.

upvoted 1 times

earthMover 2 years, 1 month ago

Selected Answer: D

D Seems like the best answer. When answer is considered correct who is making that call an is there any justification provided for us to learn from?

upvoted 2 times

Valcilio 2 years, 4 months ago

Selected Answer: D

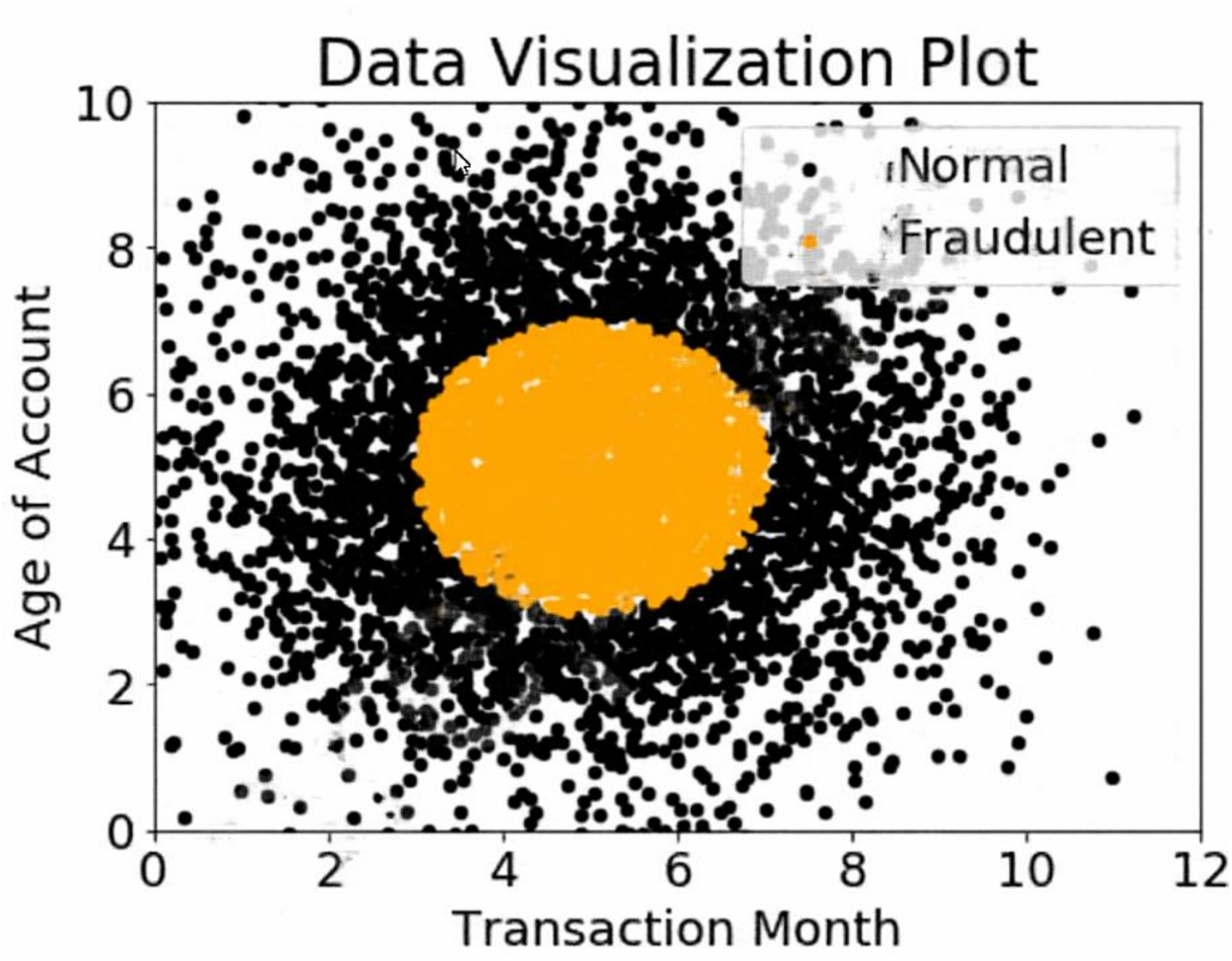
It's about parameters, not about dimensionality.

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 25 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 25
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST recall with respect to the fraudulent class?

A. Decision tree

B. Linear support vector machine (SVM)

C. Naive Bayesian classifier

D. Single Perceptron with sigmoidal activation function

Show Suggested Answer

by cnethers at *Feb. 3, 2021, 12:23 p.m.*

Comments

E_aws Highly Voted 3 years, 8 months ago

C is the correct answer because gaussian naive Bayes can do this nicely.

upvoted 12 times

E_aws 3 years, 8 months ago

of course it doesn't mention the gaussian here and refers to naive bayes in general, but I'm still positive with C.

upvoted 1 times

blubb Highly Voted 3 years, 8 months ago

Answer should be A:.
B: LINEAR SVM is a linear classifier
-> All of these have a linear decision boundary (so it's just a line y = mx+b). This leads to a bad recall and so A must be the right choice.

upvoted 9 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: A

Decision Tree (Best Choice) ✅

Highly flexible: Can capture non-linear decision boundaries, making it effective when the class distribution is not linearly separable.
Maximizes recall: A decision tree can prioritize minimizing false negatives by adjusting its splits.
Handles imbalanced classes well using class weighting or pruning techniques.

upvoted 3 times

MVAS 5 months ago

Selected Answer: C

Gaussian naive Bayes is correct one

upvoted 1 times

MintTeaClarity 7 months, 4 weeks ago

Selected Answer: A

A non-linear problem would be a case where linear classifiers, such as naive Bayes, would not be suitable since the classes are not linearly separable. In such a scenario, non-linear classifiers (e.g.,instance-based nearest neighbour classifiers) should be preferred.

upvoted 1 times

egorkrash 8 months, 3 weeks ago

Selected Answer: A

decision tree can effectively maximize the recall by drawing a square (3 <= month <= 7, 3 <= age <= 7)

upvoted 2 times

MultiCloudIronMan 8 months, 3 weeks ago

Selected Answer: A

Option C. Naive Bayesian classifier is not the best choice for achieving the highest recall for the fraudulent class because it makes strong assumptions about the independence of features. In many real-world scenarios, especially with complex data like user behavior, these assumptions do not hold true, which can lead to suboptimal performance.

In contrast, a Decision tree (Option A) can handle feature interactions and is more flexible in capturing the relationships between features, making it more effective in identifying fraudulent behavior and achieving higher recall

upvoted 1 times

ML_2 11 months ago

Selected Answer: A

Answer in my opinion is A
A Decision Tree Classifier can handle complex decision boundaries and does not assume any particular distribution of data. It is well-suited for cases like this where the decision boundary is non-linear, as seen with the clear separation between the normal and fraudulent transactions.

A Naive Bayesian classifier, on the other hand, assumes independence among features and typically performs better when data is normally

distributed, which might not be the case here given the data's clustering pattern.

upvoted 1 times

ninomfr64 1 year ago

Selected Answer: C

From Claude 3 Haiku:

A. NO, decision trees may struggle to capture the linear separability of the classes.
B. NO, Linear SVM may not be able to fully exploit the class separation due to its linear decision boundary.
C. YES, The Naive Bayesian classifier tends to perform well in situations where the classes are linearly separable. This model requires the features are independent and this is the case
D. The single Perceptron with a sigmoidal activation function may not be able to capture the complex class distributions as effectively as the Naive Bayesian classifier.

upvoted 1 times

GrumpyApple 7 months, 3 weeks ago

Funny that if you ask Haiku to explain its reason step by step, it will chose A instead of C

```
Based on the information provided, the model that is likely to have the highest recall with respect to the fraudulent class is the **Decision Tree (Most Voted)**.
```

upvoted 1 times

iambasspaul 1 year, 2 months ago

Selected Answer: C

Answer by Claude3:
In contrast, the Decision Tree (A) and Linear SVM (B) models are generally more robust to overfitting and can achieve a better balance between recall and precision, but they may not necessarily have the highest recall for the minority class.
Considering the importance of maximizing recall for the fraudulent class in this use case, the Naive Bayesian Classifier (C) could be a valid choice, although it may come with the trade-off of lower precision and potentially higher false positive rates.

upvoted 1 times

rav009 1 year, 4 months ago

highest recall.
So A

upvoted 1 times

notbother123 1 year, 4 months ago

Selected Answer: A

Only A (DT) is non-linear among the mentioned algorithms.

upvoted 1 times

kyuhuck 1 year, 5 months ago

Selected Answer: A

Given the visualized data, the Decision tree (Option A) is likely the best model to achieve the highest recall for the fraudulent class. It can handle complex patterns and create rules that are more suited for clustered and potentially non-linearly separable classes. Recall is a measure of a model's ability to capture all actual positives, and a decision tree can be tuned to prioritize capturing more of the fraudulent cases at the expense of making more false-positive errors on the normal cases.

upvoted 1 times

phdykd 1 year, 6 months ago

if it was highest precision:
Given these considerations, the best model for precision would likely be a Support Vector Machine with a non-linear kernel, such as the RBF (Radial Basis Function) kernel. This model can tightly fit the boundary around the fraudulent class, minimizing the inclusion of normal transactions in the fraudulent prediction space, and thus potentially achieving high precision. Precision is sensitive to the false positives, and the flexibility of SVMs with non-linear kernels to create a tight and precise boundary can help to minimize these.

upvoted 1 times

phdykd 1 year, 6 months ago

GPT 4 Answer is Decision Tree.
Considering the goal is to achieve the highest recall for the fraudulent class, which means we aim to capture as many fraudulent cases as possible even if it means getting more false positives, a Decision Tree would likely be the best option. This is because it can adapt to the complex shape of the class distribution and encapsulate the majority of the fraudulent class within its decision boundaries. Recall is a measure of a model's ability to capture all actual positives, and the decision tree's complex boundary setting capabilities make it well-suited for maximizing recall in this case.

upvoted 2 times

taustin2 1 year, 7 months ago

Selected Answer: A

I'm going with A. As pointed out in this article, Naive Bayes performs poorly with non-linear classification problems. The picture shows a case where the classes are not linearly separable. Decision Tree will probably give better results.
https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 24 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 24
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist trained a regression model, but the first iteration needs optimizing. The Specialist needs to understand whether the model is more frequently overestimating or underestimating the target.
What option can the Specialist use to determine whether it is overestimating or underestimating the target value?

A. Root Mean Square Error (RMSE)

B. Residual plots

C. Area under the curve

D. Confusion matrix

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 4:20 a.m.*

Comments

vetal Highly Voted 2 years, 3 months ago

RMSE says about the error value but not the sign of error. The question is to find whether the model overestimates or underestimates - I guess residual plots clearly show that

answer B

upvoted 37 times

rsimham Highly Voted 2 years, 3 months ago

Answer is B. Residual plot distribution indicates over or under-estimations

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

A residual plot helps determine whether a regression model is overestimating or underestimating the target value.

Residual = Actual Value - Predicted Value
Positive residual → The model underestimated the target.
Negative residual → The model overestimated the target.
By plotting residuals, the Machine Learning Specialist can see patterns that indicate bias:

More positive residuals → The model is underestimating.

More negative residuals → The model is overestimating.
Randomly scattered residuals around zero → The model is well-calibrated.

upvoted 2 times

Valcilio 10 months, 1 week ago

Selected Answer: B

Residual plots shows mistake by mistake!

upvoted 1 times

vetaal 1 year, 12 months ago

Selected Answer: B

B - Residual plots it is - https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

upvoted 4 times

felbuch 2 years, 2 months ago

Residual Plots (B).
AUC and Confusion Matrices are used for classification problems, not regression.
And RMSE does not tell us if the target is being over or underestimated, because residuals are squared! So we actually have to look at the residuals themselves. And that's B.

upvoted 7 times

cnethers 2 years, 2 months ago

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

1) Squaring the residuals.
2) Finding the average of the residuals.
3) Taking the square root of the result.

upvoted 3 times

cnethers 2 years, 2 months ago

Residual Plots (B). would have to be my answer

upvoted 1 times

Thai_Xuan 2 years, 2 months ago

residual plot
https://docs.aws.amazon.com/machine-learning/latest/dg/regression-model-insights.html

upvoted 5 times

syu31svc 2 years, 2 months ago

https://stattrek.com/statistics/dictionary.aspx?definition=residual%20plot#:~:text=A%20residual%20plot%20is%20a,nonlinear%20model%20is%20more%20appropriate.

Answer is B

upvoted 2 times

Antriksh 2 years, 2 months ago

without a second thought residual plot

upvoted 2 times

qururu 2 years, 2 months ago

The answer is B. Refer to Exercise 7.2.1.A
https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_OpenIntro_Statistics_(Diez_et_al)./07%3A_Introduction_to_Linear_Regression/7.02%3A_Line_Fitting%2C_Residuals%2C_and_Correlation

upvoted 1 times

C10ud9 2 years, 2 months ago

Residual plot it is Option B

upvoted 1 times

roytruong 2 years, 2 months ago

Residual plot

upvoted 2 times

deep_n 2 years, 2 months ago

B is the correct answer!!!!
RMSE has the S in it that is square... that vanishes the above below factor of the prediction.
Answers C and D are for other type of problems

upvoted 4 times

swagy 2 years, 3 months ago

It should be B. The residual plot will be give whether the target value is overestimated or underestimated.
upvoted 1 times

Jayraam 2 years, 3 months ago
Answer is C.

https://www.youtube.com/watch?v=MrjWcywVEiU
upvoted 2 times

ExamTaker123456789 2 years, 2 months ago
Answer is B.
Your vid shows a technique that is useful for defining integrals and has NOTHING to do linear regression. Also, it over-/underestimates the area under the curve, NOT the target value.
upvoted 2 times

cloud_trail 2 years, 2 months ago
Good grief, AUC is used for classification not regression.
upvoted 1 times

PRC 2 years, 3 months ago
B..Residual helps to find out whether the model is underestimating or overestimating
upvoted 3 times

AKT 2 years, 3 months ago
answer is B
upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 23 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 23

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is building a convolutional neural network (CNN) that will classify 10 types of animals. The Specialist has built a series of layers in a neural network that will take an input image of an animal, pass it through a series of convolutional and pooling layers, and then finally pass it through a dense and fully connected layer with 10 nodes. The Specialist would like to get an output from the neural network that is a probability distribution of how likely it is that the input image belongs to each of the 10 classes.
Which function will produce the desired output?

A. Dropout

B. Smooth L1 loss

C. Softmax

D. Rectified linear units (ReLU)

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 4:17 a.m.*

Comments

DonaldCMLIN Highly Voted 2 years, 9 months ago

C might be much suitable
softmax is to turn numbers into probabilities.

https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d

upvoted 30 times

rsimham Highly Voted 2 years, 9 months ago

C is right. Softmax function is used for multi-class predictoins

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: C

In a multiclass classification problem (such as classifying an image into one of 10 animal categories), the model should output a probability distribution over the classes. The Softmax function achieves this by:

Taking the raw scores (logits) from the final dense layer (10 nodes, one per class).
Exponentializing each score and normalizing them so they sum to 1, effectively turning them into probabilities.

upvoted 1 times

loict 10 months ago

Selected Answer: C

A. NO - Dropout is to prevent overfitting
B. NO - L1 regularization is to prevent overfitting
C. YES - Softmax will give probabilities for each class
D. NO - Rectified linear units (ReLU) is an activation function

upvoted 2 times

DavidRou 10 months ago

Softmax is the correct answer.

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: C

Multiclassification with probabilities is about softmax!

upvoted 1 times

vbal 1 year, 6 months ago

Softmax is for probability distribution

upvoted 1 times

technoguy 2 years, 8 months ago

it should be C. Softmax
Softmax converts outputs to Probabilites of each classification

upvoted 3 times

omar8024 2 years, 8 months ago

absolutely C

upvoted 1 times

takahirokoyama 2 years, 8 months ago

Absolute C.

upvoted 4 times

cloud_trail 2 years, 8 months ago

This is as easy a question as you will likely see on the exam, Everyone has the right answer here.

upvoted 3 times

felbuch 2 years, 8 months ago

C --> Softmax.

Let's go over the alternatives:
A. Dropout --> Not really a function, but rather a method to avoid overfitting. It consists of dropping some neurons during the training process, so that the performance of our algorithm does not become very dependent on any single neuron.
B. Smooth L1 loss --> It's a loss function, thus a function to be minimized by the entire neural network. It's not an activation function.
C. Softmax --> This is the traditional function used for multi-class classification problems (such as classifying an animal into one of 10 categories)
D. Rectified linear units (ReLU) --> This activation function is often used on the first and intermediate (hidden) layers, not on the final layer. In any case, it wouldn't make sense to use it for classification because its values can exceed 1 (and probabilities can't)

upvoted 11 times

MOMoez 2 years, 8 months ago

C, Softmax is the best suitable answer
Ref: The softmax function, also known as softargmax[1]:184 or normalized exponential function,[2]:198 is a generalization of the logistic function to multiple dimensions. It is used in multinomial logistic regression and is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes, based on Luce's choice axiom.

upvoted 2 times

ybad 2 years, 8 months ago

You guys are right, the answer is C since it automatically provides the output with a confidence interval...

Relu could be used as well but it needs to be coded in to provide the probabilities

https://medium.com/@himanshuxd/activation-functions-sigmoid-relu-leaky-relu-and-softmax-basics-for-neural-networks-and-deep-8d9c70eed91e

upvoted 1 times

yeetusdeleetus 2 years, 8 months ago

Definitely C

upvoted 1 times

bidds 2 years, 9 months ago

Definitely softmax.

upvoted 1 times

hans1234 2 years, 9 months ago

Are you sure it is C?`
The output should be "[the probability that] the input image belongs to each of the 10 classes." And not the most likely class with the highest probability, which would be the result of softmax layer.

upvoted 1 times

hans1234 2 years, 9 months ago

Yes, softmax returns indeed a vector of probabilities.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 22 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 22
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A retail chain has been ingesting purchasing records from its network of 20,000 stores to Amazon S3 using Amazon Kinesis Data Firehose. To support training an improved machine learning model, training records will require new but simple transformations, and some attributes will be combined. The model needs to be retrained daily.
Given the large number of stores and the legacy data ingestion, which change will require the LEAST amount of development effort?

A. Require that the stores to switch to capturing their data locally on AWS Storage Gateway for loading into Amazon S3, then use AWS Glue to do the transformation.

B. Deploy an Amazon EMR cluster running Apache Spark with the transformation logic, and have the cluster run each day on the accumulating records in Amazon S3, outputting new/transformed records to Amazon S3.

C. Spin up a fleet of Amazon EC2 instances with the transformation logic, have them transform the data records accumulating on Amazon S3, and output the transformed records to Amazon S3.

D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.

Show Suggested Answer

by vetal at *Dec. 6, 2019, 11:59 a.m.*

Comments

cybe001 Highly Voted 2 years, 9 months ago
D is correct. Question has '"simple transformations, and some attributes will be combined" and Least development effort. Kinesis analytics can get data from Firehose, transform and write to S3
https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html
upvoted 49 times

mawsman 2 years, 9 months ago
Best explanation here, kudos.
upvoted 4 times

kakalotka 2 years, 8 months ago
I can't find any information that indicate Kinesis data analytics taking data from firehose
upvoted 2 times

Huy Highly Voted 2 years, 8 months ago

The best way to transform data is before it arrives to S3 so D should be best answer. But D is not completed. It should have another Firehose to deliver results to S3.

upvoted 9 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

D. Insert an Amazon Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream that transforms raw record attributes into simple transformed values using SQL.
Explanation:
Since the data is already flowing through Amazon Kinesis Data Firehose, the least development effort solution is to use Amazon Kinesis Data Analytics, which supports SQL-based transformations on streaming data without requiring new infrastructure.

Why is this the best choice?
No major architectural changes – Data continues flowing from stores into Kinesis Data Firehose and then to Amazon S3.
Simple SQL transformations – Since the changes are simple (e.g., attribute combinations), SQL is sufficient.
Low operational overhead – No need to manage clusters or instances.
Real-time processing – Transformed records immediately enter Amazon S3 for training.

upvoted 2 times

CKS1210 1 year ago

Ans is D
Amazon Kinesis Data Analytics provides a serverless option for real-time data processing using SQL queries. In this case, by inserting a Kinesis Data Analytics stream downstream of the Kinesis Data Firehose stream, the retail chain can easily perform the required simple transformations on the ingested purchasing records.

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: D

The best answer is to use a lambda, but the letter D can do it very good too in the absence of the lambda option.

upvoted 2 times

cloud_trail 2 years, 8 months ago

I go with D. A tough question, though. And C are definitely out. They key to the question is that it does not say that the transformed data needs to be stored again in S3. It just needs to be sent to the model for training after being transformed. So a Kinesis Data Analytics stream is appropriate to do the transformation.

upvoted 1 times

harmanbirstudy 2 years, 8 months ago

Legacy data -- Firehose -- Kinesis Analytics -- S3.This happens in near real time before the data ends up in S3.
--Legacy data -- Firehose -- S3 is already happening (mentioned in first line in question), adding Kinesis Data Analytics to do simple transformation joins using SQL on the incoming data is the LEAST amount of work needed.
Kinesis Data analytics can write o S3. here is the AWS link with working example.Even Though Udemy tutorial said it cannot write directly to S3 :)
.

https://docs.aws.amazon.com/kinesisanalytics/latest/java/examples-s3.html

upvoted 6 times

gamaX 2 years, 8 months ago

It seems that LEAST developmnet effort:
https://aws.amazon.com/fr/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/
and GRETAST development effort:
https://aws.amazon.com/fr/blogs/big-data/optimizing-downstream-data-processing-with-amazon-kinesis-data-firehose-and-amazon-emr-running-apache-spark/

upvoted 1 times

HaiHN 2 years, 8 months ago

It's D

https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/

"In some scenarios, you may need to enhance your streaming data with additional information, before you perform your SQL analysis. Kinesis Analytics gives you the ability to use data from Amazon S3 in your Kinesis Analytics application, using the Reference Data feature. However, you cannot use other data sources from within your SQL query."

upvoted 1 times

h_sahu 2 years, 8 months ago

I believe, kinesis should be used only in case of live data stream and this is not the case here. So as per me D shouldn't be the answer. I think A should be the answer as AWS storage gateway is something which is used alongwith on premise applications to move data to s3. Then glue can be used to transform the data.

upvoted 1 times

cloud_trail 2 years, 8 months ago

With option A, you would be changing the legacy data ingestion, a huge development effort. Remember, you're talking about 20,000 stores.

upvoted 2 times

hans1234 2 years, 8 months ago

It is D.

upvoted 1 times

dikers 2 years, 9 months ago

I think the answer is D, because require the LEAST amount of development effort.

upvoted 1 times

roytruong 2 years, 9 months ago

it's D, kinesis analytic can easily connect with firehose

upvoted 2 times

dreemswang 2 years, 9 months ago

why not A. it seems good to me

upvoted 2 times

ExamTaker123456789 2 years, 9 months ago

"require stores to capture data locally using S3 gateway" - for 20k stores this creates a HUUUGE operational overhead and development effort, definitely wrong

upvoted 3 times

PRC 2 years, 9 months ago

D is correct...rest all need some kind of manual intervention as well as they are not simple..Firehose allows transformation as well as moving into S3

upvoted 6 times

devsean 2 years, 9 months ago

I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.

upvoted 3 times

hailiang 2 years, 8 months ago

Its D, because with KDA you can transform the data with SQL while with EMR you need to write code, considering the requirement of "least development effort", so D

upvoted 3 times

devsean 2 years, 9 months ago

I think the answer is B. D would be correct if they didn't want to transform the legacy data from before the switch, but it seems like they do. Choosing D would mean that you'd have to use an EC2 instance or something else to transform the legacy data along with adding the Kinesis data analytics functionality. Also, there is no real-time requirement so daily transformation is fine.

upvoted 7 times

ADVIT 1 year ago

"LEAST amount of development effort" , EMR is no complicated to LEAST

upvoted 1 times

ZSun 1 year, 2 months ago

If the question is "least cost" then B, but the question is "least develope effort, then you want to keep original architeture. I agree that for daily ETL instead of real-time, and large dataset, B is better option.

upvoted 1 times

HaiHN 2 years, 8 months ago

You can use Lambda instead of EC2. So D should be OK.
https://aws.amazon.com/blogs/big-data/preprocessing-data-in-amazon-kinesis-analytics-with-aws-lambda/

upvoted 1 times

am7 2 years, 9 months ago

can be B

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 21 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 21

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.
Which services are integrated with Amazon SageMaker to track this information? (Choose two.)

A. AWS CloudTrail

B. AWS Health

C. AWS Trusted Advisor

D. Amazon CloudWatch

E. AWS Config

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 5:41 a.m.*

Comments

rsimham Highly Voted 3 years, 3 months ago

AD is correct

upvoted 27 times

eji Highly Voted 3 years, 2 months ago

CloudTrail is use to track scientist how ofthe they deploy a model
CloudWatch for monitoring GPU and CPU
so answer is A & D

upvoted 10 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: AD

To monitor SageMaker model deployments, track resource utilization, and log errors, the Machine Learning Specialist should use:

AWS CloudTrail – Tracks API activity, such as:

Model deployments (e.g., CreateModel, CreateEndpoint)

Notebook access and actions
SageMaker job executions
Amazon CloudWatch – Monitors and logs operational metrics, such as:

CPU & GPU utilization of SageMaker endpoints
Invocation errors and latencies
Custom metrics from deployed models
Logs from training jobs and inference endpoints (via CloudWatch Logs)

upvoted 2 times

VR10 10 months, 4 weeks ago

I think AWS Config is still not the service designed to track how often Data Scientists are deploying models, nor does it track operational performance metrics like GPU and CPU utilization or the invocation errors of SageMaker endpoints.
and AWS CloudTrail continues to be the service that will track and record user activity and API usage, which includes deploying models in Amazon SageMaker.
So the answers are still A and D - CloudTrail and CloudWatch.

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: AD

AD is correct

upvoted 2 times

loict 1 year, 4 months ago

Selected Answer: AC

A. YES - to track deployements
B. NO - AWS Health is to track AWS Cloud itself (eg. is a zone down ?)
C. NO - AWS Trusted Advisor to give recommendations on infra
D. YES - for errors
E. AWS Config

upvoted 2 times

DavidRou 1 year, 4 months ago

I also believe that A and D are correct. Can someone please explain to me the main differences between CloudWatch and CloudTrail? I find the documentation a bit confusing about it

upvoted 1 times

CKS1210 1 year, 6 months ago

Option E AWS Config to record all resource types, then the new resources will be automatically recorded in your account.
Option A CloudTrail is use to track scientist how of the they deploy a model
Option D CloudWatch for monitoring GPU and CPU

upvoted 1 times

joe3232 1 year, 11 months ago

Log Amazon Sagemaker API Calls with AWS CloudTrail - https://docs.aws.amazon.com/sagemaker/latest/dg/logging-using-cloudtrail.html

upvoted 2 times

f4bi4n 2 years, 6 months ago

I wouldn't be so sure about CloudTrail, AWS Configs also tracks Sagemaker and the resource "AWS::Sagemaker::Model"

upvoted 1 times

f4bi4n 2 years, 6 months ago

just seen, this was release 4 days ago...
https://aws.amazon.com/about-aws/whats-new/2022/06/aws-config-15-new-resource-types/

upvoted 2 times

yogesh1 2 years, 11 months ago

Selected Answer: AD

A&D
CloudWatch and ClouTrail

upvoted 3 times

hess 3 years ago

AD Are Correct.

upvoted 1 times

Urban_Life 3 years, 2 months ago

absolutely

upvoted 1 times

roytruong 3 years, 2 months ago

cloudtrail and cloudwatch, no thinking

upvoted 6 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 20 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 20

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.
What should the Specialist do to meet these requirements?

A. Create one-hot word encoding vectors.

B. Produce a set of synonyms for every word using Amazon Mechanical Turk.

C. Create word embedding vectors that store edit distance with every other word.

D. Download word embeddings pre-trained on a large corpus.

Show Suggested Answer

by vetal at *Dec. 6, 2019, 11:47 a.m.*

Comments

JayK Highly Voted 3 years, 9 months ago

the solution is word embedding. As it is a interactive online dictionary, we need pre-trained word embedding thus the answer is D. In addition, there is no mention that the online dictonary is unique and does not have a pre-trained word embedding.
Thus I strongly feel the answer is D

upvoted 31 times

cybe001 Highly Voted 3 years, 9 months ago

D is correct. It is not a specialized dictionary so use the existing word corpus to train the model

upvoted 16 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

D. Download word embeddings pre-trained on a large corpus.
Reason :
For a nearest neighbor model that finds words used in similar contexts, word embeddings are the best choice. Pre-trained word embeddings capture semantic relationships and contextual similarity between words based on a large text corpus (e.g., Wikipedia, Common Crawl).

The Specialist should:

Use pre-trained word embeddings like Word2Vec, GloVe, or FastText.
Load the embeddings into the model for efficient similarity comparisons.

Use a nearest neighbor search algorithm (e.g., FAISS, k-d tree, Annoy) to quickly find similar words.

upvoted 2 times

Ajose0 9 months, 3 weeks ago

Selected Answer: D

D. Download word embeddings pre-trained on a large corpus.

Word embeddings are a type of dense representation of words, which encode semantic meaning in a vector form. These embeddings are typically pre-trained on a large corpus of text data, such as a large set of books, news articles, or web pages, and capture the context in which words are used. Word embeddings can be used as features for a nearest neighbor model, which can be used to find words used in similar contexts.

Downloading pre-trained word embeddings is a good way to get started quickly and leverage the strengths of these representations, which have been optimized on a large amount of data. This is likely to result in more accurate and reliable features than other options like one-hot encoding, edit distance, or using Amazon Mechanical Turk to produce synonyms.

upvoted 6 times

loict 9 months, 3 weeks ago

Selected Answer: D

A. NO - one-hot encoding is a very early featurization stage
B. NO - we don't want human labelling
C. NO - too costly to do from scratch
D. YES - leverage exiting training; the word embeddings will provide vectors than be used to measure distance in the downstream nearest neighbor model

upvoted 3 times

game_changer 9 months, 3 weeks ago

Selected Answer: D

Pre-trained word embeddings, such as Word2Vec, GloVe, or FastText, capture the semantic and contextual meaning of words based on a large corpus of text data. By downloading pre-trained word embeddings, the Specialist can leverage the semantic relationships between words to provide meaningful word features for the nearest neighbor model powering the widget. Utilizing pre-trained word embeddings allows the model to understand and display words used in similar contexts effectively.

upvoted 2 times

game_changer 9 months, 3 weeks ago

Selected Answer: D

A. One-hot word encoding vectors: These vectors represent words by marking them as present or absent in a fixed-length binary vector. However, they don't capture relationships between words or their meanings.

B. Producing synonyms: This would involve generating similar words for each word manually, which could be time-consuming and might not cover all possible contexts.

C. Word embedding vectors based on edit distance: This approach focuses on how similar words are in terms of their spelling or characters, not necessarily their meaning or context in sentences.

D. Downloading pre-trained word embeddings: These are vectors that represent words based on their contextual usage in a large dataset, capturing relationships between words and their meanings.

upvoted 5 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: D

correct D ay tupoy

upvoted 1 times

sonoluminescence 1 year, 8 months ago

Selected Answer: D

words that are used in similar contexts will have vectors that are close in the embedding space

upvoted 1 times

Mickey321 1 year, 11 months ago

Selected Answer: D

D is correct

upvoted 1 times

DavidRou 1 year, 11 months ago

I also believe that D is the correct answer. No reason to create word embeddings from scratch

upvoted 1 times

ortamina 2 years ago

Selected Answer: D

1. One-hot encoding will blow up the feature space - it is not recommended for a high cardinality problem domain.

2. One still needs to train the word features on large bodies of text to map context to each word

upvoted 1 times

Shailendraa 2 years, 10 months ago

12-sep exam

upvoted 1 times

helpaws 2 years, 10 months ago

Selected Answer: D

DDDDDDDDDDDD

upvoted 3 times

engomaradel 3 years, 8 months ago

D for sure

upvoted 2 times

yeetusdeleetus 3 years, 8 months ago

Definitely D.

upvoted 3 times

weslleylc 3 years, 8 months ago

A)It requires that document text be cleaned and prepared such that each word is one-hot encoded.
Ref:https://machinelearningmastery.com/what-are-word-embeddings/

upvoted 1 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 19 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 19
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is building a logistic regression model that will predict whether or not a person will order a pizza. The Specialist is trying to build the optimal model with an ideal classification threshold.
What model evaluation technique should the Specialist use to understand how different classification thresholds will impact the model's performance?

A. Receiver operating characteristic (ROC) curve

B. Misclassification rate

C. Root Mean Square Error (RMSE)

D. L1 norm

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 5:35 a.m.*

Comments

rsimham Highly Voted 3 years, 9 months ago

Ans. A is correct

upvoted 20 times

AKT Highly Voted 3 years, 9 months ago

Answer is A.
An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds

upvoted 9 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: A

Explanation:
The ROC curve is the best technique to evaluate how different classification thresholds impact the model's performance. It plots True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values.

Why the ROC Curve?
Logistic regression outputs probabilities, and we need to select a classification threshold to decide between "order pizza" (1) and "not order pizza" (0).

Changing the threshold impacts the trade-off between sensitivity (recall) and specificity.
The ROC curve helps visualize this trade-off and select the best threshold based on the business goal (e.g., maximizing recall vs. minimizing false positives).
The Area Under the ROC Curve (AUC-ROC) is a useful metric to measure the model's discrimination ability.

upvoted 2 times

GeeBeeEl 9 months, 3 weeks ago

A is indeed correct see https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
• True Positive Rate
• False Positive Rate
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:
TPR = TP/TP+FN
False Positive Rate (FPR) is defined as follows:
FPR = FP/FP+TN

upvoted 4 times

Mickey321 9 months, 3 weeks ago

Selected Answer: A

The reason for this choice is that a ROC curve is a graphical plot that illustrates the performance of a binary classifier across different values of the classification threshold1. A ROC curve plots the true positive rate (TPR) or sensitivity against the false positive rate (FPR) or 1-specificity for various threshold values2. The TPR is the proportion of positive instances that are correctly classified, while the FPR is the proportion of negative instances that are incorrectly classified.

upvoted 2 times

Valcilio 2 years, 4 months ago

Selected Answer: A

ROC curve is for defining the threshold.

upvoted 2 times

spamicho 3 years, 8 months ago

A surely

upvoted 1 times

syu31svc 3 years, 8 months ago

Question is about classification so confusion matrix would come into mind; A is the answer

upvoted 1 times

hans1234 3 years, 9 months ago

It is A.

upvoted 1 times

roytruong 3 years, 9 months ago

obviously A

upvoted 1 times

bitiyaha 3 years, 9 months ago

Root Mean Square Error (RMSE) Ans. c

upvoted 1 times

bzhao 3 years, 9 months ago

I think RMSE is for regression model

upvoted 5 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 18 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 18

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is packaging a custom ResNet model into a Docker container so the company can leverage Amazon SageMaker for training. The
Specialist is using Amazon EC2 P3 instances to train the model and needs to properly configure the Docker container to leverage the NVIDIA GPUs.
What does the Specialist need to do?

A. Bundle the NVIDIA drivers with the Docker image.

B. Build the Docker container to be NVIDIA-Docker compatible.

C. Organize the Docker container's file structure to execute on GPU instances.

D. Set the GPU flag in the Amazon SageMaker CreateTrainingJob request body.

Show Suggested Answer

---

by heihei at *Dec. 6, 2019, 3:12 a.m.*

Comments

vetal Highly Voted 3 years, 9 months ago

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf
page 55:
If you plan to use GPU devices, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers. Don't bundle NVIDIA drivers with the image. For more information about nvidia-docker, see NVIDIA/nvidia-docker.

So the answer is B

upvoted 52 times

devsean 3 years, 9 months ago

Yeah, it's B. But the page in the developer guide is page number 201 (209 in pdf). Second bullet point at the top.

upvoted 6 times

AKT Highly Voted 3 years, 9 months ago

Answer is B. below is from AWS documentation,
If you plan to use GPU devices for model training, make sure that your containers are nvidia-docker compatible. Only the CUDA toolkit should be included on containers; don't bundle NVIDIA drivers with the image.

https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html

upvoted 14 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: B

When using Amazon SageMaker with GPU-based EC2 instances (e.g., P3 instances), you must ensure that your custom Docker container can leverage NVIDIA GPUs. NVIDIA-Docker (now part of Docker with nvidia-container-runtime) allows containers to access GPU resources without needing to bundle NVIDIA drivers inside the container.

To make a custom Docker container GPU-compatible, the Machine Learning Specialist should:

Use NVIDIA CUDA and cuDNN in the Dockerfile.
Ensure the container is built using the NVIDIA Container Toolkit (nvidia-docker).
Use nvidia-container-runtime as the runtime.

upvoted 2 times

Ajose0 9 months, 3 weeks ago

Selected Answer: B

To leverage the NVIDIA GPUs on Amazon EC2 P3 instances for training with Amazon SageMaker, the Docker container must be built to be compatible with NVIDIA-Docker.

NVIDIA-Docker is a wrapper around Docker that makes it easier to use GPUs in containers by providing GPU-aware functionality.

To build a Docker container that is compatible with NVIDIA-Docker, the Specialist should install the NVIDIA GPU drivers in the Docker container and install the NVIDIA-Docker runtime on the EC2 instances.

upvoted 1 times

bakarys 9 months, 3 weeks ago

Selected Answer: B

NVIDIA-Docker is a Docker container runtime plugin that allows the Docker container to access the GPU resources on the host machine. By building the Docker container to be NVIDIA-Docker compatible, the Docker container will have access to the NVIDIA GPU resources on the Amazon EC2 P3 instances, allowing for accelerated training of the ResNet model.

upvoted 1 times

Mickey321 9 months, 3 weeks ago

Selected Answer: B

The reason for this choice is that NVIDIA-Docker is a tool that enables GPU-accelerated containers by automatically configuring the container runtime to use NVIDIA GPUs1. NVIDIA-Docker allows you to build and run Docker containers that can fully access the GPUs on your host system. This way, you can run GPU-intensive applications, such as deep learning frameworks, inside containers without any performance loss or compatibility issues.

upvoted 1 times

loict 9 months, 3 weeks ago

Selected Answer: B

A. NO - the drivers are not necessary (https://docs.aws.amazon.com/sagemaker/latest/dg/your-algorithms-training-algo-dockerfile.html)
B. YES - it is about using the CUDA library, need to use proper base image (https://medium.com/@jgleeee/building-docker-images-that-require-nvidia-runtime-environment-1a23035a3a58)
C. NO - file structure irrelavant to GPU
D. NO - SageMaker config, irrelevant to Docker

upvoted 2 times

6ff83cb 1 year, 4 months ago

Selected Answer: B

https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-dg.pdf
page 55

upvoted 1 times

iambasspaul 1 year, 2 months ago

page 570
On a GPU instance, the image is run with the --gpus option. Only the CUDA toolkit should be included in the image not the NVIDIA drivers. For more information, see NVIDIA User Guide.

upvoted 1 times

Crypt0zknight 1 year, 9 months ago

Answer B
Load the CUDA toolkit only, not the drivers. Ref GPU section : https://docs.aws.amazon.com/sagemaker/latest/dg/studio-byoi-specs.html

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

jackzhao 2 years, 4 months ago

B is correct!

upvoted 1 times

Sylzys 2 years, 4 months ago

Selected Answer: B

As per aws documentation, answer is B, and A is even explicitly not recommended

upvoted 1 times

Sorrybutnotsorry 3 years, 6 months ago

Selected Answer: B

As referred in other comments ans is B

upvoted 1 times

hussamS 3 years, 6 months ago

Selected Answer: B

ANS B
As mentioned byi other users

upvoted 1 times

sachin80 3 years, 8 months ago

As per me answer is B

upvoted 1 times

konradL 3 years, 9 months ago

The answer is for sure B - as mentioned by others. And this is clearly stated in the docs

upvoted 1 times

takahirokoyama 3 years, 9 months ago

Ans. is B.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 17 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 17
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

An employee found a video clip with audio on a company's social media feed. The language used in the video is Spanish. English is the employee's first language, and they do not understand Spanish. The employee wants to do a sentiment analysis.
What combination of services is the MOST efficient to accomplish the task?

A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend

B. Amazon Transcribe, Amazon Comprehend, and Amazon SageMaker seq2seq

C. Amazon Transcribe, Amazon Translate, and Amazon SageMaker Neural Topic Model (NTM)

D. Amazon Transcribe, Amazon Translate and Amazon SageMaker BlazingText

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 3:24 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 9 months ago
the MOST efficient means to you don't need to coding, building infra
All of sevices are manage by AWS is good,
Transcribe, Amazon Translate, and Amazon Comprehend

Answer is A
upvoted 44 times

WWODIN 3 years, 9 months ago
Agree, Answer is A
upvoted 9 times

Pg690 Highly Voted 2 years, 7 months ago
A is not 100% correct. You don't need to translate Spanish. Amazon Comprehend supports Spanish.
upvoted 8 times

cpal012 2 years, 3 months ago
Arguably, you still need a translation since the person doesn't speak Spanish.
upvoted 2 times

tonton3 2 years ago

I think there is no need to use Amazon translate because sometimes the translation is not accurate.
It means some information gets lost.

upvoted 1 times

kaike_reis 1 year, 11 months ago

Given the question, I believe that is necessary: look at the enphase of not understanding spanish. besides that, even with some information lost, you will at least understand something.

upvoted 1 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: A

A. Amazon Transcribe, Amazon Translate, and Amazon Comprehend
Explanation of the Process:
Amazon Transcribe – Converts the Spanish audio in the video into text.
Amazon Translate – Translates the Spanish text to English.
Amazon Comprehend – Performs sentiment analysis on the translated English text.

upvoted 3 times

ADVIT 9 months, 3 weeks ago

It's A:
1.Amazon Transcribe - to convert Spanish speech to Spanish text.
2.Amazon Translate - to translate Spanish text to English text
3.Amazon Comprehend - to analyze text for sentiments

upvoted 2 times

loict 9 months, 3 weeks ago

Selected Answer: A

A. YES - Comprehend is supervised so user must understand through Translate
B. NO - seq2seq is for generation and not classification
C. NO - Amazon SageMaker Neural Topic Model is unsupervised topic extraction, will not give sentiment against user-defined classes
D. NO - BlazingText is word2vec, does not give sentiment classes

upvoted 1 times

Mickey321 9 months, 3 weeks ago

Selected Answer: A

It's A:
1.Amazon Transcribe - to convert Spanish speech to Spanish text.
2.Amazon Translate - to translate Spanish text to English text
3.Amazon Comprehend - to analyze text for sentiments

upvoted 2 times

DavidRou 1 year, 11 months ago

It's A 100%

upvoted 1 times

CKS1210 2 years ago

Transcribe: Speech to text
Translate: Any language to any language
Comprehend: offers a range of capabilities for extracting insights and meaning from unstructured text data. Ex: Sentiment analysis, entity recognition, KeyPhrase Extraction, Language Detection, Document Classification

upvoted 1 times

soonmo 2 years, 1 month ago

absolutely need STT(transcribe), translation(translate), and sentimental analysis(comprehend)

upvoted 1 times

gnolam 2 years, 10 months ago

Selected Answer: A

A - confirmed by ACG

upvoted 2 times

KM226 3 years, 6 months ago

Selected Answer: A

I agree that the answer is A

upvoted 1 times

in4976 3 years, 7 months ago

Selected Answer: A

answer is a

upvoted 1 times

Dr_Kiko 3 years, 8 months ago

A; D is wrong because The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms.

upvoted 1 times

harmanbirstudy 3 years, 8 months ago

The Question/Anwser is not poorly as someone mentioned.
--Even though Comprehend can do the analysis directly on Spanish (no need of translate) but if comprehend does analysis and the resulting words are still in spanish , it will no help the employee as he doesnt know Spanish.So the transalate after transcribe will help Employee understand what is being analyzed by Comprehend in next step.
So read the question carefully before jumping to conclusions. it will save you an Exam :)

upvoted 1 times

senseikimoji 3 years, 8 months ago

I don't get this question. Comprehend supports Spanish natively. There is no need for Translate, and translate would actually reduce effectiveness of sentimental analysis. However, BCD are all invalid choices.

upvoted 3 times

ybad 3 years, 8 months ago

A
because Comprehend can provide sentiment analysis

upvoted 2 times

FastTrack 3 years, 8 months ago

A,
https://aws.amazon.com/getting-started/hands-on/analyze-sentiment-comprehend/

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 16 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 16
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

During mini-batch training of a neural network for a classification problem, a Data Scientist notices that training accuracy oscillates.
What is the MOST likely cause of this issue?

A. The class distribution in the dataset is imbalanced.

B. Dataset shuffling is disabled.

C. The batch size is too big.

D. The learning rate is very high.

Show Suggested Answer

by ozan11 at *Jan. 20, 2020, 12:07 p.m.*

Comments

gaku1016 Highly Voted 3 years, 9 months ago

Answer is D.
Should the weight be increased or reduced so that the error is smaller than the current value? You need to examine the amount of change to know that. Therefore, we differentiate and check whether the slope of the tangent is positive or negative, and update the weight value in the direction to reduce the error. The operation is repeated over and over so as to approach the optimal solution that is the goal. The width of the update amount is important at this time, and is determined by the learning rate.

upvoted 17 times

ozan11 Highly Voted 3 years, 9 months ago

maybe D ?

upvoted 8 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

D. The learning rate is very high.

Explanation:
When the learning rate is too high, the optimization process may overshoot the optimal weights in parameter space. Instead of gradually converging, the model updates weights in a highly unstable manner, causing fluctuations in training accuracy. The network fails to settle into a minimum because the updates are too aggressive.

upvoted 2 times

AjoseO 9 months, 3 weeks ago

Selected Answer: D

A high learning rate can cause oscillations in the training accuracy because the optimizer makes large updates to the model parameters in each iteration, which can cause overshooting the optimal values. This can result in the model oscillating back and forth across the optimal solution.

upvoted 3 times

Mickey321 9 months, 3 weeks ago

Selected Answer: D

If the learning rate is too high, the model weights may overshoot the optimal values and bounce back and forth around the minimum of the loss function. This can cause the training accuracy to oscillate and prevent the model from converging to a stable solution. The training accuracy is the proportion of correct predictions made by the model on the training data.

upvoted 2 times

Rejju 9 months, 3 weeks ago

When the learning rate is set too high, it can lead to oscillations or divergence during training. Here's why:

High Learning Rate: A high learning rate means that the model's parameters are updated by a large amount in each training step. This can cause the model to overshoot the optimal parameter values, leading to instability in training.

Oscillations: If the learning rate is excessively high, the model's updates can become unstable, causing it to oscillate back and forth between parameter values. This oscillation can prevent the model from converging to an optimal solution.

To address this issue, you can try reducing the learning rate. It's often necessary to experiment with different learning rates to find the one that works best for your specific problem and dataset. Learning rate scheduling techniques, such as reducing the learning rate over time, can also help stabilize training.

upvoted 2 times

CKS1210 2 years ago

Answer is A.
A high learning rate means that the model parameters are being updated by large magnitudes in each iteration. As a result, the optimization process may struggle to converge to the optimal solution, leading to erratic behavior and fluctuations in training accuracy.

upvoted 1 times

soonmo 2 years, 1 month ago

Selected Answer: D

If learning rate is high, the accuracy is fluctuated because the value of loss function moves back and forth over the global minimum.

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: D

The big learning rating overshoot in true minima.

upvoted 2 times

Tomatoteacher 2 years, 5 months ago

Selected Answer: D

D Learning rate is too high. Textbook example of learning rate being too high. Lower Learning_rate will take more iterations, or longer to train, but will settle in place.

upvoted 1 times

Shailendraa 2 years, 10 months ago

12-sep exam

upvoted 1 times

Sam1610 3 years ago

D: per supuesto

upvoted 1 times

missionml 3 years, 4 months ago

A company sells thousands of products on a public website and wants to automatically identify products with potential durability problems. The company has 1.000 reviews with date, star rating, review text, review summary, and customer email fields, but many reviews are incomplete and have empty fields. Each review has already been labeled with the correct durability result.
A machine learning specialist must train a model to identify reviews expressing concerns over product durability. The first model needs to be trained and ready to review in 2 days.
What is the MOST direct approach to solve this problem within 2 days?
A.
Train a custom classifier by using Amazon Comprehend.
B.
Build a recurrent neural network (RNN) in Amazon SageMaker by using Gluon and Apache MXNet.
C.
Train a built-in BlazingText model using Word2Vec mode in Amazon SageMaker.
D.
Use a built-in seq2seq model in Amazon SageMaker.

upvoted 1 times

missionml 3 years, 4 months ago

Is A valid option?

upvoted 1 times

btsql 3 years, 8 months ago

D is correct. big batch size make local minia.

upvoted 1 times

jeetss1 3 years, 8 months ago

it is a multiple answer question and answer should be both A and D

upvoted 1 times

syu31svc 3 years, 8 months ago

Answer is D 100%; learning rate too high will cause such an event

upvoted 3 times

deep_n 3 years, 8 months ago

The answer is D, from the Coursera deep learning specialization (course 2 - improving Deep NN)

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 15 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 15
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist at a company sensitive to security is preparing a dataset for model training. The dataset is stored in Amazon S3 and contains
Personally Identifiable Information (PII).
The dataset:
⇨ Must be accessible from a VPC only.
⇨ Must not traverse the public internet.
How can these requirements be satisfied?

A. Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.

B. Create a VPC endpoint and apply a bucket access policy that allows access from the given VPC endpoint and an Amazon EC2 instance.

C. Create a VPC endpoint and use Network Access Control Lists (NACLs) to allow traffic between only the given VPC endpoint and an Amazon EC2 instance.

D. Create a VPC endpoint and use security groups to restrict access to the given VPC endpoint and an Amazon EC2 instance

Show Suggested Answer

by JayK at *Jan. 2, 2020, 6:49 p.m.*

Comments

rajs Highly Voted 3 years, 9 months ago

Important things to note here is that

1. "The Data in S3 Needs to be Accessible from VPC"
2. "Traffic should not Traverse internet"

To fulfill Requirement #2 we need a VPC endpoint
To RESTRICT the access to S3/Bucket
- Access allowed only from VPC via VPC Endpoint

Even though Sagemaker uses EC2 - we are NOT asked to secure the EC2 :)

So the answer is A

upvoted 41 times

sdsfsdsf Highly Voted 3 years, 9 months ago

Between A & B, the answer should be A. From here:
https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html#vpc-endpoints-s3-bucket-policies
We can see that we restrict access using DENY if sourceVpce (vpc endpoint), or sourceVpc (vpc) is not equal to our VPCe/VPC. So we are using a DENY (choice A) and not an ALLOW policy (choice B).

Choices C, D we eliminate because they don't address S3 access at all.

upvoted 12 times

JonSno Most Recent 4 months, 4 weeks ago

Selected Answer: A

Create a VPC endpoint and apply a bucket access policy that restricts access to the given VPC endpoint and the VPC.
Why is this correct?
VPC endpoint for S3 allows private connectivity between Amazon S3 and the VPC without using the public internet.
Bucket access policy can be written to allow access only from this VPC endpoint.
This ensures maximum security by:
Preventing access from outside the VPC.
Blocking public access.

upvoted 2 times

Ajose0 9 months, 3 weeks ago

Selected Answer: A

In Option A, the Machine Learning Specialist would create a VPC endpoint for Amazon S3, which would allow traffic to flow directly between the VPC and Amazon S3 without traversing the public internet. Access to the S3 bucket containing PII can then be restricted to the VPC endpoint and the VPC using a bucket access policy. This would ensure that only instances within the VPC can access the data, and that the data does not traverse the public internet.

Option B and D, allowing access from an Amazon EC2 instance, would not meet the requirement of not traversing the public internet, as the EC2 instance would be accessible from the internet. Option C, using Network Access Control Lists (NACLs) to allow traffic between only the VPC endpoint and an EC2 instance, would also not meet the requirement of not traversing the public internet, as the EC2 instance would still be accessible from the internet.

upvoted 1 times

loict 9 months, 3 weeks ago

Selected Answer: A

A. YES - We first create a S3 endpoint in the VPC subnet so traffic does not flow through the Internet, then on the S3 bucket create an access policy that restricts access to the given VPC based on its ID
B. NO - we don't want to be specific to an instance
C. NO - the S3 bucket is on AWS network, you cannot change the NACL for it
D. NO - not all instances in a VPC will necessarily have the same principal that can be specified in the policy

upvoted 2 times

Mickey321 1 year, 11 months ago

Selected Answer: A

Definetly A

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: A

Well, but removing methodology, only A remains: The question never cited EC2

upvoted 3 times

ADVIT 2 years ago

Per https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html it's A

upvoted 1 times

exam887 3 years, 1 month ago

Selected Answer: A

The question do not mention EC2 at all, so should be A

upvoted 4 times

dunhill 3 years, 6 months ago

I think it should be B. Traning instance is a EC2 instance and need to be set an endpoint to load the data from S3.

upvoted 1 times

[Removed] 3 years, 7 months ago

Selected Answer: B

AWS security is a conservative security model, which implies that access are denied by default rather than granted by default. We have to explicitly allow access to a AWS resource. Additionally, B talks about allowing access FROM the VPC to S3 while A talks about allowing access from S3 to VPC (which is not what we need).
So, B.

upvoted 2 times

cpal012 2 years, 3 months ago

Um, no. A VPC endpoint is outbound from the VPC to a supported AWS service.

upvoted 1 times

technoguy 3 years, 8 months ago

Will go with B

upvoted 1 times

spamicho 3 years, 8 months ago

Betting on B here, we should control access from VPC, not to VPC.

upvoted 1 times

achiko 3 years, 8 months ago

A!
Restricting access to a specific VPC endpoint
The following is an example of an Amazon S3 bucket policy that restricts access to a specific bucket, awsexamplebucket1, only from the VPC endpoint with the ID vpce-1a2b3c4d. The policy denies all access to the bucket if the specified endpoint is not being used. The aws:SourceVpce condition is used to specify the endpoint.

https://docs.aws.amazon.com/AmazonS3/latest/userguide/example-bucket-policies-vpc-endpoint.html

upvoted 2 times

senseikimoji 3 years, 8 months ago

Can't be B. You simple cannot enable access to an endpoint to some selected instance. So A.

upvoted 1 times

Huy 3 years, 8 months ago

We shouldn't use private IP in bucket policy.

upvoted 1 times

cloud_trail 3 years, 8 months ago

B does not say enable access TO the VPC endpoint. It says to allow access FROM the endpoint. So B is the correct answer. A talks about restricting access TO the VPC endpoint, so that option is irrelevant. We're worried about access TO the S3 bucket, not access to the VPC. The question is not poorly-worded, but it is tricky and you need to read it carefully.

upvoted 1 times

yeetusdeleetus 3 years, 8 months ago

I also vote A.

upvoted 1 times

Thai_Xuan 3 years, 8 months ago

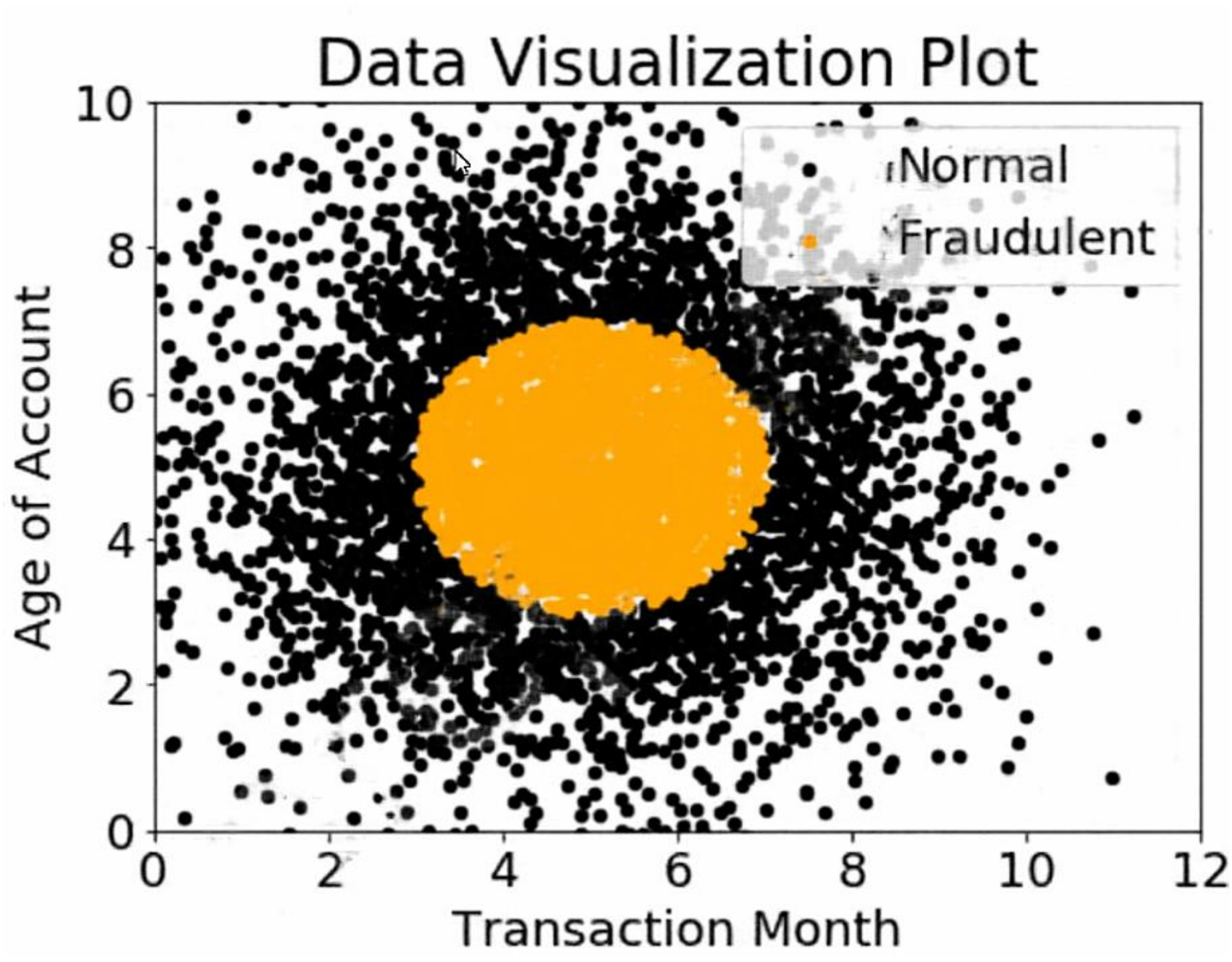
A
found here
"You can control which VPCs or VPC endpoints have access to your buckets by using Amazon S3 bucket policies. For examples of this type of bucket policy access control, see the following topics on restricting access."
https://docs.aws.amazon.com/AmazonS3/latest/dev/example-bucket-policies-vpc-endpoint.html

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 14 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 14
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a Machine Learning Specialist would like to build a binary classifier based on two features: age of account and transaction month. The class distribution for these features is illustrated in the figure provided.



Based on this information, which model would have the HIGHEST accuracy?

A. Long short-term memory (LSTM) model with scaled exponential linear unit (SELU)

B. Logistic regression

C. Support vector machine (SVM) with non-linear kernel

D. Single perceptron with tanh activation function

Show Suggested Answer

by cnethers at *Feb. 3, 2021, 11:30 a.m.*

Comments

mizuakari Highly Voted 3 years, 9 months ago

Answer is C. SVM sample use case is to put the dimensions into a higher hyperplane that can separates it. Seeing how separable it is, SVM can be used for it.

upvoted 21 times

JonSno Most Recent 4 months, 4 weeks ago

Selected Answer: C

Support Vector Machine (SVM) with Non-Linear Kernel --> Non-linear Data
Why?
SVM is powerful for classification and works well even with small datasets.
If the data has a non-linear decision boundary, using an SVM with a non-linear kernel (like RBF or polynomial) can improve accuracy.
Works well in low-dimensional feature spaces (since we have only 2 features: age of account & transaction month).
Optimal choice if the data has a non-linear decision boundary.

upvoted 1 times

MultiCloudIronMan 8 months, 3 weeks ago

Selected Answer: C

SVMs are particularly effective for binary classification tasks and can handle non-linear relationships between features1.

upvoted 2 times

Mickey321 9 months, 3 weeks ago

Selected Answer: C

You can use a support vector machine (SVM) when your data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.

upvoted 2 times

kaike_reis 1 year, 11 months ago

Well, C is the correct answer. This example is a classical one to use SVM.

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: C

SVM for RBF mode is the answer!

upvoted 1 times

Broncomailo 3 years, 5 months ago

Selected Answer: C

Answer is C

upvoted 4 times

Dr_Kiko 3 years, 8 months ago

Textbook C

upvoted 3 times

halfway 3 years, 8 months ago

C. more reading for using non-linear kernel and separate samples with a hyperplane in a higher dimension space: https://medium.com/pursuitnotes/day-12-kernel-svm-non-linear-svm-5fdefe77836c

upvoted 2 times

spamicho 3 years, 9 months ago

C seems right

upvoted 1 times

Juka3lj 3 years, 9 months ago

answer is C

upvoted 2 times

omar_bahrain 3 years, 9 months ago

Agree. The answer is A.

https://www.surveypractice.org/article/2715-using-support-vector-machines-for-survey-research

upvoted 1 times

cnethers 3 years, 9 months ago

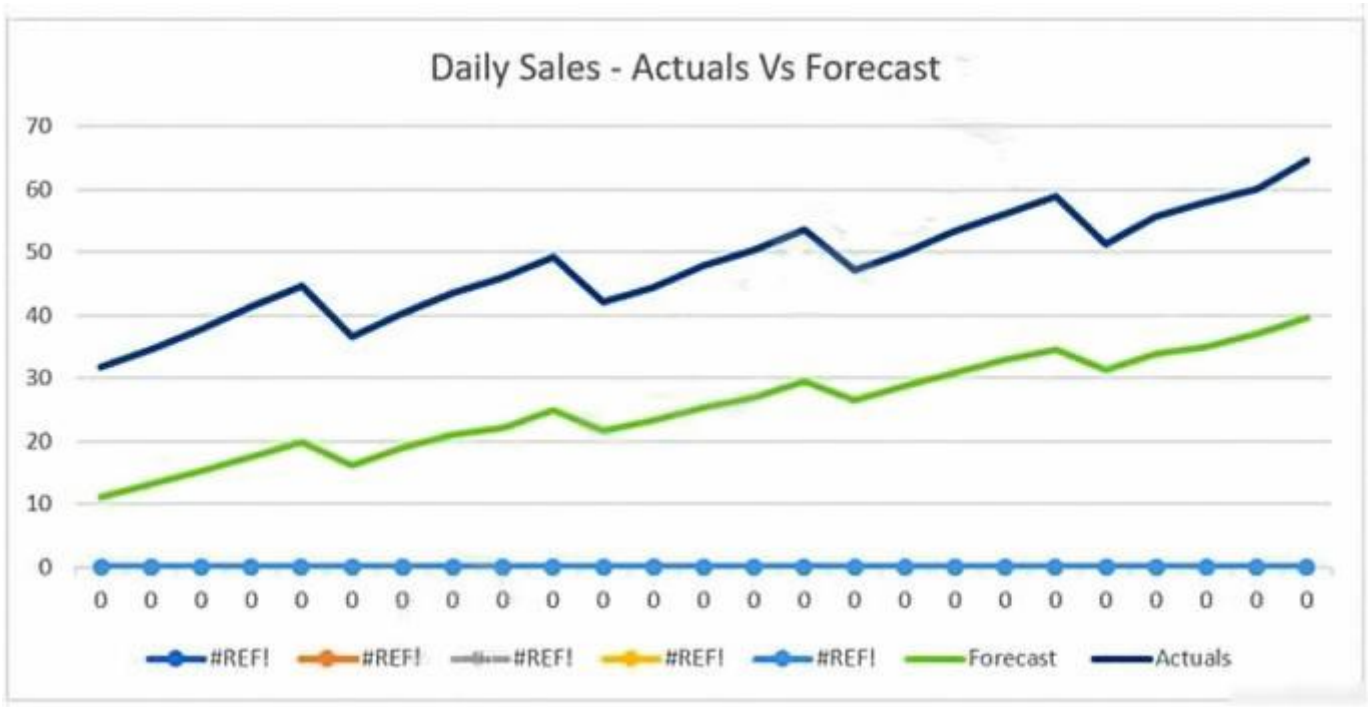
This is a good explanation of SVM
https://uk.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 13 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 13
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

The displayed graph is from a forecasting model for testing a time series.



Considering the graph only, which conclusion should a Machine Learning Specialist make about the behavior of the model?

A. The model predicts both the trend and the seasonality well

B. The model predicts the trend well, but not the seasonality.

C. The model predicts the seasonality well, but not the trend.

D. The model does not predict the trend or the seasonality well.

Show Suggested Answer

---

by ashlash at *Feb. 22, 2021, 9:04 a.m.*

Comments

jeetss1 Highly Voted 3 years, 3 months ago

A is correct answer.
Please Refer: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

upvoted 16 times

Dr_Kiko Highly Voted 3 years, 2 months ago

A; the problem is bias, not trends

upvoted 8 times

apache007 Most Recent 3 months, 2 weeks ago

Selected Answer: B

B. The model predicts the trend well, but not the seasonality.

Here's what we can observe:
The predicted mean line closely follows the general upward trend of the observed line.
The predicted mean line does not capture the high frequency up and down changes of the observed line.

upvoted 1 times

VR10 10 months, 4 weeks ago

agreed, this seems to be A. there is similarity between the blue and green lines as far as capturing trend and seasonality is concerned. It just seems that if assumption is that the model is a linear regression model then just the intercept is off by a few units.

upvoted 2 times

Mickey321 1 year, 5 months ago

Selected Answer: A

A. The model predicts both the trend and the seasonality well

upvoted 1 times

Valcilio 1 year, 10 months ago

Selected Answer: A

The problem is Bias not trends or sesonality!

upvoted 2 times

spamicho 3 years, 2 months ago

A is right, both trend (rising) and seasonality is there

upvoted 3 times

btsql 3 years, 2 months ago

C is correct answer

upvoted 1 times

btsql 3 years, 2 months ago

A is correct answer. Not C

upvoted 2 times

Kuntazulu 3 years, 3 months ago

The trend is up, so isn't it correctly predicted? And the seasonality is also in sync, the amplitude is wrong.

upvoted 3 times

georschi 3 years, 3 months ago

A is right. trend and seasonality are fine, level is the one the model gets wrong

upvoted 4 times

NotAnMLProfessional 3 years, 3 months ago

Should be C

upvoted 1 times

ashlash 3 years, 3 months ago

Should be A

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 12 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 12
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is working with a large company to leverage machine learning within its products. The company wants to group its customers into categories based on which customers will and will not churn within the next 6 months. The company has labeled the data available to the Specialist.
Which machine learning model type should the Specialist use to accomplish this task?

A. Linear regression

B. Classification

C. Clustering

D. Reinforcement learning

Show Suggested Answer

by rsimham at *Dec. 9, 2019, 2:13 a.m.*

Comments

rsimham Highly Voted 3 years, 9 months ago

B seems to be okay

upvoted 14 times

JonSno Most Recent 4 months, 4 weeks ago

Selected Answer: B

CLASSIFICATION - Binary Classification - Supervised Learning to be precise
The company wants to predict customer churn (whether a customer will leave or stay).
The data is labeled, meaning we have historical outcomes (churn or no churn).
The task involves categorizing customers into two groups:
Customers who will churn (leave)
Customers who will not churn (stay)
This means the problem is a Supervised Learning problem, specifically a binary classification problem.
The company wants to predict customer churn (whether a customer will leave or stay).
The data is labeled, meaning we have historical outcomes (churn or no churn).
The task involves categorizing customers into two groups:
Customers who will churn (leave)
Customers who will not churn (stay)
This means the problem is a Supervised Learning problem, specifically a binary classification problem.

upvoted 2 times

Mickey321 9 months, 3 weeks ago

Selected Answer: B

The reason for this choice is that classification is a type of supervised learning that predicts a discrete categorical value, such as yes or no, spam or not spam, or churn or not churn1. Classification models are trained using labeled data, which means that the input data has a known target attribute that indicates the correct class for each instance2. For example, a classification model that predicts customer churn would use data that has a label indicating whether the customer churned or not in the past.

Classification models can be used for various applications, such as sentiment analysis, image recognition, fraud detection, and customer segmentation2. Classification models can also handle both binary and multiclass problems, depending on the number of possible classes in the target attribute3.

upvoted 1 times

Sharath1783 9 months, 3 weeks ago

Selected Answer: B

Option B. This is a scenario for supervised learning model as data is labelled and only A, B are supervised learning algorithms from the options. Linear learning is to predict time series data and distribution is selecting which class the input belongs to. Hence most suitable is to use Binomial distribution model in this case.

upvoted 1 times

loict 9 months, 3 weeks ago

Selected Answer: B

A. NO - Linear regression is not best for classification
B. YES - Classification
C. NO - we want supervised classification
D. NO - there is nothing to Reinforce from

upvoted 1 times

mirik 2 years ago

The question is not clear. Actually we have 2 tasks here - group into categories (clustering) and predict if customers will churn/not churn (classification). If we had to simply do classification, why there was mentioned to group into categories?

upvoted 3 times

ovokpus 3 years ago

Selected Answer: B

This is definitely a classification problem

upvoted 4 times

Sivadharan 3 years, 2 months ago

Selected Answer: B

B is correct

upvoted 2 times

FabG 3 years, 8 months ago

B - it's a Binary Classification problem. Will the customer churn: Yes or No

upvoted 4 times

syu31svc 3 years, 8 months ago

100% is B since it is about labelled data

upvoted 1 times

eji 3 years, 8 months ago

i think the key is "the company has labeled the data" so this is classification, so it's B

upvoted 3 times

roytruong 3 years, 9 months ago

B is okey

upvoted 2 times

cybe001 3 years, 9 months ago

B is correct

upvoted 3 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 11 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 11

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist receives customer data for an online shopping website. The data includes demographics, past visits, and locality information. The
Specialist must develop a machine learning approach to identify the customer shopping patterns, preferences, and trends to enhance the website for better service and smart recommendations.
Which solution should the Specialist recommend?

A. Latent Dirichlet Allocation (LDA) for the given collection of discrete data to identify patterns in the customer database.

B. A neural network with a minimum of three layers and random initial weights to identify patterns in the customer database.

C. Collaborative filtering based on user interactions and correlations to identify patterns in the customer database.

D. Random Cut Forest (RCF) over random subsamples to identify patterns in the customer database.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 3:13 a.m.*

Comments

WWODIN Highly Voted 3 years, 9 months ago

answer should be C
Collaborative filtering is for recommendation, LDA is for topic modeling

upvoted 21 times

syu31svc Highly Voted 3 years, 8 months ago

In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set

Neural network is used for image detection

Answer is C

upvoted 12 times

Vernoxx Most Recent 2 months, 3 weeks ago

Selected Answer: C

I think it should be c

upvoted 1 times

JonSno 4 months, 4 weeks ago

Selected Answer: C

Collab filtering it is..
Collaborative filtering is the most widely used approach for recommendation systems.
It uses customer interactions (purchases, clicks, ratings) to determine preferences based on similar users or items.
Implicit collaborative filtering (based on user behavior) and explicit collaborative filtering (based on ratings) can effectively personalize recommendations.

upvoted 1 times

loict 9 months, 3 weeks ago

Selected Answer: C

A. NO - LDA is for topic modeling
B. NO - NN is a too generic term, you want Neural Collaborative
C. YES - Collaborative filtering best fit
D. NO - Random Cut Forest (RCF) for anomalities

upvoted 3 times

Mickey321 9 months, 3 weeks ago

Selected Answer: C

Collaborative filtering is a machine learning technique that recommends products or services to users based on the ratings or preferences of other users. This technique is well-suited for identifying customer shopping patterns and preferences because it takes into account the interactions between users and products.

upvoted 1 times

killermouse0 1 year, 3 months ago

Selected Answer: A

From the doc: "You can use LDA for a variety of tasks, from clustering customers based on product purchases to automatic harmonic analysis in music."

https://docs.aws.amazon.com/sagemaker/latest/dg/lda-how-it-works.html

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: C

C, always when talk about recommendation you can think about collaborative patterns!

upvoted 2 times

stjokerli 2 years, 4 months ago

A
LDA used before collaborative filtering is largely adopted.
1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some of the items, which is NOT what we have
2) recommendation is just one thing that we want to do. What about trends?
3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

upvoted 2 times

Shailendraa 2 years, 10 months ago

collaborative

upvoted 1 times

apprehensive_scar 3 years, 5 months ago

C. Easy question.

upvoted 1 times

technoguy 3 years, 8 months ago

its a appropriate use case of Collaborative filtering

upvoted 1 times

roytruong 3 years, 9 months ago

this is C

upvoted 1 times

sdsfsdsf 3 years, 9 months ago

I'm thinking that it is A because:
1) the input data that we have doesn't lend itself to collaborative filtering - it requires a set of items and a set of users who have reacted to some

of the items, which is NOT what we have
2) recommendation is just one thing that we want to do. What about trends?
3) collaborative filtering isn't one of the pre-built algorithms (weak argument, admittedly)

upvoted 6 times

cybe001 3 years, 9 months ago

Answer is C, demographics, past visits, and locality information data, LDA is appropriate

upvoted 3 times

cybe001 3 years, 9 months ago

Collaborative filtering is appropriate

upvoted 4 times

DonaldCMLIN 3 years, 9 months ago

Answer A might be more suitable than other

https://docs.aws.amazon.com/zh_tw/sagemaker/latest/dg/lda-how-it-works.html

upvoted 4 times

rsimham 3 years, 9 months ago

Not convinced with A. Answer C seems to be a better fit than A for recommendation model (LDA appears to be a topic-based model on unavailable data with similar patterns)
https://aws.amazon.com/blogs/machine-learning/extending-amazon-sagemaker-factorization-machines-algorithm-to-predict-top-x-recommendations/

upvoted 10 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 10 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 10
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist has completed a proof of concept for a company using a small data sample, and now the Specialist is ready to implement an end- to-end solution in AWS using Amazon SageMaker. The historical training data is stored in Amazon RDS.
Which approach should the Specialist use for training a model using that data?

A. Write a direct connection to the SQL database within the notebook and pull data in

B. Push the data from Microsoft SQL Server to Amazon S3 using an AWS Data Pipeline and provide the S3 location within the notebook.

C. Move the data to Amazon DynamoDB and set up a connection to DynamoDB within the notebook to pull data in.

D. Move the data to Amazon ElastiCache using AWS DMS and set up a connection within the notebook to pull data in for fast access.

Show Suggested Answer

by JayK at *Jan. 4, 2020, 1:27 p.m.*

Comments

JayK Highly Voted 9 months, 2 weeks ago
Answer is B as the data for a SageMaker notebook needs to be from S3 and option B is the only option that says it. The only thing with option B is that it is talking of moving data from MS SQL Server not RDS
upvoted 31 times

mlyu 3 years, 9 months ago
https://www.slideshare.net/AmazonWebServices/train-models-on-amazon-sagemaker-using-data-not-from-amazon-s3-aim419-aws-reinvent-2018
upvoted 2 times

HaiHN 3 years, 8 months ago
Please look at the slide 14 of that link, although the data source from DynamoDB or RDS, it is still need to use AWS Glue to move the data to S3 for SageMaker to use.

So, the right anwser should be B.
upvoted 2 times

jasonsunbao 3 years, 9 months ago
I agree. As from the ML developer guide I just read, it is the MYSQL RDS that can be used as SQL datasource.
upvoted 2 times

Rama_Adim Most Recent 1 month, 3 weeks ago

Selected Answer: A

Sagemaker can read from RDS directly - Maybe this is a new feature. Please check.
https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/

upvoted 1 times

JonSno 4 months, 4 weeks ago

Selected Answer: B

Amazon SageMaker does not natively connect to Amazon RDS. Instead, training jobs work best with data stored in Amazon S3.
Amazon S3 is the preferred data source for SageMaker because:
It integrates seamlessly with SageMaker's training job infrastructure.
It supports distributed training for large datasets.
It is cost-effective and decouples storage from compute.
Best practice → Export RDS data to Amazon S3 and train using SageMaker.

upvoted 2 times

SophieSu 9 months, 3 weeks ago

B is the correct answer.

Official AWS Documentation:
"Amazon ML allows you to create a datasource object from data stored in a MySQL database in Amazon Relational Database Service (Amazon RDS). When you perform this action, Amazon ML creates an AWS Data Pipeline object that executes the SQL query that you specify, and places the output into an S3 bucket of your choice. Amazon ML uses that data to create the datasource."

upvoted 2 times

AjoseO 9 months, 3 weeks ago

Selected Answer: B

In Option B approach, the Specialist can use AWS Data Pipeline to automate the movement of data from Amazon RDS to Amazon S3. This allows for the creation of a reliable and scalable data pipeline that can handle large amounts of data and ensure the data is available for training.

In the Amazon SageMaker notebook, the Specialist can then access the data stored in Amazon S3 and use it for training the model. Using Amazon S3 as the source of training data is a common and scalable approach, and it also provides durability and high availability of the data.

upvoted 2 times

Mickey321 9 months, 3 weeks ago

Selected Answer: B

This approach is the most scalable and reliable way to train a model using data stored in Amazon RDS. Amazon S3 is a highly scalable and durable object storage service, and Amazon Data Pipeline is a managed service that makes it easy to move data between different AWS services. By pushing the data to Amazon S3, the Specialist can ensure that the data is available for training the model even if the Amazon RDS instance is unavailable.

upvoted 2 times

loict 9 months, 3 weeks ago

Selected Answer: B

A. NO - SageMaker can only read from S3
B. YES - AWS Data Pipeline can moved from SQL Server to S3
C. NO - SageMaker can only read from S3 and not DynamoDB
D. NO - SageMaker can only read from S3 and not ElastiCache

upvoted 2 times

shammous 9 months, 3 weeks ago

Selected Answer: B

Option B (exporting to S3) is typically more flexible and cost-effective for large-scale or complex data needs (Which is our case - production), while Option A (direct connection) can be simpler and more immediate for real-time or smaller-scale scenarios like testing.

upvoted 1 times

ninomfr64 9 months, 3 weeks ago

Selected Answer: B

A. NO. It is doable, but this is not the best approach.
B. YES
C. NO. Pushing data to DynamoDB would not make it easier to access data
D. NO. Pushing data to ElastiCache would not make it easier to access data

upvoted 1 times

Denise123 1 year, 3 months ago

Selected Answer: A

For Amazon S3, you can import data from an Amazon S3 bucket as long as you have permissions to access the bucket.

For Amazon Athena, you can access databases in your AWS Glue Data Catalog as long as you have permissions through your Amazon Athena workgroup.

For Amazon RDS, if you have the AmazonSageMakerCanvasFullAccess policy attached to your user's role, then you'll be able to import data from your Amazon RDS databases into Canvas.

https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-connecting-external.html

upvoted 4 times

Aja1 1 year, 2 months ago

https://aws.amazon.com/about-aws/whats-new/2024/04/amazon-sagemaker-studio-notebooks-data-sql-query/

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

It's B, even if Microsoft SQL Server is a strange name for RDS, it's a possible database to use there and the data for sagemaker needs to be in S3!

upvoted 1 times

cnethers 3 years, 8 months ago

While B is a valid answer, It is also possible to make a SQL connection in a notebook and create a data object so A could be a valid answer too
https://stackoverflow.com/questions/36021385/connecting-from-python-to-sql-server
https://www.mssqltips.com/sqlservertip/6120/data-exploration-with-python-and-sql-server-using-jupyter-notebooks/

upvoted 2 times

gcpwhiz 3 years, 8 months ago

you need to choose the best answer, not any valid answer. Often, many of the answers are valid solutions, but are not best practice.

upvoted 2 times

scuzzy2010 3 years, 8 months ago

B is correct. MS SQL Server is also under RDS.

upvoted 2 times

roytruong 3 years, 8 months ago

B is right

upvoted 2 times

bhavesh0124 3 years, 9 months ago

B it is

upvoted 1 times

cybe001 3 years, 9 months ago

I'll go with B

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 9 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 9
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is developing a custom video recommendation model for an application. The dataset used to train this model is very large with millions of data points and is hosted in an Amazon S3 bucket. The Specialist wants to avoid loading all of this data onto an Amazon SageMaker notebook instance because it would take hours to move and will exceed the attached 5 GB Amazon EBS volume on the notebook instance.
Which approach allows the Specialist to use all the data to train the model?

A. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.

B. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to the instance. Train on a small amount of the data to verify the training code and hyperparameters. Go back to Amazon SageMaker and train using the full dataset

C. Use AWS Glue to train a model using a small subset of the data to confirm that the data will be compatible with Amazon SageMaker. Initiate a SageMaker training job using the full dataset from the S3 bucket using Pipe input mode.

D. Load a smaller subset of the data into the SageMaker notebook and train locally. Confirm that the training code is executing and the model parameters seem reasonable. Launch an Amazon EC2 instance with an AWS Deep Learning AMI and attach the S3 bucket to train the full dataset.

Show Suggested Answer

by cmm103 at *Dec. 3, 2019, 7:03 p.m.*

Comments

JayK Highly Voted 3 years, 9 months ago

Answer is A. The answer to this question is about Pipe mode from S3. The only options are A and C. As AWS Glue cannot be use to create models which is option C.
The correct answer is A

upvoted 31 times

liangfb Highly Voted 3 years, 9 months ago

Answer is A.

upvoted 13 times

JonSno Most Recent 4 months, 4 weeks ago

Selected Answer: A

Training locally on a small dataset ensures the training script and model parameters are working correctly.
Amazon SageMaker training jobs allow direct access to S3 data without downloading everything.
Pipe input mode efficiently streams data from S3 to the training instance, reducing disk space requirements and speeding up training.

upvoted 4 times

reginav 6 months, 4 weeks ago

Selected Answer: A

Only Pipe mode can stream data from S3

upvoted 1 times

Mickey321 9 months, 3 weeks ago

Selected Answer: A

The reason for this choice is that Pipe input mode is a feature of Amazon SageMaker that allows you to stream data directly from an Amazon S3 bucket to your training instances without downloading it first1. This way, you can avoid the time and space limitations of loading a large dataset onto your notebook instance. Pipe input mode also offers faster start times and better throughput than File input mode, which downloads the entire dataset before training1.

upvoted 3 times

loict 9 months, 3 weeks ago

Selected Answer: A

A. YES - pipe mode is best to start inference before the entire data is transferred; the only drawback is if multiple training jobs are done in sequence (eg. different hyperparamater), the data will be downloaded again
B. NO - we want to use SageMaker first for initial training
C. NO - We first want to test things in SageMaker
D. NO - the SageMaker notebook will not use the AMI so the testing done is useless

upvoted 1 times

kyuhuck 1 year, 5 months ago

Selected Answer: B

B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.

This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: A

I think it should be a

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: A

It's A, pipe mode is for dealing with very big data.

upvoted 2 times

yemauricio 2 years, 6 months ago

Selected Answer: A

A, PIPE is to do that sort of modeling

upvoted 2 times

Shailendraa 2 years, 10 months ago

When data is already in S3 and next it should move to Sagemaker.. so option A is suitable

upvoted 1 times

Huy 3 years, 8 months ago

Answer is A. B, C & D can be dropped because there is no integration from/to Sage Maker train job (model).

upvoted 1 times

cloud_trail 3 years, 8 months ago

Gotta be A. You need to use Pipe mode but Glue cannot train a model.

upvoted 2 times

bobdylan1 3 years, 8 months ago

AAAAAAAAAAa

upvoted 1 times

Willnguyen22 3 years, 9 months ago

ans is A

upvoted 1 times

GeeBeeEl 3 years, 9 months ago

Will you run AWS Deep Learning AMI for all cases where the data is very large in S3? Also what role is Glue playing here? Is there a transformation? These are the two issues for options B C and D. I believe they do not represent what is required to satisfy the requirements in the question. The answer definitely requires the pipe mode, but not with Glue. I go with A https://aws.amazon.com/blogs/machine-learning/using-pipe-input-mode-for-amazon-sagemaker-algorithms/

upvoted 3 times

roytruong 3 years, 9 months ago

go for A

upvoted 2 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 8 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 8
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A manufacturing company has structured and unstructured data stored in an Amazon S3 bucket. A Machine Learning Specialist wants to use SQL to run queries on this data.
Which solution requires the LEAST effort to be able to query this data?

A. Use AWS Data Pipeline to transform the data and Amazon RDS to run queries.

B. Use AWS Glue to catalogue the data and Amazon Athena to run queries.

C. Use AWS Batch to run ETL on the data and Amazon Aurora to run the queries.

D. Use AWS Lambda to transform the data and Amazon Kinesis Data Analytics to run queries.

Show Suggested Answer

by cybe001 at *Jan. 11, 2020, 4:44 p.m.*

Comments

cybe001 Highly Voted 3 years, 9 months ago
B is correct
upvoted 24 times

dhs227 Highly Voted 3 years, 9 months ago
The correct answer HAS TO be B
Using Glue Use AWS Glue to catalogue the data and Amazon Athena to run queries against data on S3 are very typical use cases for those services.

D is not ideal, Lambda can surely do many things but it requires development/testing effort, and Amazon Kinesis Data Analytics is not ideal for ad-hoc queries.
upvoted 9 times

JonSno Most Recent 4 months, 4 weeks ago
Selected Answer: B
B. Use AWS Glue to catalog the data and Amazon Athena to run queries.
Why is this the best choice?
AWS Glue can automatically catalog both structured and unstructured data in S3.
Amazon Athena is a serverless SQL query service that allows direct SQL queries on S3 data without moving it.
No infrastructure setup is required—just define a Glue Data Catalog and start querying with Athena.

upvoted 2 times

reginav 6 months, 4 weeks ago

Selected Answer: B

S3 query === athena , to catalog data glue

upvoted 1 times

AjoseO 9 months, 3 weeks ago

Selected Answer: B

AWS Glue is a fully managed ETL service that makes it easy to move data between data stores. It can automatically crawl, catalogue, and classify data stored in Amazon S3, and make it available for querying and analysis. With AWS Glue, you don't have to worry about the underlying infrastructure and can focus on your data.

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. It integrates with AWS Glue, so you can use the catalogued data directly in Athena without any additional data movement or transformation.

upvoted 3 times

Mickey321 1 year, 11 months ago

Selected Answer: B

The reason for this choice is that AWS Glue is a fully managed service that provides a data catalogue to make your data in S3 searchable and queryable1. AWS Glue crawls your data sources, identifies data formats, and suggests schemas and transformations1. You can use AWS Glue to catalogue both structured and unstructured data, such as relational data, JSON, XML, CSV files, images, or media files2.

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

SK27 2 years, 7 months ago

Selected Answer: B

B is the easiest. We can use Glue crawler.

upvoted 2 times

ryuhei 2 years, 9 months ago

Selected Answer: B

Answer B

upvoted 2 times

vetaal 3 years, 5 months ago

Selected Answer: B

Querying data in S3 with SQL is almost always Athena.

upvoted 3 times

gcpwhiz 3 years, 8 months ago

If AWS asks the question of querying unstructured data in an efficient manner, it is almost always Athena

upvoted 2 times

cloud_trail 3 years, 8 months ago

B. I don't think that you even need Glue to transform anything. Just use Glue to define the schemas and then use Athena to query based on those schemas.

upvoted 2 times

Willnguyen22 3 years, 8 months ago

answer is B

upvoted 1 times

syu31svc 3 years, 8 months ago

SQL on S3 is Athena so answer is B for sure

upvoted 1 times

roytruong 3 years, 8 months ago

B is right

upvoted 2 times

Jayraam 3 years, 9 months ago

Answer is B.

Queries Against an Amazon S3 Data Lake
Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. If you want to build your own custom Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data.

https://aws.amazon.com/glue/

upvoted 1 times

PRC 3 years, 9 months ago

Correct Ans is D...Kinesis Data Analytics can use Lambda to transform and then run the SQL queries..

upvoted 1 times

Urban_Life 3 years, 8 months ago

May I know why you are taking complex route?

upvoted 10 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 7 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 7
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is building a model that will perform time series forecasting using Amazon SageMaker. The Specialist has finished training the model and is now planning to perform load testing on the endpoint so they can configure Auto Scaling for the model variant.
Which approach will allow the Specialist to review the latency, memory utilization, and CPU utilization during the load test?

A. Review SageMaker logs that have been written to Amazon S3 by leveraging Amazon Athena and Amazon QuickSight to visualize logs as they are being produced.

B. Generate an Amazon CloudWatch dashboard to create a single view for the latency, memory utilization, and CPU utilization metrics that are outputted by Amazon SageMaker.

C. Build custom Amazon CloudWatch Logs and then leverage Amazon ES and Kibana to query and visualize the log data as it is generated by Amazon SageMaker.

D. Send Amazon CloudWatch Logs that were generated by Amazon SageMaker to Amazon ES and use Kibana to query and visualize the log data.

Show Suggested Answer

by mlyu at *Jan. 7, 2020, 6:56 a.m.*

Comments

mlyu Highly Voted 3 years, 9 months ago
Agreed. Ans is B
upvoted 17 times

JonSno Most Recent 4 months, 4 weeks ago
Selected Answer: B
Generate an Amazon CloudWatch dashboard to create a single view for latency, memory utilization, and CPU utilization
Why?
Amazon SageMaker automatically pushes latency and instance utilization metrics to CloudWatch.
CloudWatch dashboards provide a single real-time view of these key metrics during load testing.
You can configure custom CloudWatch alarms to trigger auto scaling based on the load.
upvoted 2 times

teka112233 9 months, 3 weeks ago
Selected Answer: B

the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics
https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html

upvoted 2 times

Mickey321 9 months, 3 weeks ago

Selected Answer: B

The reason for this choice is that Amazon CloudWatch is a service that monitors and manages your cloud resources and applications. It collects and tracks metrics, which are variables you can measure for your resources and applications1. Amazon SageMaker automatically reports metrics such as latency, memory utilization, and CPU utilization to CloudWatch2. You can use these metrics to monitor the performance and health of your SageMaker endpoint during the load test.

upvoted 1 times

teka112233 1 year, 11 months ago

the question is clear that the specialist is seeking for latency, memory utilization, and CPU utilization during the load test and the ideal answer for all of these is amazon cloud watch which give you all these metrics
https://docs.aws.amazon.com/sagemaker/latest/dg/monitoring-cloudwatch.html

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

It's B, even the resources that aren't visible in a first try are visible if you use cloudwatch agent.

upvoted 3 times

DS2021 2 years, 6 months ago

Selected Answer: B

Should be B

upvoted 2 times

ystotest 2 years, 7 months ago

Selected Answer: B

agreed with B

upvoted 2 times

apprehensive_scar 3 years, 5 months ago

B is the ans

upvoted 1 times

anttan 3 years, 7 months ago

Should be C right, as Cloudwatch does not have metrics for memory utilization.

upvoted 2 times

anttan 3 years, 7 months ago

After further research, I think answer is B. While indeed true that Cloudwatch does not have metrics for memory utilization by default, you can achieve by installing ClouldWatch agent on the EC2. The EC2 used by Sagemaker is pre-installed with Cloudwatch Agent.

upvoted 2 times

Willnguyen22 3 years, 8 months ago

answer is B

upvoted 1 times

syu31svc 3 years, 8 months ago

Answer is B 100%; very straightforward method

upvoted 1 times

scuzzy2010 3 years, 8 months ago

B is correct. Don't need to use Kibana or QuickSight.

upvoted 1 times

roytruong 3 years, 9 months ago

ans is B

upvoted 3 times

cybe001 3 years, 9 months ago

B is correct

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 6 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 6
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is using an Amazon SageMaker notebook instance in a private subnet of a corporate VPC. The ML Specialist has important data stored on the Amazon SageMaker notebook instance's Amazon EBS volume, and needs to take a snapshot of that EBS volume. However, the ML Specialist cannot find the Amazon SageMaker notebook instance's EBS volume or Amazon EC2 instance within the VPC.
Why is the ML Specialist not seeing the instance visible in the VPC?

A. Amazon SageMaker notebook instances are based on the EC2 instances within the customer account, but they run outside of VPCs.

B. Amazon SageMaker notebook instances are based on the Amazon ECS service within customer accounts.

C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.

D. Amazon SageMaker notebook instances are based on AWS ECS instances running within AWS service accounts.

Show Suggested Answer

by mlyu at *Jan. 7, 2020, 6:25 a.m.*

Comments

mlyu Highly Voted 3 years, 9 months ago
I think the answer should be C
upvoted 25 times

dhs227 Highly Voted 3 years, 9 months ago
The correct answer HAS TO be A

The instances are running in customer accounts but it's in an AWS managed VPC while exposing ENI to customer VPC if it was chosen.
See explanation at https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/
upvoted 18 times

scuzzy2010 3 years, 9 months ago
Can't be A because A says "but they run outside of VPCs", which is not correct. They are attached to VPC, but it can either be AWS Service VPC or Customer VPC, or Both, as per the explanation url you provided.
upvoted 10 times

cloud_trail 3 years, 8 months ago
This is exactly right. According to that document, if the notebook instance is not in a customer VPC, then it has to be in the Sagemaker

managed VPC. See Option 1 in that document.
upvoted 1 times

mawsman 3 years, 9 months ago
Actually your link says: The notebook instance is running in an Amazon SageMaker managed VPC as shown in the above diagram. That means the correct answer is C. An Amazon SageMaker managed VPC can only be created in an Amazon managed Account.
upvoted 18 times

JonSno Most Recent 4 months, 4 weeks ago
Selected Answer: C
C. Amazon SageMaker notebook instances are based on EC2 instances running within AWS service accounts.
Why?
Amazon SageMaker does use EC2 instances, but they are not directly managed within the customer's AWS account.
Instead, these instances are provisioned within AWS-managed service accounts, which is why they do not appear within the customer's VPC or EC2 console.
The only way to access the underlying EBS volume is via SageMaker APIs, rather than the EC2 console.
upvoted 5 times

liquen14 5 months ago
Selected Answer: B
Although I'd go with Glue and option B I'm pretty sure that this is one of those "15 unscored questions that do not affect your score. AWS collects information about performance on these unscored questions to evaluate these questions for future use as scored questions"

Just for fun I asked perplexity, chatgpt, gemini, deepseek and claude: all gave D as first response

When I pointed out that "according to this https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html Kinesis can't convert directly cvs to parquet. It needs a Lambda" each model responded in a different way (some of them contradictory).

My reasoning is that D (Kinesis + Firehose) is incorrect because Firehose does not support direct CSV-to-Parquet conversion and needs a Lambda not mentioned in the option. But discussing about questions like this one is nothing but I big waste of time ;-P
upvoted 1 times

liquen14 4 months, 4 weeks ago
Forget about this please I posted this here incorrectly. This corresponds to Question 3. Apologies
upvoted 1 times

reginav 6 months, 4 weeks ago
Selected Answer: A
Amazon SageMaker notebook instances are indeed based on EC2 instances, but they are managed by the SageMaker service and do not appear as standard EC2 instances in the customer's VPC. Instead, they run in a managed environment that abstracts away the underlying EC2 instances, which is why the ML Specialist cannot see the instance in the VPC.
upvoted 1 times

Mickey321 9 months, 3 weeks ago
Selected Answer: C
The explanation for this choice is that Amazon SageMaker notebook instances are fully managed by AWS and run on EC2 instances that are not visible to customers. These EC2 instances are launched in AWS-owned accounts and are isolated from customer accounts by using AWS PrivateLink1. This means that customers cannot access or manage these EC2 instances directly, nor can they see the EBS volumes attached to them.
upvoted 1 times

loict 9 months, 3 weeks ago
Selected Answer: C
A. NO - AEC2 instances within the customer account are necessarily in a VPCb
B. NO - Amazon ECS service is not within customer accounts
C. YES - EC2 instances running within AWS service accounts are not visible to customer account
D. NO - SageMaker manages EC2 instance, not ECS
upvoted 6 times

ninomfr64 9 months, 3 weeks ago
Selected Answer: C
A. NO. If the EC2 instance of the notebook was in the customer account, customer would be able to see it. Also, "they run outside VPCs" isn't true as they run in service managed VPC or can be also attached to customer provided VPC -> https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/
B. NO, Notebooks are based on EC2 + EBS
C. YES -> https://aws.amazon.com/blogs/machine-learning/understanding-amazon-sagemaker-notebook-instance-networking-configurations-and-advanced-routing-options/
D. NO, Notebooks are based on EC2 + EBS

I also actually tested it in my account: I created a Notebook and attached it to my VPC, I was not able to see the EC2 instance behind the Notebook but I was able to see the its ENI with the following description "[Do not delete] Network Interface created to access resources in your VPC for SageMaker Notebook Instance ..."
upvoted 1 times

Rejju 1 year, 10 months ago

Selected Answer: A

already given below

upvoted 1 times

Rejju 1 year, 10 months ago

I am pretty sure the answer is A : Amazon SageMaker notebook instances are indeed based on EC2 instances, and these instances are within your AWS customer account. However, by default, SageMaker notebook instances run outside of your VPC (Virtual Private Cloud), which is why they may not be visible within your VPC. SageMaker instances are designed to be easily accessible for data science and machine learning tasks, which is why they typically do not reside within a VPC. If you need them to operate within a VPC, you can configure them accordingly, but this is not the default behavior.

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

I think it should be c

upvoted 1 times

ADVIT 2 years ago

Per https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html it's C

upvoted 1 times

BeCalm 2 years, 2 months ago

Selected Answer: C

Notebooks can run inside AWS managed VPC or customer managed VPC

upvoted 1 times

Maaayaaa 2 years, 3 months ago

Selected Answer: C

C, check the digram in https://docs.aws.amazon.com/sagemaker/latest/dg/studio-notebooks-and-internet-access.html

upvoted 5 times

oso0348 2 years, 3 months ago

Selected Answer: A

When a SageMaker notebook instance is launched in a VPC, it creates an Elastic Network Interface (ENI) in the subnet specified, but the underlying EC2 instance is not visible in the VPC. This is because the EC2 instance is managed by AWS, and it is outside of the VPC. The ENI acts as a bridge between the VPC and the notebook instance, allowing network connectivity between the notebook instance and other resources in the VPC. Therefore, the EBS volume of the notebook instance is also not visible in the VPC, and you cannot take a snapshot of the volume using VPC-based tools. Instead, you can create a snapshot of the EBS volume directly from the SageMaker console, AWS CLI, or SDKs.

upvoted 2 times

ZSun 2 years, 2 months ago

what you described is C
"This is because the EC2 instance is managed by AWS, and it is outside of the VPC."

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: C

Notebooks run inside a VPC not outside!

upvoted 1 times

krzyhoo 2 years, 4 months ago

Selected Answer: C

Definitely C

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 5 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 5
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Engineer needs to build a model using a dataset containing customer credit card information
How can the Data Engineer ensure the data remains encrypted and the credit card information is secure?

A. Use a custom encryption algorithm to encrypt the data and store the data on an Amazon SageMaker instance in a VPC. Use the SageMaker DeepAR algorithm to randomize the credit card numbers.

B. Use an IAM policy to encrypt the data on the Amazon S3 bucket and Amazon Kinesis to automatically discard credit card numbers and insert fake credit card numbers.

C. Use an Amazon SageMaker launch configuration to encrypt the data once it is copied to the SageMaker instance in a VPC. Use the SageMaker principal component analysis (PCA) algorithm to reduce the length of the credit card numbers.

D. Use AWS KMS to encrypt the data on Amazon S3 and Amazon SageMaker, and redact the credit card numbers from the customer data with AWS Glue. **Most Voted**

Hide Answer

Suggested Answer: *D*

Community vote distribution
D (100%)

by vetal at *Dec. 6, 2019, 10:45 a.m.*

Comments

vetal **Highly Voted** 3 years, 9 months ago
Why not D? When the data encrypted on S3 and SageMaker uses the same AWS KMS key it can use encrypted data there.
upvoted 36 times

WWODIN 3 years, 9 months ago
should be D
upvoted 12 times

zzeng 3 years, 8 months ago
Should be D.
Use Glue to do ETL to Hash the card number

upvoted 8 times

Antriksh 3 years, 9 months ago

Answer would be D

upvoted 9 times

cybe001 Highly Voted 3 years, 9 months ago

D is correct

upvoted 8 times

Ganshank Most Recent 4 months, 2 weeks ago

Selected Answer: D

https://aws.amazon.com/blogs/big-data/detect-and-process-sensitive-data-using-aws-glue-studio/
AWS Glue can be used for detecting and processing sensitive data.

upvoted 2 times

JonSno 4 months, 4 weeks ago

Selected Answer: D

Use AWS KMS for encryption and AWS Glue to redact credit card numbers
Reasoning:
AWS KMS (Key Management Service) encrypts data at rest in Amazon S3 and during processing in Amazon SageMaker.
AWS Glue can be used to redact sensitive data before processing, ensuring that credit card numbers are removed from datasets before being used for ML.
Complies with PCI DSS requirements for handling payment information securely.

upvoted 2 times

Mickey321 9 months, 3 weeks ago

Selected Answer: D

The reason for this choice is that AWS KMS is a service that allows you to easily create and manage encryption keys and control the use of encryption across a wide range of AWS services and in your applications1. By using AWS KMS, you can encrypt the data on Amazon S3, which is a durable, scalable, and secure object storage service2, and on Amazon SageMaker, which is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly3. This way, you can protect the data at rest and in transit.

upvoted 2 times

loict 1 year, 10 months ago

Selected Answer: D

A. NO - no need for custom encryption
B. NO - IAM Policies are not to encrypt
C. NO - launch configuration is not to encrypt
D. YES

upvoted 2 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: D

I think d is correct

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: D

It's D, KMS key can be used for encrypting the data at rest!

upvoted 4 times

ystotest 2 years, 7 months ago

Selected Answer: D

agreed with D

upvoted 3 times

jerto97 3 years, 8 months ago

IMHO, the problem with the question is that it is not clear whether the credit card number is used in the model. In that case discarding is never a good option. Hashing should be a safe option to keep it in the learning path

upvoted 1 times

cloud_trail 3 years, 8 months ago

It's gotta be D but C is a clever fake answer. Use PCA to reduce the length of the credit card number? That's a clever joke, as if reducing the length of a character string is the same as reducing dimensionality in a feature set.

upvoted 3 times

cnethers 3 years, 8 months ago

Can Glue do redaction?

upvoted 1 times

cloud_trail 3 years, 8 months ago

Just have the Glue job remove the credit card column.

upvoted 1 times

syu31svc 3 years, 8 months ago

Encryption on AWS can be done using KMS so D is the answer

upvoted 1 times

roytruong 3 years, 9 months ago

D is correct

upvoted 1 times

PRC 3 years, 9 months ago

D..KMS fully managed and other options are too whacky..

upvoted 4 times

AKT 3 years, 9 months ago

D is correct

upvoted 1 times

bhavesh0124 3 years, 9 months ago

Ans D is correct

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 2 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 2
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is designing a system for improving sales for a company. The objective is to use the large amount of information the company has on users' behavior and product preferences to predict which products users would like based on the users' similarity to other users.
What should the Specialist do to meet this objective?

A. Build a content-based filtering recommendation engine with Apache Spark ML on Amazon EMR

B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

C. Build a model-based filtering recommendation engine with Apache Spark ML on Amazon EMR

D. Build a combinative filtering recommendation engine with Apache Spark ML on Amazon EMR

Show Suggested Answer

by mlyu at *Jan. 2, 2020, 9:05 a.m.*

Comments

mlyu Highly Voted 3 years, 9 months ago

B
see https://en.wikipedia.org/wiki/Collaborative_filtering#Model-based

upvoted 21 times

kalyanvarma Highly Voted 3 years, 8 months ago

Content-based filtering relies on similarities between features of items, whereas colloborative-based filtering relies on preferences from other users and how they respond to similar items.

upvoted 14 times

Manju_Bn Most Recent 9 months ago

Answer is B : Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.
Collaborative filtering focuses on user behavior and preferences therefore it is perfect for predicting products based on user similarities.

upvoted 2 times

AjoseO 9 months, 3 weeks ago

Selected Answer: B

B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

Collaborative filtering is a technique used to recommend products to users based on their similarity to other users. It is a widely used method for building recommendation engines. Apache Spark ML is a distributed machine learning library that provides scalable implementations of collaborative filtering algorithms. Amazon EMR is a managed cluster platform that provides easy access to Apache Spark and other distributed computing frameworks.

upvoted 1 times

solution123 9 months, 3 weeks ago

Selected Answer: B

Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR. (TRUE)

Collaborative filtering is a commonly used method for recommendation systems that aims to predict the preferences of a user based on the behavior of similar users. In the case described, the objective is to use users' behavior and product preferences to predict which products they want, making collaborative filtering a good fit.

Apache Spark ML is a machine learning library that provides scalable, efficient algorithms for building recommendation systems, while Amazon EMR provides a cloud-based platform for running Spark applications.
You can find more detail in https://www.udemy.com/course/aws-certified-machine-learning-specialty-2023

upvoted 2 times

ychaabane 9 months, 4 weeks ago

Selected Answer: B

collaborative filtering

upvoted 1 times

james2033 1 year, 4 months ago

Selected Answer: B

'Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users.'

Source: https://realpython.com/build-recommendation-engine-collaborative-filtering/#what-is-collaborative-filtering

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

A. NO - content-based filtering looks at similarities with items the user already looked at, not activities of other users
B. YES - state of the art
C. NO - too generic terms, everything is a model
D. NO - combinative filtering does not exist

upvoted 4 times

Mickey321 1 year, 11 months ago

Selected Answer: B

Collaborative filtering is a technique used by recommendation engines to make predictions about the interests of a user by collecting preferences or taste information from many users. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

upvoted 2 times

Mickey321 1 year, 11 months ago

Selected Answer: B

B. Build a collaborative filtering recommendation engine with Apache Spark ML on Amazon EMR.

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: B

I think it should be b

upvoted 1 times

brunokiyoshi 2 years, 3 months ago

Selected Answer: B

Content-based recommendations rely on product similarity. If a user likes a product, products that are similar to that one will be recommended. Collaborative recommendations are based on user similarity. If you and other users have given similar reviews to a range of products, the model assumes it is likely that other products those other people have liked but that you haven't purchased should be a good recommendation for you.

upvoted 4 times

dreswardev 2 years, 6 months ago

feature engineering is required, use model based

upvoted 1 times

ryuhei 2 years, 9 months ago

Selected Answer: B

Answer is "B"

upvoted 1 times

roytruong 3 years, 8 months ago

go for B

upvoted 2 times

cybe001 3 years, 8 months ago

B is correct
https://aws.amazon.com/blogs/big-data/building-a-recommendation-engine-with-spark-ml-on-amazon-emr-using-zeppelin/

upvoted 6 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 4 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 4
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A city wants to monitor its air quality to address the consequences of air pollution. A Machine Learning Specialist needs to forecast the air quality in parts per million of contaminates for the next 2 days in the city. As this is a prototype, only daily data from the last year is available.
Which model is MOST likely to provide the best results in Amazon SageMaker?

A. Use the Amazon SageMaker k-Nearest-Neighbors (kNN) algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.

B. Use Amazon SageMaker Random Cut Forest (RCF) on the single time series consisting of the full year of data.

C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.

D. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of classifier.

Show Suggested Answer

by ozan11 at *Jan. 20, 2020, 1:36 p.m.*

Comments

ozan11 Highly Voted 3 years, 9 months ago
answer should be C
upvoted 17 times

roytruong Highly Voted 3 years, 9 months ago
go for C
upvoted 6 times

robctsgps Most Recent 2 months, 1 week ago
Selected Answer: C
kNN is not specifically designed for time series forecasting. The best choice is C. Use the Amazon SageMaker Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
upvoted 1 times

JonSno 4 months, 4 weeks ago
Selected Answer: C

Amazon SageMaker Linear Learner (Regressor)
Why?
The Linear Learner algorithm can be used for time series regression.
Using predictor_type=regressor, it learns trends and patterns in historical data and extrapolates future values.
Given limited historical data (only 1 year), a simple linear regression model might perform well as a baseline.
While deep learning models (like Amazon Forecast) may be more advanced, Linear Learner is easier to implement and train for a prototype.

upvoted 2 times

loict 9 months, 3 weeks ago

Selected Answer: C

A. NO - kNN is not forecasting, it is similarities
B. NO - RCF is for anomality detection
C. YES - Linear Regression good for forecasting
D. NO - we don't want to classify

upvoted 3 times

Mickey321 9 months, 3 weeks ago

Selected Answer: C

The reason for this choice is that the Linear Learner algorithm is a versatile algorithm that can be used for both regression and classification tasks1. Regression is a type of supervised learning that predicts a continuous numeric value, such as the air quality in parts per million2. The predictor_type parameter specifies whether the algorithm should perform regression or classification3. Since the goal is to forecast a numeric value, the predictor_type should be set to regressor.

upvoted 3 times

ninomfr64 1 year ago

Selected Answer: D

A. Managing Kafka on EC2 is not compatible with least effort requirement
B. Doable (in 2024) as Glue supports streaming ETL to consumes streams and supports CSV records -> https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html
C. Managing an EMR cluster imo is no compatible with least effort requirement
D. Firehose supports kinesis data stream as source and it can use lambda to convert CSV records into parquet -> https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html

I guess this is a bit old question, pre Glue streaming ETL support (2023) -> https://aws.amazon.com/about-aws/whats-new/2023/03/aws-glue-4-0-streaming-etl/

Thus I'll go for D

upvoted 1 times

LocalHero 1 year, 7 months ago

This blog wrote Japanese.
but its said using LinearLearner for air pollution prediction.
https://aws.amazon.com/jp/blogs/news/build-a-model-to-predict-the-impact-of-weather-on-urban-air-quality-using-amazon-sagemaker/

upvoted 2 times

jyrajan69 1 year, 11 months ago

The HyperParameter is . Either “binary_classifier” or “multiclass_classifier” or “regressor”., there is no classifier so the answer is C

upvoted 1 times

Venkatesh_Babu 1 year, 11 months ago

Selected Answer: C

Ans should be c

upvoted 1 times

ortamina 2 years ago

a kNN will require a large value of k to avoid overfitting and we only have 1 year's worth of data - kNNs also face a difficult time extrapolating if the air quality series contains a trend

If we had assurances there is no trend in the air quality series (no extrapolation), and we had enough data, then kNN should beat a linear model ... I am inclined to go for C just going off of the cue that "only daily data from last year is available"

upvoted 1 times

ninomfr64 1 year ago

Agree with you analysis, to further expand it: we don't have info about dataset features based on "only daily data from last year is available" this let me think we could be in a situation where our dataset is made up by timestamp and pollution_value so KNN would be pretty useless in this situation.

upvoted 1 times

brunokiyoshi 2 years, 3 months ago

Selected Answer: C

Random cut forests in timeseries are used for anomaly detection, and not for forecasting. KNN's are classification algorithms. You would use the Linear Learner as a regressor, since forecasting falls into the domain of regression.

upvoted 3 times

brunokiyoshi 2 years, 3 months ago

I mean, you could use KNN's for regression, but for forecasting I don't think so

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: C

KNN isn't for time series predicting, go for A!

upvoted 2 times

Valcilio 2 years, 4 months ago

Im sorry, I wanted to say go for C!

upvoted 2 times

rockyykrish 2 years, 4 months ago

Creating a machine learning model to predict air quality
To start small, we will follow the second approach, where we will build a model that will predict the NO2 concentration of any given day based on wind speed, wind direction, maximum temperature, pressure values of that day, and the NO2 concentration of the previous day. For this we will use the Linear Learner algorithm provided in Amazon SageMaker, enabling us to quickly build a model with minimal work.

Our model will consist of taking all of the variables in our dataset and using them as features of the Linear Learner algorithm available in Amazon SageMaker

upvoted 1 times

Ajose0 2 years, 5 months ago

Selected Answer: A

Answer should be A.

k-Nearest-Neighbors (kNN) algorithm will provide the best results for this use case as it is a good fit for time series data, especially for predicting continuous values. The predictor_type of regressor is also appropriate for this task, as the goal is to forecast a continuous value (air quality in parts per million of contaminants). The other options are also viable, but may not provide as good of results as the kNN algorithm, especially with limited data.

using the Amazon SageMaker Linear Learner algorithm with a predictor_type of regressor, may still provide reasonable results, but it assumes a linear relationship between the input features and the target variable (air quality), which may not always hold in practice, especially with complex time series data. In such cases, non-linear models like kNN may perform better. Furthermore, the kNN algorithm can handle irregular patterns in the data, which may be present in the air quality data, and provide more accurate predictions.

upvoted 3 times

ryuhei 2 years, 9 months ago

Selected Answer: C

Answer is "C" !!!

upvoted 1 times

yemauricio 2 years, 10 months ago

answer C

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 3 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 3

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Mobile Network Operator is building an analytics platform to analyze and optimize a company's operations using Amazon Athena and Amazon S3.

The source systems send data in .CSV format in real time. The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3.

Which solution takes the LEAST effort to implement?

A. Ingest .CSV data using Apache Kafka Streams on Amazon EC2 instances and use Kafka Connect S3 to serialize data as Parquet

B. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Glue to convert data into Parquet.

C. Ingest .CSV data using Apache Spark Structured Streaming in an Amazon EMR cluster and use Apache Spark to convert data into Parquet.

D. Ingest .CSV data from Amazon Kinesis Data Streams and use Amazon Kinesis Data Firehose to convert data into Parquet.

Show Suggested Answer

by DonaldCMLIN at *Nov. 16, 2019, 3:04 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 9 months ago

Answer is B

upvoted 31 times

Antriksh 3 years, 9 months ago

you cannot use AWS glue for streaming data. Clearly B is incorrect.

upvoted 3 times

scuzzy2010 3 years, 9 months ago

Even if the exam's answer is based on solution before AWS implemented the capability of AWS glue to process streaming data, this answer is still correct as Kinesis would output the data to S3 and Glue will pick it up from there and covert to parquet. Question does not say data must be converted to parquet in real time, it only says the csv data is received as a stream in real time.

upvoted 2 times

GeeBeeEl 3 years, 8 months ago

Actually question says "The source systems send data in CSV format in real time The Data Engineering team wants to transform the data to the Apache Parquet format before storing it on Amazon S3" same as saying data must be converted real time

upvoted 5 times

zzeng 3 years, 9 months ago

AWS Glue can do it now (2020 May)
https://aws.amazon.com/jp/blogs/news/new-serverless-streaming-etl-with-aws-glue/

upvoted 6 times

hamimelon 2 years, 6 months ago

This link is in Japanese

upvoted 3 times

OmarSaadEldien 3 years, 8 months ago

the Approve Of B
https://aws.amazon.com/blogs/aws/new-serverless-streaming-etl-with-aws-glue/

upvoted 7 times

vetal Highly Voted 3 years, 9 months ago

D is wrong as kinesis firehose can convert from JSON to parquet but here we have CSV.
B is correct and here is another proof link: https://medium.com/searce/convert-csv-json-files-to-apache-parquet-using-aws-glue-a760d177b45f

upvoted 24 times

zzeng 3 years, 9 months ago

https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html
You are right.
https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html
If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first

upvoted 8 times

samy666 3 years, 2 months ago

But there is no Lambda in D

upvoted 2 times

AdolinKholin 2 years, 9 months ago

But there's a D in Lambda

upvoted 3 times

daveclear Most Recent 2 months, 2 weeks ago

Selected Answer: B

Answer is B
- Firehose cannot convert from csv to parquet without a lambda: https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html
- Glue can handle streaming data: https://docs.aws.amazon.com/glue/latest/dg/add-job-streaming.html

upvoted 2 times

LalBSingh 4 months, 3 weeks ago

Selected Answer: D

Kinesis Data Firehose supports real-time streaming ingestion and can automatically convert CSV to Parquet before storing it in S3.

upvoted 3 times

daveclear 2 months, 2 weeks ago

It requires a lambda to go from csv to parquet

upvoted 1 times

JonSno 4 months, 4 weeks ago

Selected Answer: D

Amazon Kinesis Data Streams + Amazon Kinesis Data Firehose
Effort: Lowest effort
Why?
Amazon Kinesis Data Firehose natively supports real-time CSV ingestion and automatic conversion to Parquet.
Fully managed, serverless, and directly integrates with Amazon S3.
Requires zero infrastructure management compared to other solutions.

upvoted 1 times

JonSno 4 months, 4 weeks ago

I take this back .. ans shd be B.. on researching further it is JSON or ORC to Parque that KDS supports.. So answer is B - not optimal but close to suitable
. Amazon Kinesis Data Streams + AWS Glue AWS Glue can batch-process CSV and convert it to Parquet for S3. However, Glue is batch-oriented, not real-time.

upvoted 1 times

liquen14 4 months, 4 weeks ago

Selected Answer: B

Although I'd go with Glue and option B I'm pretty sure that this is one of those "15 unscored questions that do not affect your score. AWS collects information about performance on these unscored questions to evaluate these questions for future use as scored questions"

Just for fun I asked perplexity, chatgpt, gemini, deepseek and claude: all gave D as first response

When I pointed out that "according to this https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html Kinesis can't convert directly cvs to parquet. It needs a Lambda" each model responded in a different way (some of them contradictory).

My reasoning is that D (Kinesis + Firehose) is incorrect because Firehose does not support direct CSV-to-Parquet conversion and needs a Lambda not mentioned in the option. But discussing about questions like this one is nothing but I big waste of time ;-P

upvoted 2 times

AbimbolaOlaniran 6 months, 3 weeks ago

Selected Answer: D

D
Kinesis Data Firehose is designed specifically for streaming data delivery to destinations like S3. It has built-in support for data format conversion, including CSV to Parquet. This eliminates the need for managing separate transformation services like Glue or Spark. The setup is significantly simpler: you configure a Firehose delivery stream, specify the data format conversion, and point it to your S3 bucket.

Therefore, option D requires the least implementation effort because it leverages a fully managed service (Kinesis Data Firehose) with built-in functionality for data format conversion.

upvoted 1 times

venksters 6 months, 4 weeks ago

Selected Answer: B

Amazon Kinesis Data Firehose can only convert from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3.

upvoted 1 times

TinTinAWS 9 months, 2 weeks ago

Answer B,
Yes, Amazon Kinesis Data Firehose can convert CSV to Apache Parquet, but you need to use a Lambda function to transform the CSV to JSON first: here the question is least effort to build, so B is the right answer with least effort to build the solution

upvoted 1 times

Keya 9 months, 3 weeks ago

Selected Answer: B

Use Amazon Kinesis Data Streams to ingest customer data and configure a Kinesis Data Firehose delivery stream as a consumer to convert the data into Apache Parquet is incorrect. Although this could be a valid solution, it entails more development effort as Kinesis Data Firehose does not support converting CSV files directly into Apache Parquet, unlike JSON.

upvoted 2 times

geoan13 9 months, 3 weeks ago

Selected Answer: B

Amazon Kinesis Data Firehose can convert the format of your input data from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3. Parquet and ORC are columnar data formats that save space and enable faster queries compared to row-oriented formats like JSON. If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first.
https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html

upvoted 1 times

rav009 1 year, 2 months ago

Selected Answer: D

Between B and D chose D.
Because Firehose can't handle csv directly.

upvoted 1 times

rav009 1 year, 2 months ago

Between B and D chose B.
Because Firehose can't handle csv directly.

upvoted 1 times

s_k_aws 1 year, 3 months ago

Answer is B.
https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html
"If you want to convert an input format other than JSON, such as comma-separated values (CSV) or structured text, you can use AWS Lambda to transform it to JSON first."

upvoted 1 times

chewasa 1 year, 4 months ago

Selected Answer: B

u need glue to convert to parquet

upvoted 1 times

0c47783 1 year, 4 months ago

D for sure, Firehose can convert csv to parquet

upvoted 3 times

vkbajoria 1 year, 4 months ago

Answer is unfortunately B. firehose cannot convert coma separated CSV to parquet directly.

upvoted 1 times

kyuhuck 1 year, 5 months ago

Selected Answer: D

b is not goog but - >given the context of "finding the solution that requires the least effort to implement," option D is the most suitable choice. Ingesting data from Amazon Kinesis Data Streams and using Amazon Kinesis Data Firehose to convert the data to Parquet format is a serverless approach. It allows for automatic data transformation and storage in Amazon S3 without the need for additional development or management of data conversion logic. Therefore, under the given conditions, option D is considered the solution that requires the "least effort" to implement

upvoted 3 times

shammous 11 months, 1 week ago

Kinesis Data Firehose doesn't convert anything, it rather calls a lambda function to do so which is the overhead we want to avoid. B is the correct answer.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 51 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 51

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A manufacturing company has a large set of labeled historical sales data. The manufacturer would like to predict how many units of a particular part should be produced each quarter.

Which machine learning approach should be used to solve this problem?

A. Logistic regression

B. Random Cut Forest (RCF)

C. Principal component analysis (PCA)

D. Linear regression

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 3:06 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

HOW MANY/MUCH, THOSE ARE REGRESSION TOPIC,
LOGISTIC FOR 0/1,YES/NO

https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/regression-model-insights.html

THE ANSWER SHOULD BE D.

upvoted 62 times

rsimham 3 years, 3 months ago

agree. RCF is mostly used for anomaly detection or separate outliers

upvoted 10 times

syu31svc Highly Voted 3 years, 3 months ago

Amazon SageMaker Random Cut Forest (RCF) is an unsupervised algorithm for detecting anomalous data points within a data set
Answer is D 100%

upvoted 10 times

JonSno Most Recent 4 months, 3 weeks ago

Selected Answer: D

The problem involves predicting the number of units to be produced each quarter based on historical sales data. This is a continuous numerical prediction, making it a regression problem.

Linear regression is ideal for forecasting when there is a linear relationship between input variables (e.g., past sales, seasonal trends) and the target variable (units to be produced).
It helps model the relationship between past sales and future demand.
If there are seasonal effects, a time-series model (like ARIMA or Prophet) could be considered as well.

upvoted 1 times

t47 9 months, 1 week ago

D should be the answer

upvoted 1 times

endeesa 1 year, 1 month ago

Selected Answer: D

How many units should give this away as Linear regression

upvoted 1 times

AmeeraM 1 year, 3 months ago

Selected Answer: D

I do not see any hint of anomalies here, we are looking for a number to be predicted, this seems to be the reason of the correct answer
https://docs.aws.amazon.com/quicksight/latest/user/how-does-rcf-generate-forecasts.html

upvoted 1 times

DavidRou 1 year, 4 months ago

Selected Answer: D

How can the right answer be B? That Random Cut Forest is an algorithm written for anomaly detection.

upvoted 3 times

Mickey321 1 year, 4 months ago

Selected Answer: D

option D

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: D

D is the correct. B is for outlier detection only.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: D

It sounds like Linear regression problem and Random Cut is more known for anomaly detection while it can do other types of ML. The answer seems to be strange with no explanation.

upvoted 1 times

jackzhao 1 year, 10 months ago

D is correct!

upvoted 1 times

oso0348 1 year, 10 months ago

Selected Answer: D

D. Linear regression would be the appropriate machine learning approach to solve this problem of predicting the number of units of a particular part to be produced each quarter. Linear regression is a supervised learning algorithm used for predicting continuous variables based on input features. In this case, the historical sales data can be used as input features, and the number of units produced each quarter can be used as the continuous target variable.

upvoted 2 times

Nadia0012 1 year, 10 months ago

Selected Answer: D

definitely D.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: D

This is a regression problem where the goal is to predict a continuous outcome, which in this case is the number of units of a particular part that should be produced each quarter. Linear regression is a simple and commonly used approach to solve such problems, where a linear relationship is established between the independent variables (e.g., historical sales data) and the dependent variable (e.g., number of units of a part to be produced).

upvoted 2 times

Tomatoteacher 1 year, 12 months ago

Selected Answer: D

D, RCF answers here just link one article where RCF is implemented to find outliers in time series, or are able to deduce trends, but here they mention already labelled data, RCF is unsupervised, so that data would go to waste.

upvoted 1 times

hamimelon 2 years ago

Honestly, i think these are all bad answers. It should be time series modeling methods.

upvoted 2 times

Peeking 2 years, 1 month ago

Selected Answer: D

The answer is D. B is out of it as RCF is used for anomaly detection. Logistic Regression is for Classification mainly. Only linear regression can be used if Time series algorithms are not part of the options.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 100 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 100
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist is developing a binary classifier to predict whether a patient has a particular disease on a series of test results. The Data Scientist has data on
400 patients randomly selected from the population. The disease is seen in 3% of the population.
Which cross-validation strategy should the Data Scientist adopt?

A. A k-fold cross-validation strategy with k=5

B. A stratified k-fold cross-validation strategy with k=5

C. A k-fold cross-validation strategy with k=5 and 3 repeats

D. An 80/20 stratified split between training and validation

Show Suggested Answer

by scuzzy2010 at *Feb. 19, 2021, 2:26 a.m.*

Comments

scuzzy2010 Highly Voted 3 years, 9 months ago

B - stratified k-fold cross-validation will enforce the class distribution in each split of the data to match the distribution in the complete training dataset.

upvoted 16 times

SophieSu Highly Voted 3 years, 9 months ago

B is the correct answer. Use Stratified k-Fold Cross-Validation for Imbalanced Classification. Stratified train/test splits is an option too. But the question is specifically asking "cross-validation" strategy.

upvoted 9 times

MultiCloudIronMan Most Recent 8 months, 3 weeks ago

Selected Answer: B

In summary, Option B is the most appropriate strategy for handling the imbalanced dataset and ensuring reliable performance metrics for the binary classifier.

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

for imbalanced data. Stratified k-fold cross-validation ensures that the distribution of the target variable is the same in each fold. This is important for binary classification problems, where the target variable is imbalanced. In this case, the disease is seen in only 3% of the population. This means that if we do not use stratified k-fold cross-validation, then there is a risk that the training and validation sets will not be representative of the actual population.

upvoted 1 times

ADVIT 2 years ago

B
https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

Stratified cross validation is for unbalanced data like this!

upvoted 1 times

AWS__Newbie 3 years, 8 months ago

Why K=5?

upvoted 2 times

eeah 3 years, 3 months ago

K=5 is just standard

upvoted 1 times

Vita_Rasta84444 3 years, 8 months ago

Yes, B...

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 99 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 99

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A company uses camera images of the tops of items displayed on store shelves to determine which items were removed and which ones still remain. After several hours of data labeling, the company has a total of 1,000 hand-labeled images covering 10 distinct items. The training results were poor.

Which machine learning approach fulfills the company's long-term needs?

A. Convert the images to grayscale and retrain the model

B. Reduce the number of distinct items from 10 to 2, build the model, and iterate

C. Attach different colored labels to each item, take the images again, and build the model

D. Augment training data for each item using image variants like inversions and translations, build the model, and iterate.

Show Suggested Answer

by lstdanagan at *April 21, 2022, 5:10 a.m.*

Comments

ovokpus Highly Voted 2 years ago

Selected Answer: D

Data Augumentation is the way to go here.

How does converting to grayscale help? What if the colors of the items are relevant in object identification???

upvoted 11 times

AmeeraM Most Recent 9 months ago

Selected Answer: D

data augemntation

upvoted 1 times

jopaca1216 10 months ago

D is correct

How can I make the decision to use gray images if the question doesn't even indicate whether the images are colored or not? and even so, colored images are important to ensure more accuracy in training than compared to gray imagens.

Due that the model is underfitting, more data like indicated the option D is the correct action.

upvoted 1 times

mirik 1 year ago

C: "Attach different colored labels to each item, take the images again, and build the model"
It is also kind of augmentation. It is even better than just inverting and translating existing samples.

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

But it's done in real life and your manual work would be lost.

upvoted 1 times

Debayandt91 1 year, 2 months ago

shouldnt it be reduced to 2 variables , taking image of empty shelf and non empty and that should do it ?

upvoted 1 times

Sylzys 1 year, 4 months ago

D is of course the right answer, grayscale only won't help anything

upvoted 3 times

PHTR 1 year, 6 months ago

D is the CORRECT ANSWER

https://research.aimultiple.com/data-augmentation/

upvoted 1 times

aScientist 1 year, 8 months ago

Selected Answer: D

Data augmentation is correct. we need more samples

upvoted 1 times

tgaos 2 years, 1 month ago

D is correct

upvoted 3 times

cron0001 2 years, 2 months ago

Selected Answer: D

D is my answer for this. A can help but it'll need more than that.

upvoted 4 times

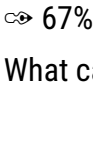
lstdanagan 2 years, 2 months ago

Selected Answer: D

D, i guess

upvoted 4 times

Amazon Discussions



Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 98 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 98
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A machine learning specialist works for a fruit processing company and needs to build a system that categorizes apples into three types. The specialist has collected a dataset that contains 150 images for each type of apple and applied transfer learning on a neural network that was pretrained on ImageNet with this dataset.
The company requires at least 85% accuracy to make use of the model.
After an exhaustive grid search, the optimal hyperparameters produced the following:
- 68% accuracy on the training set
- 67% accuracy on the validation set

What can the machine learning specialist do to improve the system's accuracy?

A. Upload the model to an Amazon SageMaker notebook instance and use the Amazon SageMaker HPO feature to optimize the model's hyperparameters.

B. Add more data to the training set and retrain the model using transfer learning to reduce the bias.

C. Use a neural network model with more layers that are pretrained on ImageNet and apply transfer learning to increase the variance.

D. Train a new model using the current neural network architecture.

Show Suggested Answer

by lstdanagan at *April 21, 2022, 5:09 a.m.*

Comments

dolorez Highly Voted 2 years, 1 month ago

Selected Answer: B

the answer is B - the model is underfitting = high bias, so we want to reduce it
C is wrong because the intention is not to increase variance which equals overfitting (using a more complex model would be good, but to reduce bias not increase variance)

upvoted 11 times

CloudGyan Most Recent 6 months ago

Selected Answer: B

The 68% accuracy on the training set and 67% accuracy on the validation set suggest that the model is biased - underfitting and does not have enough capacity or relevant information to learn the underlying patterns in the data.

upvoted 1 times

endeesa 7 months, 2 weeks ago

Selected Answer: B

I would think ImageNet network is good enough already, so more data

upvoted 1 times

loict 10 months ago

Selected Answer: B

A. NO - HPO has already been done though grid search
B. YES - 150 images is very small; need x10 that
C. NO - need bigger training set
D. NO - what would the new model be ?

upvoted 2 times

Mickey321 10 months, 2 weeks ago

Selected Answer: B

More data to training set

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: B

Letter B is the correct one. We can add more data with data augmentation. Letter A would be a repetition of what has already been done. Letter C is impractical. Letter D is starting from scratch without need.

upvoted 1 times

mirik 1 year ago

Selected Answer: D

I think it should be D: "Train a new model using the current neural network architecture".
Because apples data is very specific and ImageNet weights will be to generic there. We still can leave ImageNet weights for an initial configuration but the model should be retrained from scratch.

upvoted 1 times

cox1960 1 year, 2 months ago

Selected Answer: A

450 images should be fine. HPO for me.

upvoted 1 times

expertguru 1 year, 6 months ago

bOTH VALIdation set and train set performing equally but performance not good. So the basic problem here is high bias (train error) and high variance (test error). Ideally we want both low, but there is trade-off need to be cautious to avoid overfitting. So this problem needs solution for Low bias first (so training performance improves with decent) for later to figure out whether that leads to overfit or not when you test it,! Answer choice B

upvoted 1 times

deng113jie 1 year, 11 months ago

why not A?
https://aws.amazon.com/about-aws/whats-new/2022/07/amazon-sagemaker-automatic-model-tuning-supports-increased-limits-improve-accuracy-models/

upvoted 2 times

edvardo 2 years, 1 month ago

Given that the model can't even fit the training set properly, it would be convenient to amplify the layers that are trained. If I understood the phrasing correctly, I would go with C.

upvoted 1 times

lstdanagan 2 years, 2 months ago

Selected Answer: C

C, accuracy on training set is low, model not complex enough

upvoted 1 times

spaceexplorer 2 years, 2 months ago

B is more accurate, while adding more complexity for model is viable but you don't want to increase variance

upvoted 12 times

NeverMinda 2 years, 1 month ago

It only has 150 photos for training, more complex neural network won't help

upvoted 4 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 97 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 97
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A large company has developed a BI application that generates reports and dashboards using data collected from various operational metrics. The company wants to provide executives with an enhanced experience so they can use natural language to get data from the reports. The company wants the executives to be able ask questions using written and spoken interfaces.
Which combination of services can be used to build this conversational interface? (Choose three.)

A. Alexa for Business

B. Amazon Connect

C. Amazon Lex

D. Amazon Polly

E. Amazon Comprehend

F. Amazon Transcribe

Show Suggested Answer

by [deleted] at *Feb. 4, 2021, 6:33 a.m.*

Comments

astonm13 Highly Voted 3 years, 9 months ago

C - voice and text interface
E - understanding
F - Speech to text

upvoted 35 times

hero67 3 years, 8 months ago

Why would I need to transcribe while I have Lex that do the NLU part? It would be more reasonable to select Either Connect (B) or Polly (D) if the specs to generate output speech.

upvoted 4 times

Hariru 3 years, 8 months ago

E - is more to express the "feeling" or "mood". We would rather need something, that can speak to the customer. So my suggestion is c,d,f

upvoted 5 times

F1Fan 1 year, 2 months ago

The question states that the company wants to "provide executives with an enhanced experience so they can use natural language to get data from the reports." The key phrase here is "use natural language," which implies that the executives will be interacting with the system using human-like language, either written or spoken. To understand and interpret natural language inputs from users, whether written or spoken, the system needs to have natural language understanding (NLU) or natural language processing (NLP) capabilities. Without NLU/NLP capabilities, the system would not be able to make sense of the executives' natural language queries and extract the relevant information to retrieve data from the reports and dashboards. Services like Amazon Lex and Amazon Comprehend are specifically designed to provide NLU and NLP functionalities, respectively. Amazon Lex uses NLU models to understand the intent and extract relevant information from user inputs, while Amazon Comprehend provides NLP capabilities to analyze and extract insights from text data.

upvoted 1 times

eganilovic Highly Voted 3 years, 9 months ago

If we need to build written and spoken interfaces we need :
F - Transcribe (speech to text)
D- Polly (text ot speech)
And for chatbot:
E - Lex

upvoted 23 times

eganilovic 3 years, 9 months ago

*C - Lex

So C,D,F

upvoted 17 times

weelz 3 years, 8 months ago

I second that, the keyword here is "conversational interface". so, no conversation without Amazon Lex

upvoted 1 times

sheetalconect Most Recent 1 year ago

Selected Answer: ACD

Alexa for Business: Handles the voice interaction, converting spoken queries into text and providing the voice interface that executives use to interact with the BI application.
Amazon Lex: Processes the text input (converted by Alexa) and understands the intent behind the queries, enabling the conversational interface.
Amazon Polly: Optional but useful if you want to convert the textual responses from the BI application back into spoken responses, providing a complete voice-based interaction.

upvoted 1 times

ArchMelody 1 year, 4 months ago

Selected Answer: CDF

Lex for bot service, Polly for text-to-speech (answer) and Transcribe for speech-to-text (question).

upvoted 2 times

vkbajoria 1 year, 4 months ago

I believe Answer should be CDF
C: Lex
D: Polly
F: Transcribe

upvoted 1 times

kyuhuck 1 year, 5 months ago

Selected Answer: CDF

For a BI application where executives can ask questions using written and spoken interfaces, the following combination of services would be suitable:

Amazon Lex (Option C): To build the core conversational interface that understands and processes natural language queries.
Amazon Polly (Option D): To provide spoken responses to written queries, giving a more interactive experience for users who are not using the voice interface.
Amazon Transcribe (Option F): To convert spoken queries into text that can be understood by Amazon Lex.
These three services would work together to provide a comprehensive conversational interface that allows for both text and voice interactions, meeting the requirements of the scenario provided.

upvoted 2 times

Alice1234 1 year, 5 months ago

C. Amazon Lex: It provides advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, enabling you to build applications with highly engaging user experiences and lifelike conversational interactions.

D. Amazon Polly: This service turns text into lifelike speech using deep learning. It would enable the BI application to deliver the answers to the executives' questions in a spoken format.

F. Amazon Transcribe: This is an automatic speech recognition (ASR) service that makes it easy for developers to add speech-to-text capability to their applications. This would be necessary for the BI application to interpret spoken questions from the executives.

upvoted 1 times

CloudHandsOn 1 year, 5 months ago

Selected Answer: CDF

CDF -> CEF. you dont need comprehend in this scenario.

upvoted 3 times

CloudHandsOn 1 year, 6 months ago

Selected Answer: CDF

Amazon Lex (C): This service is crucial for building conversational interfaces. It provides the capabilities to understand and interpret user input in natural language, which is essential for understanding the questions asked by executives.

Amazon Transcribe (F): For a spoken interface, you need a service that can convert speech into text. Amazon Transcribe does exactly this, allowing the system to process spoken questions by converting them into text that can then be interpreted by Amazon Lex.

Amazon Polly (D): To enhance the user experience by responding to inquiries not only in text but also in spoken form, Amazon Polly is ideal. It converts text responses into lifelike speech, allowing the system to verbally communicate with the executives.

Together, these three services (Amazon Lex, Amazon Transcribe, and Amazon Polly) will enable a comprehensive conversational interface for the BI application, catering to both written and spoken queries and responses

upvoted 2 times

endeesa 1 year, 7 months ago

Selected Answer: CDF

why does aws use muliplt service for tts and stt?

upvoted 3 times

sukye 1 year, 7 months ago

Selected Answer: CDF

No, don't need E Comprehend because the report has already been generated.

upvoted 2 times

akgarg00 1 year, 7 months ago

Answer is CEF --> Input can be speech but the output to the user will be text (as nothing specific is mentioned) using Lex for conversational interface, Transcribe to convert speech to text (if input is speech) and Comprehend for insights from text

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: CEF

CEF is correct

upvoted 1 times

DimLam 1 year, 8 months ago

Selected Answer: CDE

I will go with:
lex for the chat interface
comprehend for getting insights from reports
Polly for text-to-speech transformation

https://aws.amazon.com/blogs/machine-learning/deriving-conversational-insights-from-invoices-with-amazon-textract-amazon-comprehend-and-amazon-lex/

upvoted 1 times

jopaca1216 1 year, 10 months ago

Amazon Polly is essential for providing spoken responses in a conversational interface, it doesn't directly handle the natural language understanding and processing aspect, which is why it wasn't included as one of the top three services for building the conversational interface in this scenario.

Correct is C, E, F

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: CDF

A. NO - Alexa for Business
B. NO - Amazon Connect for call centers
C. YES - Amazon Lex for chatbots
D. YES - Lex Text-to-Speech
E. NO - Amazon Comprehend is for topic extraction and sentiment analysis, Transcribe already does it
F. YES - Transcribe Speech-to-Text

upvoted 4 times

AmeeraM 1 year, 9 months ago

Transcribe does not do sentiment analysis and topic extraction it just generates transcript from speech so we need Amazon Comprehend

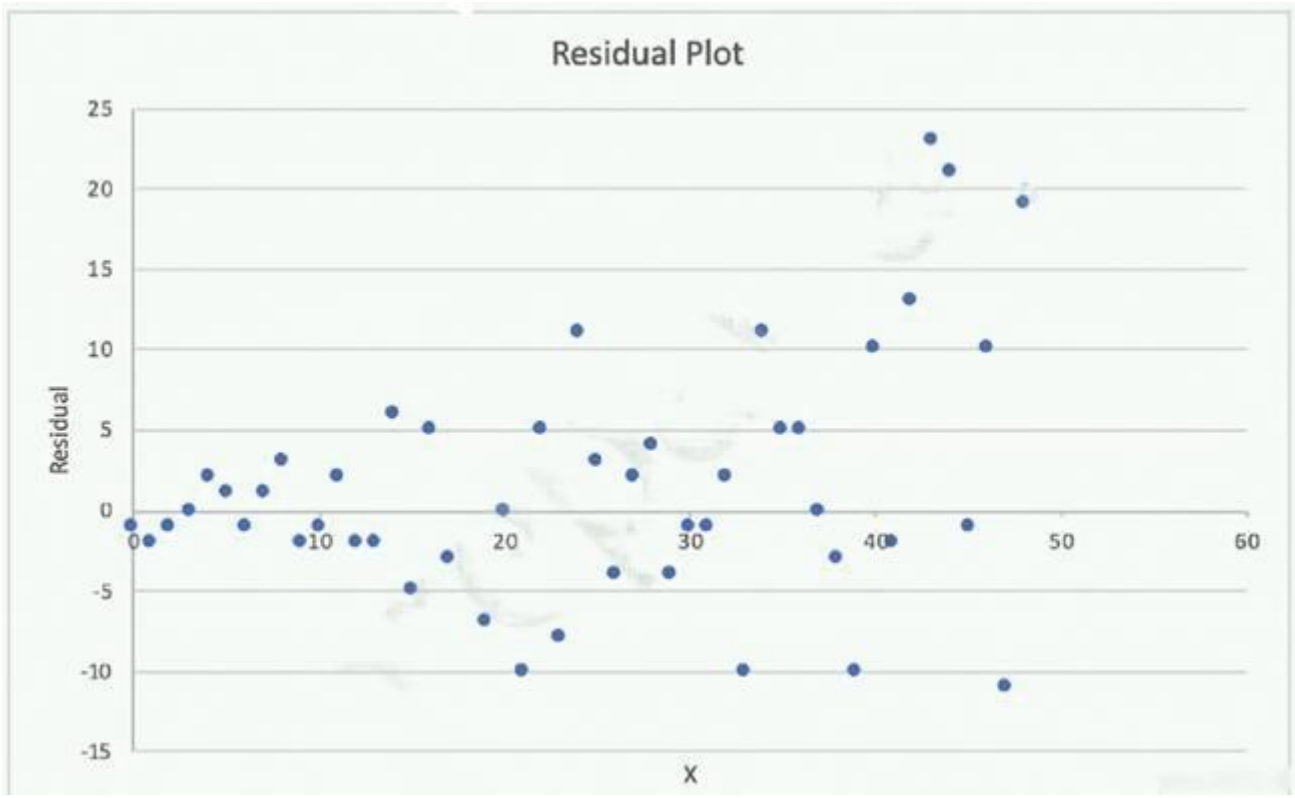
upvoted 1 times

Mickey321 1 year, 10 months ago

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 96 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 96
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is attempting to build a linear regression model.



Given the displayed residual plot only, what is the MOST likely problem with the model?

A. Linear regression is inappropriate. The residuals do not have constant variance.

B. Linear regression is inappropriate. The underlying data has outliers.

C. Linear regression is appropriate. The residuals have a zero mean.

D. Linear regression is appropriate. The residuals have constant variance.

Show Suggested Answer

by takahirokoyama at *Feb. 3, 2021, 11:35 p.m.*

Comments

takahirokoyama Highly Voted 2 years, 9 months ago
Ans. is A.

High-degree polynomial transformation.
upvoted 12 times

TrekkingMachine 2 years, 9 months ago
I think so too.
upvoted 2 times

cnethers Highly Voted 2 years, 8 months ago
Some Good Reading https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html

Ans is A
upvoted 7 times

Sadgamaya 2 years, 3 months ago
Thank you for sharing.
upvoted 1 times

geoan13 Most Recent 8 months ago
Answer A.
One of the key assumptions of linear regression is that the residuals have constant variance at every level of the predictor variable(s). If this assumption is not met, the residuals are said to suffer from heteroscedasticity. When this occurs, the estimates for the model coefficients become unreliable
https://www.statology.org/constant-variance-assumption/
upvoted 1 times

Mickey321 10 months, 2 weeks ago
Selected Answer: A
Agree with A
upvoted 1 times

Nadia0012 1 year, 4 months ago
Selected Answer: A
https://blog.minitab.com/en/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses
upvoted 1 times

Tomatoteacher 1 year, 5 months ago
Selected Answer: A
Kind of like heteroskedasticity, anyways it is A.
upvoted 1 times

jrff 1 year, 8 months ago
Selected Answer: A
Answer is A. It does not has constant variance !
upvoted 3 times

Shailendraa 1 year, 10 months ago
A is correct answer
upvoted 1 times

ovokpus 2 years ago
These images are broken. I cannot review the question properly!
upvoted 2 times

John_Pongthorn 2 years, 4 months ago
Selected Answer: D
D is best answer
all x values are scattering as a whole , no matter what x is
https://www.statisticshowto.com/residual-plot/

if you take all x values to plot histogram , it will be bel-curv.
upvoted 2 times

AddiWei 2 years, 5 months ago
And it does NOT mean linear regression is not appropriate. It means your linear regression model is biased due to several reasons.
upvoted 1 times

eeah 2 years, 3 months ago
yes, it does. One of the main assumptions is homoscedasticity.
upvoted 2 times

AddiWei 2 years, 5 months ago
100% A

upvoted 1 times

u404 2 years, 5 months ago

I will choose A , because the data is heteroscedastic. It violates a key assumption of linear regeression

upvoted 2 times

Huy 2 years, 8 months ago

A. https://www.originlab.com/doc/origin-help/residual-plot-analysis

upvoted 2 times

yummytaco 2 years, 8 months ago

Do not have content variance
https://stats.stackexchange.com/questions/52089/what-does-having-constant-variance-in-a-linear-regression-model-mean

upvoted 2 times

Vita_Rasta84444 2 years, 8 months ago

Answer is A. As x raises, the residuals become higher and higher...

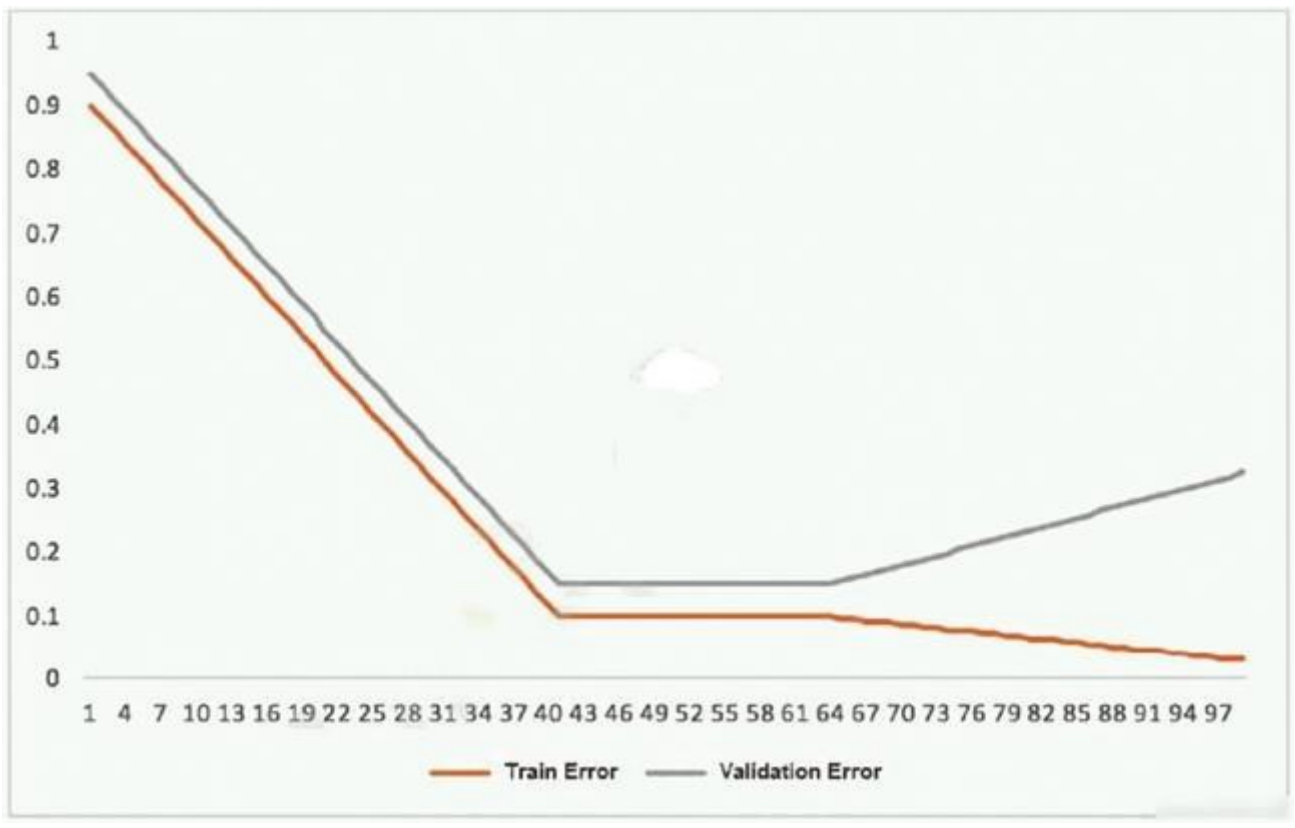
upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 95 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 95
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

This graph shows the training and validation loss against the epochs for a neural network.
The network being trained is as follows:
- Two dense layers, one output neuron
- 100 neurons in each layer
- 100 epochs

Random initialization of weights



Which technique can be used to improve model performance in terms of accuracy in the validation set?

A. Early stopping

B. Random initialization of weights with appropriate seed

C. Increasing the number of epochs

D. Adding another layer with the 100 neurons

Show Suggested Answer

---

by ahquiceno at *Feb. 3, 2021, 4:24 p.m.*

Comments

ahquiceno Highly Voted 3 years, 3 months ago

Answer A.

upvoted 22 times

eganilovic Highly Voted 3 years, 3 months ago

The answer is Early Stopping. Stopp the training before accuracy start do decrease.

upvoted 8 times

StelSen 3 years, 3 months ago

Appreciates your explanation. Cheers

upvoted 2 times

AlWave Most Recent 11 months ago

I will go with A
Early stopping is a powerful technique to prevent overfitting. It involves monitoring the model's performance on a validation dataset during training.
If the validation loss starts increasing or plateaus, early stopping stops further training. This ensures that the model doesn't overfit to the training data.
Based on the graph, if the validation loss begins to stagnate or increase after a certain number of epochs, enabling early stopping could lead to better generalization.

upvoted 1 times

AmeeraM 1 year, 2 months ago

Selected Answer: A

early stopping before error increase

upvoted 1 times

seifskl 1 year, 3 months ago

Selected Answer: A

Early stopping

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

Early stopping

upvoted 1 times

vbal 1 year, 7 months ago

A: stop the training process of a neural network before it reaches the maximum number of epochs or iterations; in this case stop close to 64 Epochs.

upvoted 1 times

Peeking 2 years, 1 month ago

Selected Answer: A

Early stopping and not increasing epochs.

upvoted 1 times

ryuhei 2 years, 3 months ago

Selected Answer: A

Answer is "A"

upvoted 1 times

chrisabc 3 years, 2 months ago

Early Stopping can improve the model?

upvoted 3 times

Vita_Rasta84444 3 years, 3 months ago

A is the answer

upvoted 2 times

astonm13 3 years, 3 months ago

I would go for A

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 94 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 94
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A health care company is planning to use neural networks to classify their X-ray images into normal and abnormal classes. The labeled data is divided into a training set of 1,000 images and a test set of 200 images. The initial training of a neural network model with 50 hidden layers yielded 99% accuracy on the training set, but only 55% accuracy on the test set.
What changes should the Specialist consider to solve this issue? (Choose three.)

A. Choose a higher number of layers

B. Choose a lower number of layers

C. Choose a smaller learning rate

D. Enable dropout

E. Include all the images from the test set in the training set

F. Enable early stopping

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 4:23 p.m.*

Comments

cnethers Highly Voted 3 years, 3 months ago

when looking at an overfitting issue :
https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html
1. Simplifying The Model (reduce number of layers)
2. Early Stopping
3. Use Data Augmentation
4. Use Regularization (L1 + L2)
5. Use Dropouts

So looking at the options:
B, D, F

upvoted 48 times

SophieSu Highly Voted 3 years, 2 months ago

BDF !!!

upvoted 8 times

asthamishra Most Recent 12 months ago

looking at last 100 questions many answers were wrong , thanks to the discussion forum to provide correct answer

upvoted 2 times

AmeeraM 1 year, 2 months ago

Selected Answer: BDF

I would say BCD or BDF

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: BDF

Agree with BDF

upvoted 1 times

JK1977 1 year, 7 months ago

Selected Answer: BDF

Over fitting problem. All the options B, D, F reduce over fitting.

upvoted 1 times

Debayandt91 1 year, 8 months ago

In what world is ACE the answer ?

upvoted 1 times

Dota_addict 1 year, 8 months ago

Selected Answer: BDF

BDF is the answer

upvoted 1 times

Aninina 2 years ago

Selected Answer: BDF

BDF is the correct

upvoted 1 times

Peeking 2 years, 1 month ago

Selected Answer: BDF

ADE is absolutely wrong. 50 layers is already overfitting the model. We cannot increase the number of layers again.

upvoted 1 times

GauravLahotiML 2 years, 2 months ago

Selected Answer: BDF

BDF is the correct answer

upvoted 1 times

exam887 2 years, 7 months ago

Selected Answer: BDF

should be BDF

upvoted 1 times

NILKK 2 years, 8 months ago

One of the correct answer is showing as A. I wanted to understand how A(Choose Higher Number of Layers) is the correct Answer ?

upvoted 1 times

KM226 3 years ago

Selected Answer: BCE

I believe the answer is BCE because the model is overfitting.

upvoted 1 times

windy9 1 year, 3 months ago

C might not be, because the model yielded 99% accuracy on the training set

upvoted 2 times

johnvik 3 years, 2 months ago

choose smaller learning rate c, d, f,

upvoted 1 times

johnvik 3 years, 2 months ago

ignore answer is correct BDF

upvoted 1 times

Vita_Rasta84444 3 years, 2 months ago

BDF!!!

upvoted 3 times

astonm13 3 years, 3 months ago

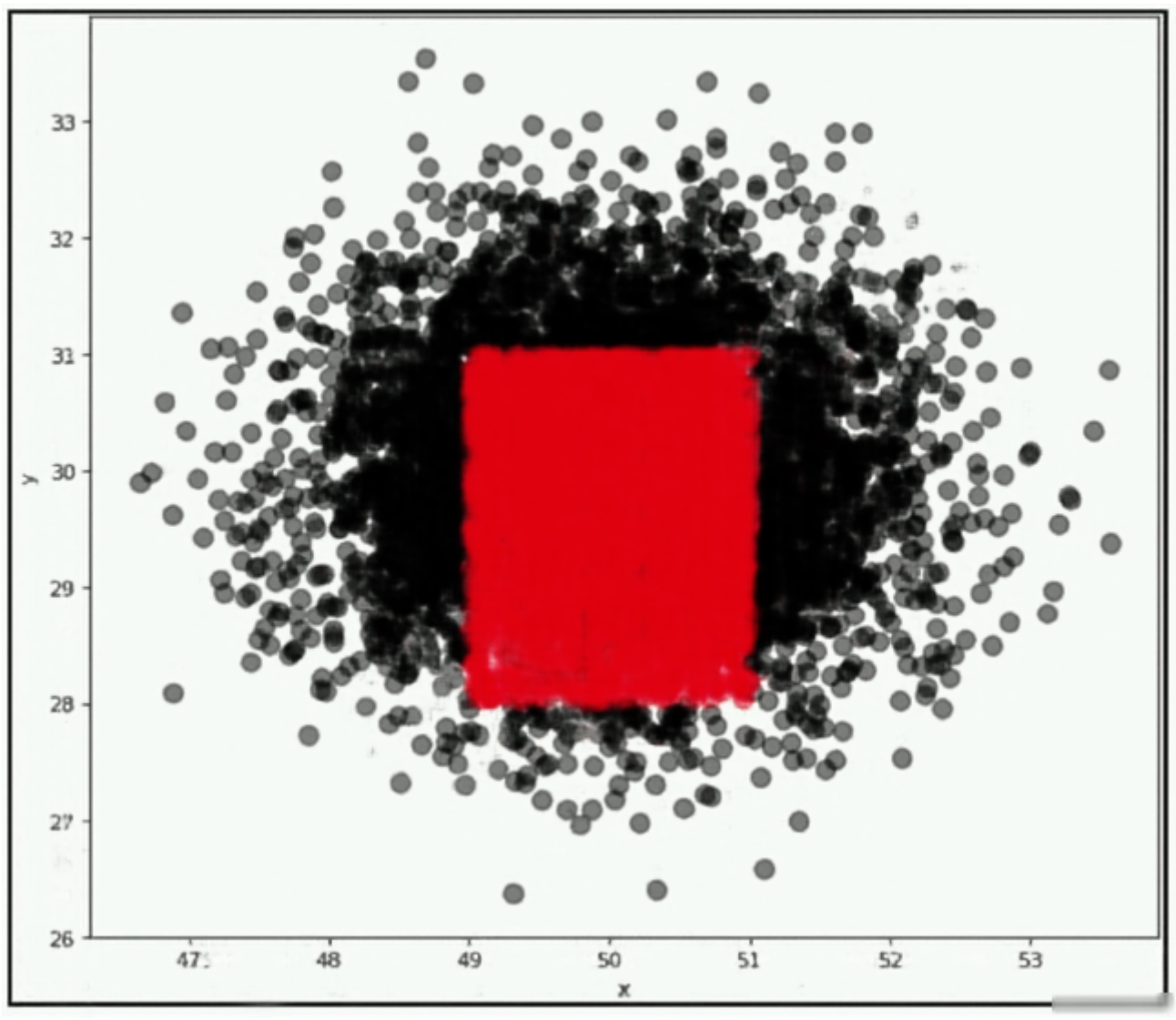
It is supposed to be BDF

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 93 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 93
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company wants to classify user behavior as either fraudulent or normal. Based on internal research, a machine learning specialist will build a binary classifier based on two features: age of account, denoted by x, and transaction month, denoted by y. The class distributions are illustrated in the provided figure. The positive class is portrayed in red, while the negative class is portrayed in black.



Which model would have the HIGHEST accuracy?

A. Linear support vector machine (SVM)

B. Decision tree

C. Support vector machine (SVM) with a radial basis function kernel

D. Single perceptron with a Tanh activation function

Show Suggested Answer

by [deleted] at *Feb. 3, 2021, 4:36 a.m.*

Comments

[Removed] Highly Voted 3 years, 9 months ago

Due to straight angles, I would choose Decision tree. See https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

upvoted 24 times

MrCarter 3 years, 8 months ago

From your link it is obvious that the best answer is still SVM with RBF kernel. In your link the SVM-RBF got 88% accuracy on the 'square-like' dataset whereas the Decision tree achieved only 80%. Answer is SVM with RBF kernel

upvoted 16 times

ttsun 3 years, 7 months ago

note the data from sklearn link is shaped as a ball of mass not a square. the RBF kernel would be better but the question shows a square. Decision tree should be better fit for this problem.

upvoted 7 times

SophieSu Highly Voted 3 years, 9 months ago

B - Decision tree - is not the best answer. If you use decision tree to do clustering, every time you need to partition the space into 2 parts. Hence you will split the space into 3*3. The red points in the center box and the black points will fall into the 8 boxes around it. The black points will be identified as 8 different classes.

C is the correct answer. SVM with non-linear kernel is appropriate for non-linear clustering. Even if the shape is close to rectangular. SVM with non-linear kernel will be ale to approximate the rectangular boundary shape.

upvoted 18 times

robotgeek 1 year, 10 months ago

Your statement "The black points will be identified as 8 different classes" does not make a lot of sense because the leaf node in a tree with be 1 of 2 classes, not 8 different classes just because they are visually in one place or the other

upvoted 1 times

Madwyn 3 years, 8 months ago

The tree works like this with this branch with 4 nodes:
Age > 49? Y
Age > 51? N
Transaction > 28? Y
Transaction > 31? N
Positive

Correct answer is B.

upvoted 10 times

nick3332 Most Recent 3 months, 1 week ago

Selected Answer: B

Tip: When details are missing, assume ideal conditions, so assume no overfitting issues. Therefore is B is better than C. If there are overfitting issues or a posibility of overfitting then C is the right answer.

upvoted 1 times

2eb8df0 4 months, 2 weeks ago

Selected Answer: B

Decision tree makes more sense, this decision boundary isn't complex at all and there is no risk of overfitting, all the points are inside the square

upvoted 2 times

MultiCloudIronMan 8 months, 3 weeks ago

Selected Answer: C

This is because the RBF kernel can handle non-linear relationships between features, which is often necessary for complex classification tasks.

upvoted 2 times

MJSY 9 months, 2 weeks ago

Selected Answer: C

Decision Tree can treat the training data well but will have a risk of overfitting. the SVM with RBF kernel will be more robust.

upvoted 2 times

rookiee1111 1 year, 2 months ago

Selected Answer: B

As the positive cases can be interpreted and separated from non positive ones by decision tree easily. SVM would have made sense if the two classes were inseparable or had complex relationship in data.

upvoted 1 times

vkbajoria 1 year, 3 months ago

Selected Answer: C

It is C SVM with RBF Kernel can classify this image. For decision tree, it will be more difficult

upvoted 2 times

kyuhuck 1 year, 5 months ago

Selected Answer: C

From the visual information provided, an SVM with an RBF kernel (Option C) would likely be the best choice because it can handle the circular class distribution. The RBF kernel is especially good at dealing with such scenarios where the boundary between classes is not linear.

upvoted 2 times

Alice1234 1 year, 5 months ago

Answer C

B. Decision Tree: Decision trees can capture non-linear patterns and are capable of splitting the feature space in complex ways. They can be very effective if the decision boundary is not linear, but they might also overfit if the decision boundary is too complex.

C. SVM with RBF Kernel: An SVM with a radial basis function (RBF) kernel is designed to handle non-linear boundaries by mapping input features into higher-dimensional spaces where the classes are more likely to be separated by a hyperplane. Given the clustered nature of the classes in the image, an SVM with an RBF kernel would likely be able to separate the classes with a higher degree of accuracy.

upvoted 2 times

praveenaws 1 year, 6 months ago

Selected Answer: C

SVM-RBF is the correct solution

upvoted 1 times

Neet1983 1 year, 6 months ago

Support vector machine (SVM) with a radial basis function kernel would likely have the highest accuracy for this task because it can handle the non-linear separation required by the data.

upvoted 1 times

endeesa 1 year, 7 months ago

Selected Answer: C

I will lean with C

upvoted 1 times

akgarg00 1 year, 7 months ago

Answer is B as Decision tree can attain 100% accuracy in this case.

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: C

SVM with RBF and proper C and Gamma value can accomodate this square shape (https://vitalflux.com/svm-rbf-kernel-parameters-code-sample/)

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: C

confusing between SVN or devision tree. learning towards C

upvoted 1 times

rags1482 2 years, 1 month ago

Answer C
In general, SVMs are a good choice for tasks where accuracy is critical, such as fraud detection and medical diagnosis. Decision trees are a good choice for tasks where interpretability is important, such as customer segmentation and product recommendation.

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 92 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 92

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A company wants to predict the sale prices of houses based on available historical sales data. The target variable in the company's dataset is the sale price. The features include parameters such as the lot size, living area measurements, non-living area measurements, number of bedrooms, number of bathrooms, year built, and postal code. The company wants to use multi-variable linear regression to predict house sale prices.
Which step should a machine learning specialist take to remove features that are irrelevant for the analysis and reduce the model's complexity?

A. Plot a histogram of the features and compute their standard deviation. Remove features with high variance.

B. Plot a histogram of the features and compute their standard deviation. Remove features with low variance.

C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores.

D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 4:08 p.m.*

Comments

puffpuff Highly Voted 2 years, 8 months ago

D should be the more comprehensive answer. If it's not correlated, you can't make use of it in a linear regression
A lot of others say B, but low variance can also be due to the nature/typical magnitudes of the variable itself

upvoted 29 times

hamimelon 1 year, 6 months ago

I think the problem with B is that what is considered "low variance"? The features are on different scales.

upvoted 1 times

V_B_ 1 year, 11 months ago

Correlation indicates only linear relation, but, there might be non linear as well. To exploit it in the Linear Regression, you can take the variables to some power or run some non linear preprocessing on it, and you don't have to change the algorithm for it.
So, answer B seem much more solid for me.

upvoted 2 times

ahquiceno Highly Voted 2 years, 9 months ago

Answer B. Is not the best solucion prior can use other analysis. https://community.dataquest.io/t/feature-selection-features-with-low-variance/2418
If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model. In that case, it

should be removed. Or if only a handful of observations differ from a constant value, the variance will also be very low.

upvoted 16 times

fshkkento 2 years, 7 months ago

Low variance does not mean the feature is not important, right?
If variance of target true value is also small and the correlation between above feature and target, the feature can be important feature.

upvoted 7 times

rb39 1 year, 10 months ago

it does. If feature and target are correlated and you expect the target to change, the feature must have some sort of variance. Otherwise it means feature is almost constant so does target.

upvoted 1 times

akgarg00 Most Recent 7 months, 3 weeks ago

Selected Answer: D

D is the best answer as it is mentioned multivariable linear regression applied where correlation is strong between dependent and independent variables.

upvoted 1 times

mirik 1 year ago

Selected Answer: D

D: We should remove features that are strongly correlated with each other and weakly correlated with the target:

https://androidkt.com/find-correlation-between-features-and-target-using-the-correlation-matrix/
You can evaluate the relationship between each feature and target using a correlation and selecting those features that have the strongest relationship with the target variable.

upvoted 2 times

HunterZ9527 1 year, 2 months ago

Selected Answer: D

I think D is the correct answer. If I remember correctly, Benjamini-Hochberg Method is essentially answer D if you consider the Hypothesis to be: the feature is powerfully influential to the target.
My problem with B is that the variance can be easily affected by the scale. In the question, the number of bedroom's variance is very low, while the sqrt of the house has a high variance, both of these could be very useful. Furthermore, zip codes are included, and it is safe to assume the variance of zip codes can be high, but the information is very limited, especially if you use them as numerical instead of categorical features.

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: D

B is correct but the answer in D is better.

upvoted 2 times

AjoseO 1 year, 5 months ago

Selected Answer: D

D is preferred over C because the goal is to predict the sale price of houses, which is the target variable. By checking the correlation of each feature against the target variable, the machine learning specialist can identify which features are most relevant to the prediction of the sale price and which are less relevant. Removing features with low correlation to the target variable helps reduce the complexity of the model and potentially improve its accuracy.

On the other hand, a heatmap showing the correlation of the dataset against itself (C) doesn't directly address the relevance of the features to the target variable, and so it's not as effective in reducing the complexity of the model.

upvoted 3 times

expertguru 1 year, 6 months ago

Answer should be D, THIS is feature elimination /selection during feature Enginering. Choice c is so close just to confuse test takers to pick the wrong choice! See below C and D answers -- C should have been correct if the question asked about how to visualize correlation among independent variables! PROVIDED second sentence in C needs to be removed or to say which feature you will eliminate in such case then the one with low correlation against target out of those two.

C. Build a heatmap showing the correlation of the dataset against itself. Remove features with low mutual correlation scores.
D. Run a correlation check of all features against the target variable. Remove features with low target variable correlation scores.

upvoted 1 times

Ob1KN0B 1 year, 11 months ago

Selected Answer: D

The multiple regression model is based on the following assumptions:

There is a linear relationship between the dependent variables and the independent variables
The independent variables are not too highly correlated with each other
yi observations are selected independently and randomly from the population
Residuals should be normally distributed with a mean of 0 and variance σ

upvoted 5 times

wakuwaku 2 years, 4 months ago

I think the answer is D.
If the model is a decision tree or something like that, I don't think it is possible to make a decision based only on the direct correlation with the target variable.
But in multiple linear regression, the only thing that matters is the relationship between the target variable and the feature variable.
B, if the standard deviation is small but not zero, then we have information.

upvoted 3 times

apprehensive_scar 2 years, 5 months ago

Selected Answer: B

B is correct.

upvoted 2 times

Peasfull 2 years, 6 months ago

To eliminate extraneous information. So, the answer is D.

upvoted 2 times

Asrivastava3 2 years, 7 months ago

Correct answer is D. The reason B is wrong because it is difficult to reason out why would you plot a histogram? Absolutely unnecessary step and distraction choice.

upvoted 4 times

MikkyO 2 years, 8 months ago

Answer is D.
https://deep-r.medium.com/difference-between-variance-co-variance-and-correlation-ea0b7ddbaa1

upvoted 5 times

Huy 2 years, 8 months ago

Answer C. Heatmaps is used to visualize for correlation matrix https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

upvoted 1 times

mahmoudai 2 years, 8 months ago

but is mentioned, "Remove features with low mutual correlation scores." which is wrong you should drop features with high correlation scores. so Answer is D

upvoted 5 times

hero67 2 years, 8 months ago

The problem with correlation tasks is it capture linear relations only. So, I would go with B

upvoted 1 times

YJ4219 2 years, 9 months ago

I think the answer is D, because the correlation between each feature and the target, if this feature has low variance (as B suggests) this correlation calculation will be low and thus will be removed.
In short i think D is a more general and more accurate answer.

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 91 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 91
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A manufacturer of car engines collects data from cars as they are being driven. The data collected includes timestamp, engine temperature, rotations per minute
(RPM), and other sensor readings. The company wants to predict when an engine is going to have a problem, so it can notify drivers in advance to get engine maintenance. The engine data is loaded into a data lake for training.
Which is the MOST suitable predictive model that can be deployed into production?

A. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a recurrent neural network (RNN) to train the model to recognize when an engine might need maintenance for a certain fault.

B. This data requires an unsupervised learning algorithm. Use Amazon SageMaker k-means to cluster the data.

C. Add labels over time to indicate which engine faults occur at what time in the future to turn this into a supervised learning problem. Use a convolutional neural network (CNN) to train the model to recognize when an engine might need maintenance for a certain fault.

D. This data is already formulated as a time series. Use Amazon SageMaker seq2seq to model the time series.

Show Suggested Answer

by ac71 at *Feb. 3, 2021, 4:04 a.m.*

Comments

ac71 Highly Voted 2 years, 9 months ago
This is a supervised problem and needs labels. Can't use clustering to find when faults can happen. CNN is for images not for timeseries data here. Hence, A seems appropriate.
upvoted 53 times

youjun 2 years, 6 months ago
AGREE WITH YOU
upvoted 2 times

astonm13 2 years, 9 months ago
Agree, the answer is A
upvoted 7 times

loict Most Recent 10 months ago
Selected Answer: A

A. YES - RNN good for time series as we want to use previous input
B. NO - we know the class (fault) ahead of time, it is supervised
C. NO - CNN is for images
D. NO - seq2seq is for word generation

upvoted 1 times

Mickey321 10 months, 2 weeks ago

Selected Answer: A

Answer is A

upvoted 1 times

AjoseO 1 year, 5 months ago

Selected Answer: A

A recurrent neural network (RNN) is a more suitable choice than a convolutional neural network (CNN) because the data collected from the engines is a sequence of values over time, and the goal is to predict a future event (an engine fault). RNNs are designed to handle sequential data and can learn patterns and dependencies over time, making them well-suited for time-series data like this.

On the other hand, CNNs are designed for image processing and are not ideal for sequential data.

upvoted 3 times

spidy20 1 year, 10 months ago

Selected Answer: A

Answer should be A

upvoted 1 times

Morsa 1 year, 12 months ago

Selected Answer: A

It can only be A. Agree with the comments before

upvoted 1 times

irimala 2 years, 1 month ago

Selected Answer: A

Obviously A

upvoted 2 times

apprehensive_scar 2 years, 5 months ago

Selected Answer: A

A - obviously.

upvoted 1 times

bitsplease 2 years, 5 months ago

Seq2Seq also uses RNN under the hood, BUT option D. did not mention anything about "adding labels"--which is required here--hence --> A

upvoted 3 times

geekgirl007 2 years, 6 months ago

Selected Answer: A

A is correct. CNN is for images and RNN is for timeseries.

upvoted 1 times

loyor94478 2 years, 8 months ago

AAAAAAAAAAA
https://towardsdatascience.com/how-to-implement-machine-learning-for-predictive-maintenance-4633cdbe4860

upvoted 2 times

omar8024 2 years, 8 months ago

I think A is correct

upvoted 1 times

Vita_Rasta84444 2 years, 8 months ago

It is A

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 90 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 90

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is planning to create a long-running Amazon EMR cluster. The EMR cluster will have 1 master node, 10 core nodes, and 20 task nodes. To save on costs, the Specialist will use Spot Instances in the EMR cluster.
Which nodes should the Specialist launch on Spot Instances?

A. Master node

B. Any of the core nodes

C. Any of the task nodes

D. Both core and task nodes

Show Suggested Answer

by ac71 at *Feb. 3, 2021, 4:01 a.m.*

Comments

[Removed] Highly Voted 2 years, 9 months ago

Answer is C. https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

upvoted 23 times

Sneep 1 year, 6 months ago

It's definitely C. The fact that this site indicates A is a clear sign that answers are just randomly selected, it would make zero sense to spot-instance the master node for an EMR cluster. Make sure you look at discussions for all of these questions.

upvoted 4 times

SophieSu Highly Voted 2 years, 8 months ago

C is the correct answer.
"Long-Running Clusters and Data Warehouses
If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances. You can launch your master and core instance groups as On-Demand Instances to handle the normal capacity and launch task instance groups as Spot Instances to handle your peak load requirements."

upvoted 10 times

teka112233 Most Recent 10 months, 3 weeks ago

Selected Answer: C

According to :https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

The task nodes process data but do not hold persistent data in HDFS. If they terminate because the Spot price has risen above your maximum Spot price, no data is lost and the effect on your cluster is minimal.

When you launch one or more task instance groups as Spot Instances, Amazon EMR provisions as many task nodes as it can, using your maximum Spot price. This means that if you request a task instance group with six nodes, and only five Spot Instances are available at or below your maximum Spot price, Amazon EMR launches the instance group with five nodes, adding the sixth later if possible.

upvoted 1 times

Khalil11 1 year, 3 months ago

Selected Answer: C

The correct answer is C

upvoted 1 times

Sylzys 1 year, 4 months ago

Selected Answer: C

I don't get why the wrong answer are still not updated after more than 1 year of everyone showing docs proving answer C..

upvoted 3 times

gusta_dantas 11 months, 3 weeks ago

1 and a half year and still wrong.. Incredible!

upvoted 2 times

AjoseO 1 year, 5 months ago

Selected Answer: C

Long-running clusters and data warehouses

If you are running a persistent Amazon EMR cluster that has a predictable variation in computational capacity, such as a data warehouse, you can handle peak demand at lower cost with Spot Instances.

You can launch your primary and core instance groups as On-Demand Instances to handle the normal capacity and launch the task instance group as Spot Instances to handle your peak load requirements.

upvoted 1 times

SK27 1 year, 7 months ago

Selected Answer: C

Only task nodes can be deleted without losing data.

upvoted 1 times

Twist3d 1 year, 7 months ago

C, If you want to cut cost on an EMR cluster in the most efficient way, use spot instances on the task nodes because it, task nodes do not store data so no risk of data loss

upvoted 1 times

ovokpus 2 years ago

Selected Answer: C

For Long running jobs, you do not want to compromise the Master node(sudden termination) or the core nodes (HDFS data loss).

Spot Instances on 20 task nodes are enough cost savings without compromising the job.

Hence, C

upvoted 3 times

Jump09 2 years ago

If your primary concern is the cost, then you can run the master node on spot instances.

upvoted 1 times

Jump09 2 years ago

Adding the related reference from the AWS documentation:
Master node on a Spot Instance
The master node controls and directs the cluster. When it terminates, the cluster ends, so you should only launch the master node as a Spot Instance if you are running a cluster where sudden termination is acceptable. This might be the case if you are testing a new application, have a cluster that periodically persists data to an external store such as Amazon S3, or are running a cluster where cost is more important than ensuring the cluster's completion.

upvoted 1 times

Jump09 2 years ago

In the question , there are no specific conditions mentioned except the concern with the COST, thus I think the answer should be A.

upvoted 1 times

benson2021 2 years, 8 months ago

Answer: C. https://aws.amazon.com/getting-started/hands-on/optimize-amazon-emr-clusters-with-ec2-spot/
Amazon recommends using On-Demand instances for Master and Core nodes unless you are launching highly ephemeral workloads.

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 89 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 89
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is given a structured dataset on the shopping habits of a company's customer base. The dataset contains thousands of columns of data and hundreds of numerical columns for each customer. The Specialist wants to identify whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible.
What approach should the Specialist take to accomplish these tasks?

A. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a scatter plot.

B. Run k-means using the Euclidean distance measure for different values of k and create an elbow plot.

C. Embed the numerical features using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and create a line graph.

D. Run k-means using the Euclidean distance measure for different values of k and create box plots for each numerical column within each cluster.

Show Suggested Answer

by ac71 at *Feb. 3, 2021, 3:55 a.m.*

Comments

ac71 Highly Voted 2 years, 9 months ago

A is correct. tSNE can do segmentation or grouping as well. Refer: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

upvoted 21 times

SophieSu Highly Voted 2 years, 8 months ago

A is definitely the correct answer.
Pay attention to what the question is asking:
"whether there are natural groupings for these columns across all customers and visualize the results as quickly as possible"

The key point is to visualize the "groupings"(exactly what t-SNE scatter plot does, it visualize high-dimensional data points on 2D space).
The question does not ask to visualize how many groups you would classify (K-Means Elbow Plot does not visualize the groupings, it is used to determine the optimal # of groups=K).

upvoted 18 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: A

option A

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

B doesn't even answer the question: how are you going to see your customer groups in an elbow plot

upvoted 1 times

windy9 9 months, 2 weeks ago

Elbow plot helps you identify the correct number of clusters during K-Means clustering. The clustering happens basis of all the features and thus group employees. This is to help your understanding. And the correct answer however is still tSNE becuase the question focuses on identifying relationships/similarities between the features / columns in the dataset. The correct answer is A

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: A

Euclidean Distance suffers for high dimensional data. tSNE can suffers as well, but from my perspective is the correct one.

upvoted 1 times

Sylzys 1 year, 4 months ago

Selected Answer: A

Elbow plot will not help visualize groups, only try to predict an optimal number of clusters.
I think A is a better choice here

upvoted 2 times

AjoseO 1 year, 5 months ago

Selected Answer: A

A.

The t-SNE algorithm is a popular tool for visualizing high-dimensional datasets, as it can transform high-dimensional data into a 2D scatter plot, which makes it easier to visualize and understand the relationships between data points.

The scatter plot produced by t-SNE can be interpreted as a map that reveals the structure of the data, showing whether there are natural groupings or clusters within the data.

Option A is the quickest and simplest way to visualize the data in a meaningful way, allowing the Specialist to gain insights into the data more efficiently.

upvoted 3 times

minkhant19 1 year, 7 months ago

A is correct

upvoted 1 times

Shailendraa 1 year, 10 months ago

12-sep exam

upvoted 3 times

Morsa 1 year, 12 months ago

Selected Answer: A

A as k-means elbow is erroneous. It does not helping here. Scatter plot and t-sne is the right answer

upvoted 2 times

ovokpus 2 years ago

Selected Answer: A

An elbow plot (B) will not give you what the question is asking for. A scatter plot will, and t-SNE is first for visualizing before dimensionality reduction.

upvoted 2 times

Sadgamaya 2 years, 3 months ago

A is correct as k means suffer from curse of dimensionality and t-she will be a better option.

upvoted 1 times

Mircuz 2 years, 4 months ago

Selected Answer: A

The B,C,D plots are meaningless wrt the problem —> A

upvoted 2 times

Mircuz 2 years, 4 months ago

Selected Answer: B

t-SNE suffers curse of dimensionality and is indicated for small datasets

upvoted 1 times

AddiWei 2 years, 5 months ago

Additionally the numeric features don't require "embedding". I think they meant to write "standardize"

upvoted 1 times

apprehensive_scar 2 years, 5 months ago

Rooting for A

upvoted 1 times

bitsplease 2 years, 5 months ago

B & D are wrong--because data contains "thousands of columns" and using k-means with euclidean suffers from "curse of dimensionality"

Thus leaving A & C, you CANNOT viz clusters/groups/segments in a line graph so correct answer is A

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 88 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 88

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A web-based company wants to improve its conversion rate on its landing page. Using a large historical dataset of customer visits, the company has repeatedly trained a multi-class deep learning network algorithm on Amazon SageMaker. However, there is an overfitting problem: training data shows 90% accuracy in predictions, while test data shows 70% accuracy only.
The company needs to boost the generalization of its model before deploying it into production to maximize conversions of visits to purchases.
Which action is recommended to provide the HIGHEST accuracy model for the company's test and validation data?

A. Increase the randomization of training data in the mini-batches used in training

B. Allocate a higher proportion of the overall data to the training dataset

C. Apply L1 or L2 regularization and dropouts to the training

D. Reduce the number of layers and units (or neurons) from the deep learning network

Show Suggested Answer

by knightknt at *April 20, 2022, 12:42 p.m.*

Comments

knightknt Highly Voted 2 years, 2 months ago

I think C will be answer, because we even don't know how many layers now, so apply L1,L2 and dropouts layer will be first resort to solve overfitting. If it still does not work, then to reduce layers

upvoted 11 times

mamun4105 Most Recent 10 months, 1 week ago

D: D is the correct answer. C could be the answer only if it is a regression problem. You cannot apply L1 (Lasso regression) and L2 (Ridge regression) to classification problems. However, you can use dropout here.

upvoted 1 times

DimLam 8 months, 2 weeks ago

Why do you think it works only for regression problems? L1/L2 regularizations are just adding penalties to loss functions. I don't see any problems with applying it to DL model

upvoted 1 times

Mickey321 10 months, 2 weeks ago

Selected Answer: C

C Regulization

upvoted 2 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: C

if you see overfit think regularization.

upvoted 1 times

Khalil11 1 year, 3 months ago

Selected Answer: C

C is the correct answer: The overfitting problem can be addressed by applying regularization techniques such as L1 or L2 regularization and dropouts. Regularization techniques add a penalty term to the cost function of the model, which helps to reduce the complexity of the model and prevent it from overfitting to the training data. Dropouts randomly turn off some of the neurons during training, which also helps to prevent overfitting.

upvoted 2 times

Valcilio 1 year, 4 months ago

Selected Answer: C

D can work, but C is a better answer!

upvoted 2 times

drcok87 1 year, 4 months ago

C and D both seems to be correct but, seems like removing layer is first step in to optimization
https://www.kaggle.com/general/175912
d

upvoted 2 times

AjoseO 1 year, 5 months ago

Selected Answer: C

C. Apply L1 or L2 regularization and dropouts to the training" because regularization can help reduce overfitting by adding a penalty to the loss function for large weights, preventing the model from memorizing the training data.

Dropout is a regularization technique that randomly drops out neurons during the training process, further reducing the risk of overfitting.

upvoted 1 times

albu44 1 year, 6 months ago

Selected Answer: D

"The first step when dealing with overfitting is to decrease the complexity of the model. To decrease the complexity, we can simply remove layers or reduce the number of neurons to make the network smaller."
https://www.kdnuggets.com/2019/12/5-techniques-prevent-overfitting-neural-networks.html

upvoted 1 times

Peeking 1 year, 7 months ago

Selected Answer: D

Deep learning tuning order:
1. Number of layers
2. Number of neurons (indirectly implements dropout)
3. L1/L2 regularization
4. Dropout

upvoted 4 times

kaike_reis 11 months, 2 weeks ago

the problem is overfitting, not HP Tuning.

upvoted 1 times

Shakespeare 7 months ago

Can be used for overfitting as well, but the problem does not say it is a deep learning algorithm being used so C would be more appropriate.

upvoted 1 times

Parth12 1 year, 11 months ago

Selected Answer: C

Here we are looking to reduce the Overfitting to improve the generalization. In order to do so, L1(or Lasso) regression has always been a good aide.

upvoted 3 times

mamun4105 10 months, 1 week ago

This is not a regression problem at all.

upvoted 1 times

mtp1993 2 years ago

Selected Answer: C

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 87 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 87
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist needs to analyze employment data. The dataset contains approximately 10 million observations on people across 10 different features. During the preliminary analysis, the Data Scientist notices that income and age distributions are not normal. While income levels shows a right skew as expected, with fewer individuals having a higher income, the age distribution also shows a right skew, with fewer older individuals participating in the workforce.
Which feature transformations can the Data Scientist apply to fix the incorrectly skewed data? (Choose two.)

A. Cross-validation

B. Numerical value binning

C. High-degree polynomial transformation

D. Logarithmic transformation

E. One hot encoding

Show Suggested Answer

by Joe_Zhang at *Feb. 2, 2021, 12:46 a.m.*

Comments

seanLu Highly Voted 3 years, 2 months ago

I would go with B,D. Refer to quantile binning and log transform below.
https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b

upvoted 31 times

OmarSaadEldien 3 years, 2 months ago

Agree with B &D
B binning for age
D for make income in normal dist

upvoted 7 times

omar_bahrain 3 years, 2 months ago

agree B&D. both are strategies to eliminate the effect of skewing

upvoted 6 times

Joe_Zhang Highly Voted 3 years, 3 months ago

SHOULD BE C,D

upvoted 10 times

Togy Most Recent 3 months ago

Selected Answer: B

Binning involves grouping numerical values into discrete intervals or bins. While it can simplify the representation of a feature and potentially make the distribution appear less skewed in a histogram, it doesn't fundamentally change the underlying skewness of the continuous data. It discretizes the data rather than transforming its distribution.

upvoted 1 times

VR10 10 months, 3 weeks ago

Selected Answer: BD

D. Logarithmic Transformation: Addresses the right-skewed income and age distributions. The log function compresses large values, reducing the impact of outliers and making the distributions closer to normal.

B. Numerical Value Binning: Useful for the age distribution. By grouping ages into bins (e.g., 20-29, 30-39, etc.), you reduce the impact of the right skew caused by fewer older individuals. While it doesn't achieve a perfectly normal distribution, it often makes the feature more interpretable and manageable for modeling.

upvoted 1 times

AmeeraM 1 year, 3 months ago

Selected Answer: BD

B and D

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: BD

Agree with B &D B binning for age D for make income in normal dist

upvoted 1 times

Shailendraa 2 years, 4 months ago

BD is correct

upvoted 2 times

Sivadharan 2 years, 7 months ago

Selected Answer: BD

B & D. Reasonable explanation in below discussion.

upvoted 3 times

angnam 2 years, 11 months ago

BD
With age, always do quantile binning
With skewed data, always use log.

upvoted 1 times

Juka3lj 3 years, 2 months ago

B because we have skewed data with few exeptions
D log transform can change distribution of data
not C - because there is no indicaiton in the text, that data is following any of the HIGH DEGREE polynomial distribution like x^ 10

upvoted 5 times

Vita_Rasta84444 3 years, 2 months ago

should be c and d

upvoted 4 times

achiko 3 years, 2 months ago

polynomial transformations can also be used for skewed data. https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/

upvoted 3 times

jiadong 3 years, 3 months ago

It seems the ans are C,D
https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53

upvoted 5 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 86 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 86

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist wants to bring a custom algorithm to Amazon SageMaker. The Specialist implements the algorithm in a Docker container supported by Amazon SageMaker.

How should the Specialist package the Docker container so that Amazon SageMaker can launch the training correctly?

A. Modify the bash_profile file in the container and add a bash command to start the training program

B. Use CMD config in the Dockerfile to add the training program as a CMD of the image

C. Configure the training program as an ENTRYPOINT named train

D. Copy the training program to directory /opt/ml/train

Show Suggested Answer

by Paul_NoName at *Feb. 3, 2021, 7:59 p.m.*

Comments

Paul_NoName Highly Voted 3 years, 3 months ago

C seems correct as per documentations.

upvoted 17 times

awsemort Most Recent 11 months, 1 week ago

Selected Answer: C

I thought it was D, but it is C. It's not D because we copy the TRAINING code into /opt/ml/code/train.py

upvoted 2 times

MaximusDecimus 11 months, 1 week ago

In Docker, the ENTRYPOINT instruction is used to specify the executable that should be run when the container starts. However, Amazon SageMaker expects the training script to be launched by specific commands provided by SageMaker itself, rather than relying solely on the Docker container's ENTRYPOINT. The convention for using Docker containers with Amazon SageMaker is to copy the training script and associated resources to specific directories within the container, such as /opt/ml/code, and let SageMaker manage the execution of the training process.

I would go with D

upvoted 1 times

cyberfriends 1 year, 2 months ago

Selected Answer: C

C is correct

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: C

C is correct

upvoted 1 times

JK1977 1 year, 7 months ago

Selected Answer: C

Amazon SageMaker requires that a custom algorithm container has an executable named train that runs your training program. This executable can be configured as an ENTRYPOINT in the Dockerfile, which specifies the default command to run when the container is launched.

upvoted 3 times

JK1977 1 year, 7 months ago

Selected Answer: B

Amazon SageMaker requires that a custom algorithm container has an executable named train that runs your training program. This executable can be configured as an ENTRYPOINT in the Dockerfile, which specifies the default command to run when the container is launched.

upvoted 1 times

Dota_addict 1 year, 8 months ago

Selected Answer: D

you are all wrong, it is D based on,https://docs.aws.amazon.com/sagemaker/latest/dg/adapt-training-container.html

upvoted 1 times

oso0348 1 year, 9 months ago

Selected Answer: C

To package a Docker container for use with Amazon SageMaker, the training program should be configured as an ENTRYPOINT named train in the Dockerfile. This means that the training program will be automatically executed when the container is launched by Amazon SageMaker, and it can be passed command-line arguments to specify hyperparameters or other training settings.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: C

The recommended option to package the Docker container for Amazon SageMaker is to configure the training program as an ENTRYPOINT named train.

This is because ENTRYPOINT allows you to specify a command that will always be executed when the Docker container is run, ensuring that the training program will always run when the container is launched by Amazon SageMaker.

Additionally, naming the ENTRYPOINT "train" is a convention used by Amazon SageMaker to identify the main training script.

upvoted 1 times

tsangckl 2 years, 1 month ago

Selected Answer: C

It's C

upvoted 1 times

gnolam 2 years, 4 months ago

C for sure
as per AWS docs:
> In your Dockerfile, use the exec form of the ENTRYPOINT instruction:
> ENTRYPOINT ["python", "k-means-algorithm.py"]

upvoted 1 times

Juka3lj 3 years, 3 months ago

C is correct

upvoted 1 times

Aashi22 3 years, 3 months ago

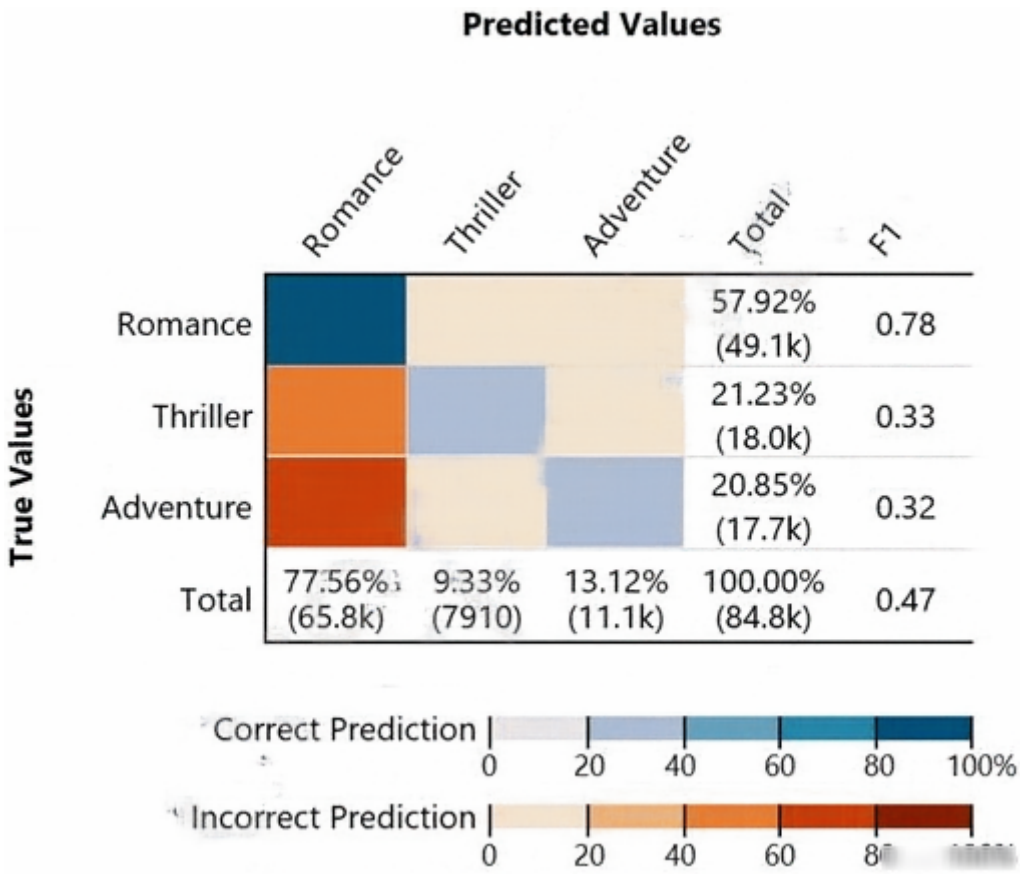
option C https://github.com/awsdocs/amazon-sagemaker-developer-guide/blob/master/doc_source/your-algorithms-training-algo-dockerfile.md

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 85 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 85
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

Given the following confusion matrix for a movie classification model, what is the true class frequency for Romance and the predicted class frequency for
Adventure?



A. The true class frequency for Romance is 77.56% and the predicted class frequency for Adventure is 20.85%

B. The true class frequency for Romance is 57.92% and the predicted class frequency for Adventure is 13.12%

C. The true class frequency for Romance is 0.78 and the predicted class frequency for Adventure is (0.47-0.32)

D. The true class frequency for Romance is 77.56% — 0.78 and the predicted class frequency for Adventure is 20.85% — 0.32

Show Suggested Answer

---

by cnethers at *Feb. 5, 2021, 2:34 p.m.*

Comments

SophieSu Highly Voted 2 years, 9 months ago

B is the correct answer. Straightforward!

upvoted 17 times

cnethers Highly Voted 2 years, 9 months ago

https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

upvoted 10 times

teka112233 Most Recent 10 months, 3 weeks ago

Selected Answer: B

to be able to understand this Multiclass Model Insights and to be able to answer this question :
True class-frequencies in the evaluation data: The second to last column shows that in the evaluation dataset, 57.92% of the observations in the evaluation data is Romance, 21.23% is Thriller, and 20.85% is Adventure.

Predicted class-frequencies for the evaluation data: The last row shows the frequency of each class in the predictions. 77.56% of the observations is predicted as Romance, 9.33% is predicted as Thriller, and 13.12% is predicted as Adventure.
REF: https://docs.aws.amazon.com/machine-learning/latest/dg/multiclass-model-insights.html

upvoted 1 times

Shailendraa 1 year, 10 months ago

12-sep exam

upvoted 2 times

milan_ml 1 year, 11 months ago

Selected Answer: B

The image can be found here:
https://vceguide.com/what-is-the-true-class-frequency-for-romance-and-the-predicted-class-frequency-for-adventure/

upvoted 1 times

obadazx 2 years ago

No image is there!

upvoted 3 times

SDikeman62 2 years, 2 months ago

WHy there is no image? Admin. Please fix it.

upvoted 6 times

Juka3lj 2 years, 8 months ago

B is correct

upvoted 1 times

NotAnMLProfessional 2 years, 8 months ago

A seems to be correct

upvoted 3 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 84 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 84

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is applying a linear least squares regression model to a dataset with 1,000 records and 50 features. Prior to training, the ML
Specialist notices that two features are perfectly linearly dependent.
Why could this be an issue for the linear least squares regression model?

A. It could cause the backpropagation algorithm to fail during training

B. It could create a singular matrix during optimization, which fails to define a unique solution

C. It could modify the loss function during optimization, causing it to fail during training

D. It could introduce non-linear dependencies within the data, which could invalidate the linear assumptions of the model

Show Suggested Answer

by Paul_NoName at *Feb. 3, 2021, 7:56 p.m.*

Comments

Paul_NoName Highly Voted 2 years, 9 months ago

B is correct answer .

upvoted 22 times

hamimelon 1 year, 6 months ago

Agree, B.

upvoted 2 times

pravv 2 years, 8 months ago

why B is the correct answer and not C?

upvoted 1 times

hamimelon 1 year, 6 months ago

For example. If you have two variables, X and Y, and you have two data points. You want to solve the problem: aX1+bY1 = Z1, aX2 + bY2 = Z2. However, if Y=2X -> Y1 = 2X1, Y2 = 2X2, then problem becomes: aX1+bY1 = Z1, a*2X1 + b*2Y1 = Z2 = 2*Z1. So you end up with only one function: aX1+bY1=Z1, meaning there will be more than one answer for (a, b).

If you are familiar with linear algebra, it's easier to express the concept.

upvoted 9 times

SophieSu 2 years, 8 months ago

A square matrix is singular, that is, its determinant is zero, if it contains rows or columns which are proportionally interrelated; in other words, one or more of its rows (columns) is exactly expressible as a linear combination of all or some other its rows (columns), the combination being without a constant term.

upvoted 7 times

Sneep Highly Voted 1 year, 6 months ago

B: If two features in the dataset are perfectly linearly dependent, it means that one feature can be expressed as a linear combination of the other. This can create a singular matrix during optimization, as the linear model would be trying to fit a linear equation to a dataset where one variable is fully determined by the other. This would lead to an ill-defined optimization problem, as there would be no unique solution that minimizes the sum of the squares of the residuals. This could lead to problems during training, as the model would not be able to find appropriate parameter values to fit the data.

upvoted 7 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: B

Option B

upvoted 1 times

AjoseO 1 year, 5 months ago

Selected Answer: B

The presence of linearly dependent features means that they are redundant, and provide no additional information to the model.

This can result in a matrix that is not invertible, which is a requirement for solving a linear least squares regression problem. The presence of a singular matrix can also cause numerical instability and make it impossible to find an optimal solution to the optimization problem.

upvoted 4 times

yemauricio 1 year, 7 months ago

Selected Answer: B

linera dependence creates singular matrix that causes problems at the moment we fit the modle

upvoted 4 times

wisoxe8356 1 year, 7 months ago

Selected Answer: B

https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea
B - two features are perfectly linearly dependent = singular matrix during optimization
Not D - Not100% correct (as Multicollinearity happens when independent variables in the regression model are highly correlated to each other) they can still be independent variables

upvoted 3 times

ovokpus 2 years ago

Selected Answer: D

Consider one of the 5 assumptions of linear regression. This situation violates the assumption of "No multicollinearity between feature variables"

Hence, D

upvoted 3 times

jerto97 2 years, 8 months ago

B. See the multicollinearity problem in wikipedia https://en.wikipedia.org/wiki/Multicollinearity (second paragraph)

upvoted 4 times

takahirokoyama 2 years, 8 months ago

This issue is overfitting.

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 83 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 83
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A real estate company wants to create a machine learning model for predicting housing prices based on a historical dataset. The dataset contains 32 features.
Which model will meet the business requirement?

A. Logistic regression

B. Linear regression

C. K-means

D. Principal component analysis (PCA)

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 2:30 p.m.*

Comments

SophieSu Highly Voted 2 years, 8 months ago
B is the correct answer.
upvoted 19 times

Mickey321 Most Recent 10 months, 2 weeks ago
Selected Answer: B
Linear regression
upvoted 1 times

kaike_reis 11 months, 2 weeks ago
Selected Answer: B
B, the only model for regression in the options.
upvoted 1 times

jonsnow777 2 years, 5 months ago
Selected Answer: B
Answer B
upvoted 2 times

ahquiceno 2 years, 9 months ago

Answer B.

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 82 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 82

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time.

Specifically, the Specialist must train a model that returns the probability that a given transaction may fraudulent.

How should the Specialist frame this business problem?

A. Streaming classification

B. Binary classification

C. Multi-category classification

D. Regression classification

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 2:29 p.m.*

Comments

ahquiceno Highly Voted 3 years, 9 months ago

Answer B.

upvoted 28 times

SophieSu Highly Voted 3 years, 9 months ago

B IS NOT CORRECT! Return the probability. Not the 1 or 0. D IS THE CORRECT ANSWER.

upvoted 14 times

mdboy93 1 year, 9 months ago

Regression Classification is a made-up term, any binary classifier makes decisions based on probability score.

upvoted 1 times

srinu3054 3 years, 9 months ago

there is nothing like regression classification. (instead it should have said logistic regression). It should be Binary. i.e., either fraud or non fraud. Even with probabilities, we have a threshold to decide the class.

upvoted 11 times

seanLu 3 years, 9 months ago

Logistic regression will give the probability, and logistic regression is a binary classification algorithm.

https://machinelearningmastery.com/logistic-regression-for-machine-learning/

upvoted 9 times

Seoyong Most Recent 8 months, 1 week ago

Streaming classification: is the process of organizing and categorizing large amounts of data that are continuously flowing. This data can include medical records, banking transactions, and internet records
Binary Classification: Logistic Regression
Multiclass Classification: Softmax regression
Regression Classification is a made-up term

upvoted 1 times

Seoyong 8 months, 1 week ago

Random forest is the most suitable model for predicting fraudulent transactions.

Answer is A

upvoted 1 times

ZumbaZim 1 year, 4 months ago

I always see that the community voting is more appropriate and the moderator answer looks

out to be on wrong side. I see this for almost in 1 out of 5 questions. Which answer should we consider here as right one ??

upvoted 1 times

endeesa 1 year, 7 months ago

Selected Answer: B

Its definitely a classification problem, and between Binary and Streaming classification. Binary classification makes more sence

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

Binary classification

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: B

B, easy.

upvoted 1 times

gusta_dantas 1 year, 11 months ago

B, obviously!

```
from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
log_reg.predict_proba(X_test)
```

=)

upvoted 3 times

rodrigus 2 years, 4 months ago

The correct solution obviously is binary classification. For the comment above that says that binary classfication doesn't returns a probablity (for example SVM(classification) only returns a class and logistic, RFClassifier, XGBoostClassifier gives a probability and also a class given a threshold), you should ask yourself if that a regressor model returns always a probability, that is, if there is a restriction in a regressor model to predict values only in [0,1].

upvoted 1 times

AjoseO 2 years, 5 months ago

Selected Answer: B

The Specialist is trying to determine whether a given transaction is fraudulent or not, which is a binary outcome (yes or no). Therefore, the problem should be framed as binary classification.

The goal is to predict the probability of a transaction being fraudulent or not, and based on that, the Specialist can make a binary decision (fraudulent or not).

upvoted 2 times

Tomatoteacher 2 years, 5 months ago

Selected Answer: B

This is just binary classification, I don't understand how it could be anything else

upvoted 3 times

Sneep 2 years, 6 months ago

It's B.

This business problem can be framed as a binary classification problem, where the goal is to predict whether a given transaction is fraudulent

(positive class) or not fraudulent (negative class). The model should output a probability for each transaction, indicating the likelihood that it is fraudulent.

upvoted 2 times

DS2021 2 years, 6 months ago

Selected Answer: D

should be D

upvoted 1 times

RLai 2 years, 6 months ago

Logistic regression models the probability of the default class (e.g. the first class).

For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height.

upvoted 1 times

theprismdata 3 years, 2 months ago

I think the answer is B, fraud has various cases which hard to define.
So, Classification result will be fraud or not fraud.
If Multi-category classfication, must define case of fraud in detaily

upvoted 1 times

theprismdata 3 years, 2 months ago

More specifically, anomaly detection model will be needed

upvoted 1 times

KM226 3 years, 6 months ago

Selected Answer: B

I believe the answer is B: it a binary classification problem because we are classifying an observation into one of two categories and the target variable in this problem is limited to two options: fraudulent or not fraudulent

upvoted 3 times

lesh3000 3 years, 8 months ago

well, regression classification is bullshit, I hope they formulate their questions better on the real exam. binary classification gives probability between 0 and 1

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 81 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 81

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist is training a multilayer perception (MLP) on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve and acceptable recall metric. The Data Scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

Which techniques should be used to meet these requirements?

A. Gather more data using Amazon Mechanical Turk and then retrain

B. Train an anomaly detection model instead of an MLP

C. Train an XGBoost model instead of an MLP

D. Add class weights to the MLP's loss function and then retrain

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 2:29 p.m.*

Comments

rhuanca Highly Voted 2 years, 1 month ago

I believe is C, because we already made all changes possible in MLP hidden layers and the results have not improved then we must change model so XGBoot seems the best option

upvoted 5 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: D

In this case, the data scientist is training a multilayer perceptron (MLP), which is a type of neural network, on a dataset with multiple classes. The target class of interest is unique compared to the other classes within the dataset, but it does not achieve an acceptable recall metric. Recall is a measure of how well the model can identify the relevant examples from the minority class. The data scientist has already tried varying the number and size of the MLP's hidden layers, which has not significantly improved the results. A solution to improve recall must be implemented as quickly as possible.

upvoted 2 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: D

The fastest one is D

upvoted 1 times

ADVIT 1 year ago

"quickly as possible" mean do not change to new stuff, so it's D.

upvoted 1 times

kukreti18 1 year, 1 month ago

Not C, as the question ask for a quick solution.
I accept D.

upvoted 1 times

vbal 1 year, 1 month ago

Answer C : https://towardsdatascience.com/boosting-techniques-in-python-predicting-hotel-cancellations-62b7a76ffa6c

upvoted 1 times

AjoseO 1 year, 5 months ago

Selected Answer: D

Adding class weights to the MLP's loss function balances the class frequencies in the cost function during training, so the optimization process focuses more on the underrepresented class, improving recall.

upvoted 3 times

Tomatoteacher 1 year, 5 months ago

Selected Answer: D

I have done this before, class weights help with unbalanced data. Only logical solution that would help if not done, XGBoost could be different, but who knows, both NNs and XGBoost have comparable performance. Answer D!

upvoted 4 times

hamuozi 1 year, 9 months ago

Selected Answer: D

In this example, it is necessary to improve recall as soon as possible, so instead of creating additional datasets, it is effective to change the weight of each class during learning.

upvoted 4 times

victorlifan 1 year, 10 months ago

C: 'distinct' indicates we can simplify this as a binary classification problem; then, NN is just overkill. plus, retraining a NN is much slower than training an XGboost model

upvoted 2 times

exam_prep 2 years, 1 month ago

I feel answer is B. Question says Target is different than the input data which is hint for anomaly detection.

upvoted 2 times

kaike_reis 11 months, 2 weeks ago

stop overthink

upvoted 1 times

KM226 2 years, 6 months ago

I believe the answer is C because we need to use hyperparameters to improve model performance.

upvoted 2 times

ksarda11 2 years, 8 months ago

In case of the quickest possible way, D seems fine. For XGBoost, it will take a bit of time to code again

upvoted 4 times

ahquiceno 2 years, 9 months ago

For me Answer A. Why no other model instead xgBoost, the model need more labeled data to be trained and learn more positive examples.

upvoted 2 times

SophieSu 2 years, 8 months ago

A is incorrect. Even if you hire Amazon Mechanical Turk, you won't have more data. This question is NOT asking about "labeling".

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 80 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 80
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A credit card company wants to build a credit scoring model to help predict whether a new credit card applicant will default on a credit card payment. The company has collected data from a large number of sources with thousands of raw attributes. Early experiments to train a classification model revealed that many attributes are highly correlated, the large number of features slows down the training speed significantly, and that there are some overfitting issues.
The Data Scientist on this project would like to speed up the model training time without losing a lot of information from the original dataset.
Which feature engineering technique should the Data Scientist use to meet the objectives?

A. Run self-correlation on all features and remove highly correlated features

B. Normalize all numerical values to be between 0 and 1

C. Use an autoencoder or principal component analysis (PCA) to replace original features with new features

D. Cluster raw data using k-means and use sample data from each cluster to build a new dataset

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 2:10 p.m.*

Comments

ahquiceno Highly Voted 3 years, 9 months ago

Answer C. Need reduce the features preserving the information on it this is achieve using PCA.

upvoted 26 times

Dr_Kiko 3 years, 8 months ago

without losing a lot of information from the original dataset
since when PCA retains information?

upvoted 3 times

VinceCar 2 years, 8 months ago

PCA helps to speed up the training

upvoted 4 times

[Removed] Highly Voted 3 years, 9 months ago

Answer is A, because one must avoid information loss that PCA or autoencoders introduce through new features (https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/). Otherwise, I would perform C.

upvoted 6 times

SophieSu 3 years, 8 months ago

If you REMOVE highly correlated features(that means in pairs), the model lost a lot of information.

upvoted 4 times

rodrigus 2 years, 4 months ago

A doesn't have sense. Self-correlation is for times series data, not for pair correlation

upvoted 2 times

xicocaio Most Recent 9 months, 2 weeks ago

Selected Answer: A

This question can be misleading.

I would choose A if self-correlation in the dataset is meaning pair-wise correlation, this is the most typical approach in real life. But if self-correlation means auto-correlation as in the time-series treatment, then it is wrong.

Issues with answer C: Autoencoders are notorious for being hard to interpret. With PCA it is possible, but definitely not easy if you have a large dataset. In real life with this scenario, you would always go with pairwise correlation as the most simple yet effective approach.

upvoted 1 times

Giodefa96 11 months, 2 weeks ago

Selected Answer: C

Answer is C

upvoted 1 times

geoan13 1 year, 8 months ago

Answer C
PCA (Principal Component Analysis) takes advantage of multicollinearity and combines the highly correlated variables into a set of uncorrelated variables. Therefore, PCA can effectively eliminate multicollinearity between features.
https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b#:~:text=PCA%20(Principal%20Component%20Analysis)%20takes,effectively%20eliminate%20multicollinearity%20between%20features.

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: C

Option C

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: C

An autoencoder is a type of neural network that can learn a compressed representation of the input data, called the latent space, by encoding and decoding the data through multiple hidden layers1. PCA is a statistical technique that can reduce the dimensionality of the data by finding a set of orthogonal axes, called the principal components, that capture the most variance in the data2. Both methods can transform the original features into new features that are lower-dimensional, uncorrelated, and informative.

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: C

C is the correct.
Self-correlation is for time series, which is not mention here. Besides that, even if was correlation only, try to do this in thousand features...

upvoted 1 times

vbal 2 years, 1 month ago

A . run correlation matrix and remove highly correlated features.

upvoted 1 times

JK1977 2 years, 1 month ago

Selected Answer: C

PCA for feature reduction

upvoted 1 times

GOSD 2 years, 2 months ago

is it just me or is every 15th answer here PCA?

upvoted 2 times

oso0348 2 years, 3 months ago

Selected Answer: C

Using an autoencoder or PCA can help reduce the dimensionality of the dataset by creating new features that capture the most important information in the original dataset while discarding some of the noise and highly correlated features. This can help speed up the training time and reduce overfitting issues without losing a lot of information from the original dataset. Option A may remove too many features and may not capture all the important information in the dataset, while option B only rescales the data and does not address the issue of highly correlated

features. Option D is not a feature engineering technique and may not be an effective way to reduce the dimensionality of the dataset.

upvoted 1 times

Paolo991 2 years, 3 months ago

Selected Answer: C

PCA builds new features starting from high correlated ones. So it matches the question

upvoted 1 times

Sneep 2 years, 6 months ago

It's C.

The Data Scientist should use principal component analysis (PCA) to replace the original features with new features. PCA is a technique that reduces the dimensionality of a dataset by projecting it onto a lower-dimensional space, while preserving as much of the original variation as possible. This can help to speed up the training time of the model and reduce overfitting issues, without losing a significant amount of information from the original dataset.

upvoted 1 times

Aninina 2 years, 6 months ago

Selected Answer: C

C: PCA is the solution

upvoted 1 times

ovokpus 3 years ago

Selected Answer: C

Correction to C. Removing correlated features from hundreds of columns will be tedious and time consuming. PCA is the way to go here.

Apologies for the flip

upvoted 2 times

ovokpus 3 years ago

Selected Answer: A

Answer is A. Eliminate features that are highly correlated. This will not compromise the quality of the feature space as much as PCA would.

upvoted 1 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 79 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 79
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A trucking company is collecting live image data from its fleet of trucks across the globe. The data is growing rapidly and approximately 100 GB of new data is generated every day. The company wants to explore machine learning uses cases while ensuring the data is only accessible to specific IAM users.
Which storage option provides the most processing flexibility and will allow access control with IAM?

A. Use a database, such as Amazon DynamoDB, to store the images, and set the IAM policies to restrict access to only the desired IAM users.

B. Use an Amazon S3-backed data lake to store the raw images, and set up the permissions using bucket policies.

C. Setup up Amazon EMR with Hadoop Distributed File System (HDFS) to store the files, and restrict access to the EMR instances using IAM policies.

D. Configure Amazon EFS with IAM policies to make the data available to Amazon EC2 instances owned by the IAM users.

Show Suggested Answer

by Paul_NoName at *Feb. 3, 2021, 7:46 p.m.*

Comments

Paul_NoName Highly Voted 3 years, 9 months ago
B is the right answer
upvoted 33 times

clawo Highly Voted 3 years, 5 months ago
Selected Answer: B
B to use as storage with policies
upvoted 7 times

xicocaio Most Recent 9 months, 2 weeks ago
Selected Answer: B
- Amazon S3-backed data lake: S3 is the best storage option for large and rapidly growing datasets like images from trucks. S3 scales easily, handles large volumes of data, and is cost-effective for long-term storage, making it a natural choice for this scenario.
- IAM access control: You can use bucket policies in S3 to set very specific access controls, ensuring that only certain IAM users have permission to access or modify the data. This satisfies the requirement for access control using IAM.
- Processing flexibility: Storing the images in S3 offers flexibility for future machine learning use cases. The data stored in S3 can easily be integrated with other AWS services like SageMaker, Athena, EMR, and more for processing and analysis.

upvoted 1 times

endeesa 1 year, 7 months ago

Selected Answer: B

EMR/HDFS is not more 'flexible' than S3

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

A. NO - volume too big for a DB
B. YES
C. NO - instance access will not control HDFS access
D. NO - EFS does not use IAM policies (it is unix)

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

S3 indeed

upvoted 1 times

JK1977 2 years, 1 month ago

Selected Answer: B

S3 always

upvoted 1 times

Nadia0012 2 years, 4 months ago

Selected Answer: B

I would say the answer is B not because of the cost on EMR,. that is also a current answer. however: "most processing flexibility" indicates that S3 is a better option. because all ML solutions and work flows integrate with S3. it hasn't spoken what the ML solution and which services so I take the safe side and go with S3

upvoted 2 times

KlaudYu 2 years, 11 months ago

Selected Answer: B

C is not affordable because it is ephemeral storage. https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html "HDFS is used by the master and core nodes. One advantage is that it's fast; a disadvantage is that it's ephemeral storage which is reclaimed when the cluster ends. It's best used for caching the results produced by intermediate job-flow steps."

upvoted 4 times

ZSun 2 years, 2 months ago

the question does not require long-term storage.

upvoted 1 times

geekgirl007 3 years, 6 months ago

Selected Answer: C

C is correct. it says real time data and to be used for ml process so EMR more suitable. also S3 bucket policies not same as IAM users so B is not correct.

upvoted 4 times

ovokpus 3 years ago

Why will you need to spin up servers (EMR) just to store visual data for ML?

upvoted 5 times

Abdo702 3 years, 8 months ago

I think Amazon EMR is more appropriate, as the data scheme stated is a big data scheme.

https://aws.amazon.com/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc

upvoted 3 times

Sourabh1703 3 years, 5 months ago

IAM support is required for storage feature , that is not possible as per options described as IAM is supported for HDFS for the instance running on top of it, hence B should be correct

upvoted 2 times

Vita_Rasta84444 3 years, 8 months ago

B is the right answer

upvoted 2 times

srinu3054 3 years, 8 months ago

S3 is the easy, scalable and secure option to store the image data.

upvoted 1 times

astonm13 3 years, 8 months ago

B is the right answer

upvoted 1 times

zzaibis 3 years, 9 months ago

B is an appropriate choice

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 78 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 78

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist previously trained a logistic regression model using scikit-learn on a local machine, and the Specialist now wants to deploy it to production for inference only.
What steps should be taken to ensure Amazon SageMaker can host a model that was trained locally?

A. Build the Docker image with the inference code. Tag the Docker image with the registry hostname and upload it to Amazon ECR.

B. Serialize the trained model so the format is compressed for deployment. Tag the Docker image with the registry hostname and upload it to Amazon S3.

C. Serialize the trained model so the format is compressed for deployment. Build the image and upload it to Docker Hub.

D. Build the Docker image with the inference code. Configure Docker Hub and upload the image to Amazon ECR.

Show Suggested Answer

by Paul_NoName at *Feb. 3, 2021, 7:44 p.m.*

Comments

arulrajjayaraj Highly Voted 2 years, 9 months ago

Ans : A Refer the below :
https://sagemaker-workshop.com/custom/containers.html

upvoted 19 times

Paul_NoName Highly Voted 2 years, 9 months ago

A
https://sagemaker-workshop.com/custom/containers.html

upvoted 7 times

endeesa Most Recent 7 months, 2 weeks ago

Selected Answer: A

You need the container to be hosted on ECR.

upvoted 1 times

loict 10 months ago

Selected Answer: A

A. YES - the inference code is built after inspecting the coefficient of the Linear Model (or, alternatively, the model can be serialized via pickle and

the inference code is simply to unserialized the mode); ECR is only registry supported by SageMaer; tagging the Docker image with the registry hostname (eg. docker tag image1 public.ecr.aws/g6h7x5m5/image1) is required so that the docker push command knows where to push the image
B. NO - no need to compress; image must be on ECR
C. NO - no need to compress; image must be on ECR
D. NO - image must be on ECR

upvoted 5 times

Khalil11 1 year, 3 months ago

Selected Answer: A

A is the right answer

upvoted 3 times

Nadia0012 1 year, 4 months ago

Selected Answer: A

For SageMaker to run a container for training or hosting, it needs to be able to find the image hosted in the image repository, Amazon Elastic Container Registry (Amazon ECR). The three main steps to this process are building locally, tagging with the repository location, and pushing the image to the repository.

upvoted 4 times

geekgirl007 2 years, 6 months ago

Selected Answer: A

A for sure.

upvoted 3 times

astonm13 2 years, 8 months ago

Answer is A.

upvoted 3 times

cnethers 2 years, 8 months ago

Docker Hub is a repository so ANS D makes no sense. Option A is the way to go.

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 77 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 77

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist needs to move and transform data in preparation for training. Some of the data needs to be processed in near-real time, and other data can be moved hourly. There are existing Amazon EMR MapReduce jobs to clean and feature engineering to perform on the data.

Which of the following services can feed data to the MapReduce jobs? (Choose two.)

A. AWS DMS

B. Amazon Kinesis

C. AWS Data Pipeline

D. Amazon Athena

E. Amazon ES

Show Suggested Answer

by Joe_Zhang at *Feb. 1, 2021, 11:43 p.m.*

Comments

Joe_Zhang Highly Voted 3 years, 3 months ago

should be BC

upvoted 32 times

Denise123 Most Recent 10 months, 3 weeks ago

It is obviously B and C, I am frustrated with the number of wrong answers. Why the moderator's answers keep being super weird?

upvoted 2 times

Mickey321 1 year, 4 months ago

Selected Answer: BC

B for near real-time
C for hourly

upvoted 2 times

Khalil11 1 year, 9 months ago

The right answer is BC

upvoted 2 times

AjoseO 1 year, 11 months ago

Selected Answer: BC

AWS Data Pipeline (Option C) can be used to move the hourly data, as it provides a way to move data from various sources to Amazon EMR for processing.

Amazon Kinesis (Option B) can be used to process data in near-real time, as it is a real-time data streaming service that can handle large amounts of incoming data from multiple sources. The data can be fed to Amazon EMR MapReduce jobs for processing.

upvoted 4 times

DS2021 2 years ago

Selected Answer: BC

should be BC

upvoted 2 times

Peeking 2 years, 1 month ago

Selected Answer: BC

Kinesis for near realtime data and pipeline for the other data moved hourly.

upvoted 3 times

John_Pongthorn 2 years, 10 months ago

Selected Answer: BC

AWS ES is an elastic search , it is nothing to do with this question.

upvoted 3 times

scsas 2 years, 10 months ago

Kinesis data into EMR: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html

upvoted 1 times

apprehensive_scar 2 years, 11 months ago

BC. easy

upvoted 2 times

KM226 3 years ago

I believe the answer is BC

upvoted 1 times

Madwyn 3 years, 2 months ago

BD.
Data Pipeline is to orchestrate the workflow, how can that feed data to the MR jobs?

upvoted 2 times

Vita_Rasta84444 3 years, 2 months ago

Answer is B and C

upvoted 1 times

astonm13 3 years, 2 months ago

Answer is for sure BC

upvoted 3 times

takahirokoyama 3 years, 3 months ago

Ans is BC.
(https://aws.amazon.com/jp/emr/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc)

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 76 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 76
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company uses a long short-term memory (LSTM) model to evaluate the risk factors of a particular energy sector. The model reviews multi-page text documents to analyze each sentence of the text and categorize it as either a potential risk or no risk. The model is not performing well, even though the Data Scientist has experimented with many different network structures and tuned the corresponding hyperparameters.
Which approach will provide the MAXIMUM performance boost?

A. Initialize the words by term frequency-inverse document frequency (TF-IDF) vectors pretrained on a large collection of news articles related to the energy sector.

B. Use gated recurrent units (GRUs) instead of LSTM and run the training process until the validation loss stops decreasing.

C. Reduce the learning rate and run the training process until the training loss stops decreasing.

D. Initialize the words by word2vec embeddings pretrained on a large collection of news articles related to the energy sector.

Show Suggested Answer

by jiadong at *Feb. 3, 2021, 4:58 a.m.*

Comments

jiadong Highly Voted 3 years, 9 months ago
I think the right answer is D
upvoted 24 times

SophieSu Highly Voted 3 years, 9 months ago
D is correct.
C is not the best the answer because the question states that tuning parameters doesn't help a lot. Transfer learning would be better solution!
upvoted 11 times

xicocaio Most Recent 9 months, 2 weeks ago
Selected Answer: D
Using word2vec embeddings would give the model more accurate representations of words at the start, potentially leading to a significant performance boost for text classification tasks.
upvoted 1 times

ninomfr64 1 year ago
Selected Answer: D

A. NO, transfer learning helps, word2vec > TD-ITF as the first keeps into account part of the word context (there is a hyperparameter for this)
B. LTSM delivers better results wrt GRU which is in turn a compromise architecture to balance accuracy with training time/cost
C. Heperparameters tuning has been already applied, this will not help
D. YEs, transfer learning will help and word3vec is better option in this scenario

upvoted 2 times

3eb0542 1 year, 5 months ago

Selected Answer: D

How are the 'correct' answers being provided? I'm seeing so many answers that seem to be wrong and usually, the community vote seems to be correct. This is kind of frustrating.

upvoted 3 times

Mickey321 1 year, 10 months ago

Selected Answer: D

Word2vec is a technique that can learn distributed representations of words, also known as word embeddings, from large amounts of text data. Word embeddings can capture the semantic and syntactic similarities and relationships between words, and can be used as input features for neural network models. Word2vec can be trained on domain-specific corpora to obtain more relevant and accurate word embeddings for a particular task.

upvoted 3 times

kaike_reis 1 year, 11 months ago

Selected Answer: D

From my perspective, B and C are wrong because the DS already tried something close to this. D is correct.

upvoted 1 times

vbal 2 years, 1 month ago

I don't think High Dimensionality is take care by C2V; TF-IDF is required. A.

upvoted 1 times

Peeking 2 years, 7 months ago

Selected Answer: D

Transfer learning, in my experience, has been a good way to boost performance when hyperparameter tuning did not work.

upvoted 2 times

Sidekick 3 years, 1 month ago

The case ask for predicting labels for sentences, the appropriate algo should be "Text Classification" Which, just as "wrod2vec,i part of Blazing Text.

upvoted 1 times

julpeg 3 years, 3 months ago

Selected Answer: D

The answer should be D. My reasoning is that by using a word embedding which is trained on domain specific material, the embeddings between two words are more domain specific. This means that relations (good or bad) are represented in a better way, which also means that the model should be able to predict the results in a more accurate way.

upvoted 3 times

bitsplease 3 years, 5 months ago

both A & D "seem" correct, but word2vec takes ORDER of words into acc (to some extent)--while TF-IDF does not. Thus max boost is from D.

B,C are wrong because the DS has tried several network architectures (aka LSTM) and hyperparameter tuning (aka option C)

upvoted 6 times

ahmedelbhy 3 years, 8 months ago

i think answer is A as The model reviews multi-page text documents

upvoted 1 times

GiyeonShin 2 years, 6 months ago

I think that the general tf-idf vectors cannot be directly adapted to the deep learning model, because of the large dimension in vector values

upvoted 1 times

puffpuff 3 years, 9 months ago

I think it should be B
A/D are false flags because the question doesn't specify what kind of data engineering is currently done on the inputs, as a baseline
Per wikipedia, for GRUs, "GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets", which fits the context of a particular energy sector

upvoted 2 times

ChanduPatil 3 years, 9 months ago

why not B??

upvoted 1 times

GiyeonShin 2 years, 6 months ago

Generally, LSTM has the better performance then GRU in large datasets such as multi-page documents. GRU has advantages of memory

allocation and training time.

upvoted 1 times

GiyeonShin 2 years, 6 months ago

Early stopping can give the model better performance, but I think that the model needs more condition like patience value for early stopping. This is because the model doesn't always show the performance at its maximum when the validation loss stops decreasing.

upvoted 1 times

jkreddy 3 years, 9 months ago

It cannot be C, because hyper parameter tuning didnt work as given in question. Also, A and D are same, however, word2vec model internally implements tf-idf much more efficiently. So answer got to be D

upvoted 4 times

YJ4219 3 years, 9 months ago

but they need to classify the whole sentence i think for such a case we use object2vec not word2vec, but since it's not available in the answers, B is the only answer left.

upvoted 2 times

tmld 3 years, 9 months ago

I go for C

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 75 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 75
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist wants to determine the appropriate SageMakerVariantInvocationsPerInstance setting for an endpoint automatic scaling configuration. The Specialist has performed a load test on a single instance and determined that peak requests per second (RPS) without service degradation is about 20 RPS. As this is the first deployment, the Specialist intends to set the invocation safety factor to 0.5.
Based on the stated parameters and given that the invocations per instance setting is measured on a per-minute basis, what should the Specialist set as the
SageMakerVariantInvocationsPerInstance setting?

A. 10

B. 30

C. 600

D. 2,400

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 1:23 p.m.*

Comments

Paul_NoName Highly Voted 2 years, 8 months ago

C is correct .
SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60
AWS recommended Saf_fac =0 .5

upvoted 14 times

ahquiceno Highly Voted 2 years, 8 months ago

Answer C: SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60
https://docs.aws.amazon.com/sagemaker/latest/dg/endpoint-scaling-loadtest.html

upvoted 6 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: C

To calculate the SageMakerVariantInvocationsPerInstance setting, we can use the following equation from the web search results1:
SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60
Where MAX_RPS is the maximum RPS that the variant can handle, SAFETY_FACTOR is the safety factor that we choose to ensure that we don't

exceed the maximum RPS, and 60 is to convert from RPS to invocations-per-minute.
Plugging in the given values, we get:
SageMakerVariantInvocationsPerInstance = (20 * 0.5) * 60 SageMakerVariantInvocationsPerInstance = 10 * 60
SageMakerVariantInvocationsPerInstance = 600
Therefore, the Specialist should set the SageMakerVariantInvocationsPerInstance setting to 600.

upvoted 1 times

jackzhao 1 year, 3 months ago

SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60

upvoted 1 times

King_Chess1 1 year, 5 months ago

SageMakerVariantInvocationsPerInstance = (MAX_RPS * SAFETY_FACTOR) * 60
(20RPS * 0.5Safety Factor) * 60
(10)*60 = 600
Answer C

upvoted 1 times

Peeking 1 year, 7 months ago

Selected Answer: C

Maximum request at peak time = 20 RPS = 20x60 = 1200RPM
Safety factor of 0.5 = 1200*0.5 = 600
Basic setting of parameter = 600 (requests per minutes)

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 74 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 74

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning team runs its own training algorithm on Amazon SageMaker. The training algorithm requires external assets. The team needs to submit both its own algorithm code and algorithm-specific parameters to Amazon SageMaker.
What combination of services should the team use to build a custom algorithm in Amazon SageMaker? (Choose two.)

A. AWS Secrets Manager

B. AWS CodeStar

C. Amazon ECR

D. Amazon ECS

E. Amazon S3

Show Suggested Answer

by ahquiceno at *Feb. 3, 2021, 1:32 p.m.*

Comments

Paul_NoName Highly Voted 3 years, 9 months ago

CE is the right answer. ECR uses ECS internally while using SGM.

upvoted 11 times

[Removed] 3 years, 9 months ago

CE based on criteria and this documentation: https://docs.aws.amazon.com/sagemaker/latest/dg/sagemaker-mkt-create-model-package.html

"For Location of inference image, type the path to the image that contains your inference code. The image must be stored as a Docker container in Amazon ECR.
For Location of model data artifacts, type the location in S3 where your model artifacts are stored."

upvoted 6 times

ZSun 2 years, 2 months ago

the answer is correct but the explanation is completely wrong.
The question is about how to create your own algorithm using container, not "put the inference in market" (which is your resource link).
the right citation should be "Adapting your own training container": create s3 to store model articraft, and push code to ECR.

upvoted 3 times

SophieSu Highly Voted 3 years, 8 months ago

CE IS THE CORRECT ANSWER 100%

upvoted 5 times

MultiCloudIronMan Most Recent 9 months, 3 weeks ago

Selected Answer: CE

Amazon ECR (Option C): Amazon Elastic Container Registry (ECR) is used to store, manage, and deploy Docker container images. The team can package their custom algorithm code into a Docker container and store it in Amazon ECR1.
Amazon S3 (Option E): Amazon Simple Storage Service (S3) is used to store external assets and data. The team can store the algorithm-specific parameters and any other required data in Amazon S3

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: CE

Amazon ECR is a fully managed container registry service that allows users to store, manage, and deploy Docker container images. Amazon SageMaker supports using custom Docker images for training and inference, which can contain the user's own training algorithm and any external assets or dependenciesAd1. The user can push their Docker image to Amazon ECR and then reference it in their Amazon SageMaker training job configurationAd1.

upvoted 1 times

jackzhao 2 years, 3 months ago

CE is correct!

upvoted 1 times

Valcilio 2 years, 4 months ago

ECR for the code, S3 for the parameters!

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: CE

C contain the algorithm's image and E contain algorithm's parameters.

upvoted 2 times

randomnamer 3 years, 8 months ago

The location of the model artifacts. Model artifacts can either be packaged in the same Docker container as the inference code or stored in Amazon S3. Not so sure.

upvoted 1 times

cnethers 3 years, 9 months ago

https://aws.amazon.com/blogs/machine-learning/bringing-your-own-custom-container-image-to-amazon-sagemaker-studio-notebooks/
If you wish to use your private VPC to securely bring your custom container, you also need the following:

A VPC with a private subnet
VPC endpoints for the following services:
Amazon Simple Storage Service (Amazon S3)
Amazon SageMaker
Amazon ECR
AWS Security Token Service (AWS STS)
CodeBuild for building Docker containers

Answer C+E

upvoted 3 times

ahquiceno 3 years, 9 months ago

For me CD. needs storage and create a custom docker using ECR to store it.

upvoted 1 times

ahquiceno 3 years, 8 months ago

Sorry, CE is correct.

upvoted 1 times

gcpwhiz 3 years, 8 months ago

Sagemaker will spin up the instances needed with the right image. No need to use ECS. CE is right

upvoted 1 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 73 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 73
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

An aircraft engine manufacturing company is measuring 200 performance metrics in a time-series. Engineers want to detect critical manufacturing defects in near- real time during testing. All of the data needs to be stored for offline analysis.
What approach would be the MOST effective to perform near-real time defect detection?

A. Use AWS IoT Analytics for ingestion, storage, and further analysis. Use Jupyter notebooks from within AWS IoT Analytics to carry out analysis for anomalies.

B. Use Amazon S3 for ingestion, storage, and further analysis. Use an Amazon EMR cluster to carry out Apache Spark ML k-means clustering to determine anomalies.

C. Use Amazon S3 for ingestion, storage, and further analysis. Use the Amazon SageMaker Random Cut Forest (RCF) algorithm to determine anomalies.

D. Use Amazon Kinesis Data Firehose for ingestion and Amazon Kinesis Data Analytics Random Cut Forest (RCF) to perform anomaly detection. Use Kinesis Data Firehose to store data in Amazon S3 for further analysis.

Show Suggested Answer

by Joe_Zhang at *Feb. 1, 2021, 10:42 p.m.*

Comments

Joe_Zhang Highly Voted 3 years, 9 months ago
D near-real time
upvoted 43 times

DimLam 1 year, 8 months ago
The main problem with D is that Amazon Kinesis Data Firehose can not be a source service for Amazon Kinesis Data Analytics.

The answer would be correct if it said
"Using Amazon Kinesis Data Stream to ingest data, using Amazon Kinesis Data Analytics for defect detection and using Amazon Kinesis Data Firehose for storing data for further Analysis"

https://docs.aws.amazon.com/firehose/latest/dev/create-name.html
upvoted 2 times

VR10 1 year, 4 months ago

Actually Kinesis Data Firehose can be used for Data Ingestion.
So the correct option is still D

upvoted 2 times

cnethers Highly Voted 3 years, 9 months ago

Glad we are all in agreement D is the correct answer

upvoted 17 times

xicocaio Most Recent 9 months, 2 weeks ago

Selected Answer: D

Amazon Kinesis Data Firehose is a fully managed service for real-time data ingestion, which fits the requirement for near-real-time defect detection. It can ingest large volumes of data from various sources and reliably load the data into other AWS services like Amazon S3 for storage.
Amazon Kinesis Data Analytics with Random Cut Forest (RCF) is highly efficient for detecting anomalies in streaming data in near real time, which is what the engineers need to catch manufacturing defects during testing.
After detecting anomalies, the data can be stored in Amazon S3 via Kinesis Data Firehose for offline analysis.

upvoted 1 times

SandyHenshaw 11 months, 3 weeks ago

Selected Answer: D

D - firehose for near realtime

upvoted 1 times

VR10 1 year, 4 months ago

Selected Answer: D

Kinesis Data Firehose is a fully managed service that can ingest streaming data and load it into destinations like S3, Redshift, Elasticsearch. and with Kinesis Data Analytics and RCF and then Data Firehose again to store on S3.
D is the best choice.

upvoted 1 times

fa0d8b7 1 year, 7 months ago

https://docs.aws.amazon.com/managed-flink/latest/java/get-started-exercise-fh.html

upvoted 2 times

endeesa 1 year, 7 months ago

Selected Answer: D

Kinesis seems like the only viable option

upvoted 1 times

akgarg00 1 year, 7 months ago

The answer is D. Since, data is continuously coming in Kinesis datafirehose is our streaming application (also we need near Real time defect detection and storage in S3) and anomaly detection can be done by kinesis data application (RCF algorithm).

upvoted 1 times

AmeeraM 1 year, 9 months ago

Selected Answer: D

D, near real-time ingestion is the key

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: D

A. NO - AWS IoT will first store the data, then make it available for Analytics/Jupyter (https://docs.aws.amazon.com/iotanalytics/latest/userguide/welcome.html); so not real-time
B. NO - not realtime to store the data before analytics
C. NO - not realtime to store the data before analytics
D. YES - real-time pipe, RCF best for anomalities

upvoted 1 times

DavidRou 1 year, 10 months ago

Selected Answer: D

How can someone use S3 for ingestion? Firehose is the right answer

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: D

This option meets the requirements of performing near-real time defect detection, storing all the data for offline analysis, and handling 200 performance metrics in a time-series. Amazon Kinesis Data Firehose is a fully managed service that can ingest streaming data from various sources and deliver it to destinations such as Amazon S3, Amazon OpenSearch Service, and Amazon Redshift. Amazon Kinesis Data Analytics is a service that can process streaming data using SQL or Apache Flink applications. Amazon Kinesis Data Analytics provides a built-in RANDOM_CUT_FOREST function, a machine learning algorithm that can detect anomalies in streaming data1. This function can handle high-dimensional data and assign an anomaly score to each record based on how distant it is from other records1. The anomaly scores can then be delivered to another destination using Kinesis Data Firehose or consumed by other applications using Kinesis Data Streams.

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 72 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 72
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist is building a model to predict customer churn using a dataset of 100 continuous numerical features. The Marketing team has not provided any insight about which features are relevant for churn prediction. The Marketing team wants to interpret the model and see the direct impact of relevant features on the model outcome. While training a logistic regression model, the Data Scientist observes that there is a wide gap between the training and validation set accuracy.
Which methods can the Data Scientist use to improve the model performance and satisfy the Marketing team's needs? (Choose two.)

A. Add L1 regularization to the classifier

B. Add features to the dataset

C. Perform recursive feature elimination

D. Perform t-distributed stochastic neighbor embedding (t-SNE)

E. Perform linear discriminant analysis

Show Suggested Answer

by bluer1 at *April 30, 2022, 7:18 p.m.*

Comments

bluer1 Highly Voted 2 years, 2 months ago

AC - correct answer

upvoted 13 times

lynn22 Most Recent 7 months ago

Selected Answer: AE

I think ACE are all correct

upvoted 1 times

loict 10 months ago

Selected Answer: AC

A. YES - standard for overfitting
B. NO - we have already too much overfitting
C. YES - feature elimination can reduce model complexity and thus overfitting
D. NO - that does dimensionnality reduction to 2D or 3D, for visualization; we want more than a few features
E. NO - LDA is an alternative to logistic regression; it may not address overfitting

upvoted 4 times

Mickey321 10 months, 2 weeks ago

Selected Answer: AC

A due to fitting
C Recursive feature elimination (RFE) is a wrapper method that iteratively removes features based on their importance scores from a classifier. RFE starts with all features and then eliminates the least important ones until a desired number of features is reached. This can help to reduce the dimensionality of the dataset and improve the model performance by removing irrelevant or redundant features. The Marketing team can then interpret the model by looking at the remaining features and their importance scores.

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: AC

AC are the correct

upvoted 1 times

earthMover 1 year, 1 month ago

Selected Answer: AC

How can we add features to the dataset provided.... we can't make them up from thin air. Hopefully the moderators can provide some insight on this. I was thinking of paying for this site but the answers are all over the place.

upvoted 1 times

bakarys 1 year, 4 months ago

Selected Answer: AC

A. Add L1 regularization to the classifier and C. Perform recursive feature elimination are the methods that can be used to improve the model performance and satisfy the Marketing team's needs.

Explanation:
A. Adding L1 regularization to the logistic regression classifier can help to improve the model performance and reduce overfitting. This can also help to highlight the relevant features for churn prediction as L1 regularization can shrink the coefficients of irrelevant features to zero.

C. Recursive feature elimination can be used to select the most relevant features for the model. This can help to improve the model performance and highlight the relevant features for churn prediction.

upvoted 3 times

AjoseO 1 year, 5 months ago

Selected Answer: AC

A. Adding L1 regularization can help to reduce overfitting by shrinking the coefficients of less important features towards zero, which can improve the model's generalization performance on the validation set.

C. Recursive feature elimination is a feature selection technique that removes the least important feature at each iteration and trains the model on the remaining features until a desired number of features is reached. This method can be used to identify the most relevant features for the prediction task and reduce the dimensionality of the dataset, leading to improved model performance and interpretability for the Marketing team.

upvoted 2 times

wisoxe8356 1 year, 7 months ago

AC -
Key: logistic regression model = non linear in terms of Odds and Probability, however it is linear in terms of Log Odds.
Key: Large gap between training & validation = overfitting
=> 5 techniques to prevent overfitting:
1. Simplifying the model | 2. Early stopping
3. Use data argumentation | 4. Use regularization | 5. Use dropouts

A - yes to avoid overfitting (although i am thinking it is talking about regressor)
Not B - add feature will lead to overfitting
C - feature elimination - prevent overfitting
Not D - t-SNE is a nonlinear dimensionality reduction technique
Not E - find feature correlation only - Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics and other fields, to find a linear combination of features that characterizes or separates two or more classes of objects or events.

upvoted 4 times

itallomd 1 year, 7 months ago

L1 won't do naturally the feature elimination?
I guess AB

upvoted 1 times

Atreides457 1 year, 10 months ago

why not A & D? or C & D?
does not t-SNE grant the marketing team's wish for visualization of relationships? or are we to presume that A&C are best as C (recursive feature elimination) grants us some visualization of feature importance.

upvoted 2 times

tgaos 2 years, 1 month ago

Selected Answer: AC

AC is correct

upvoted 3 times

NeverMinda 2 years, 1 month ago

Selected Answer: AC

overfitting: add regularization, remove features

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 71 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 71
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist needs to migrate an existing on-premises ETL process to the cloud. The current process runs at regular time intervals and uses PySpark to combine and format multiple large data sources into a single consolidated output for downstream processing.
The Data Scientist has been given the following requirements to the cloud solution:
⇨ Combine multiple data sources.
⇨ Reuse existing PySpark logic.
⇨ Run the solution on the existing schedule.
⇨ Minimize the number of servers that will need to be managed.
Which architecture should the Data Scientist use to build this solution?

A. Write the raw data to Amazon S3. Schedule an AWS Lambda function to submit a Spark step to a persistent Amazon EMR cluster based on the existing schedule. Use the existing PySpark logic to run the ETL job on the EMR cluster. Output the results to a ג€processedג€ location in Amazon S3 that is accessible for downstream use.

B. Write the raw data to Amazon S3. Create an AWS Glue ETL job to perform the ETL processing against the input data. Write the ETL job in PySpark to leverage the existing logic. Create a new AWS Glue trigger to trigger the ETL job based on the existing schedule. Configure the output target of the ETL job to write to a ג€processedג€ location in Amazon S3 that is accessible for downstream use.

C. Write the raw data to Amazon S3. Schedule an AWS Lambda function to run on the existing schedule and process the input data from Amazon S3. Write the Lambda logic in Python and implement the existing PySpark logic to perform the ETL process. Have the Lambda function output the results to a ג€processedג€ location in Amazon S3 that is accessible for downstream use.

D. Use Amazon Kinesis Data Analytics to stream the input data and perform real-time SQL queries against the stream to carry out the required transformations within the stream. Deliver the output results to a ג€processedג€ location in Amazon S3 that is accessible for downstream use.

Show Suggested Answer

by Joe_Zhang at *Feb. 1, 2021, 10:36 p.m.*

Comments

Paul_NoName Highly Voted 3 years, 9 months ago
B it is .
upvoted 29 times

[Removed] 3 years, 9 months ago

I agree, B is serverless and reuses Pyspark. Similar example shown here: https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming-python-samples-medicaid.html

upvoted 11 times

SophieSu Highly Voted 3 years, 9 months ago

A is not correct because Minimize the number of servers that will need to be managed. EMR is not server-less.
B is correct. AWS Glue supports an extension of the PySpark Python dialect for scripting extract, transform, and load...
C is not correct because using Lambda for ETL you will not be able to Reuse existing PySpark logic
D is not correct because Kinesis is not server-less. And you can not Reuse existing PySpark logic

upvoted 12 times

xicocaio Most Recent 9 months, 2 weeks ago

Selected Answer: B

Option B (using AWS Glue for the ETL process) is the best solution for the described requirements.

A: This solution requires managing an Amazon EMR cluster, which would involve more server management than AWS Glue, violating the requirement to minimize the number of servers to be managed.

C: AWS Lambda is not ideal for this use case because it has resource limitations, including memory and execution time limits (15 minutes max), which might not be suitable for large-scale ETL operations involving PySpark logic.

D: Amazon Kinesis Data Analytics is focused on real-time stream processing, which doesn't fit the described scheduled batch processing scenario.

upvoted 1 times

akgarg00 1 year, 7 months ago

Answer is A, as B clearly mentions that Pyspark code is written with leverage from already existing code. Also, the server architecture used currently is on-premises which will have more servers that solution A.

upvoted 2 times

sonoluminescence 1 year, 8 months ago

Selected Answer: B

Amazon Kinesis Data Analytics is more suited for real-time processing and streaming data. The given use case does not indicate a need for real-time processing, so this might not be the best fit. Furthermore, it doesn't support PySpark natively.

upvoted 1 times

Shenannigan 1 year, 10 months ago

Selected Answer: B

Voted B based on the serverless (minimum servers) and https://docs.aws.amazon.com/glue/latest/dg/aws-glue-programming.html

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

Indeed B using Glue

upvoted 1 times

kaike_reis 1 year, 11 months ago

B is the correct.
A you have to manage EMR, so it's wrong.
D you don't use Spark, so it's wrong.
C you will not be using Spark, so it's wrong.

upvoted 1 times

Maaayaaa 2 years, 3 months ago

Selected Answer: B

B ticks all boxes. Minimize servers -> AWS managed services -> Glue.

upvoted 2 times

bakarys 2 years, 4 months ago

Selected Answer: A

Option A would be the best response for this scenario.

This solution allows the Data Scientist to reuse the existing PySpark logic while migrating the ETL process to the cloud. The raw data is written to Amazon S3, and a Lambda function is scheduled to trigger a Spark step on a persistent EMR cluster based on the existing schedule. The PySpark logic is used to run the ETL job on the EMR cluster, and the results are output to a processed location in Amazon S3 that is accessible for downstream use. This solution minimizes the number of servers that need to be managed, and it allows for a seamless migration of the existing ETL process to the cloud.

upvoted 1 times

sqavi 2 years, 5 months ago

Selected Answer: B

Option D is wrong it should be B

upvoted 1 times

Peeking 2 years, 7 months ago

D cannot be answer as there is no streaming data or Realtime processing.

upvoted 2 times

salads 2 years, 10 months ago

Selected Answer: B

the answer is b

upvoted 2 times

Nickname_L 3 years, 8 months ago

Answer should be B. Serverless, on a regular schedule (no real time requirement), reuses PySpark code in Glue ETL script.

upvoted 4 times

gcpwhiz 3 years, 8 months ago

Answer is B as they specifically ask about reusing existing PySpark, which can be done with Glue

upvoted 3 times

Aashi22 3 years, 8 months ago

https://docs.aws.amazon.com/glue/latest/dg/creating_running_workflows.html

upvoted 1 times

astonm13 3 years, 9 months ago

It is B. ! "Minimize number of servers to be managed". B is a Serverless solution which fulfils other requirements!

upvoted 2 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 70 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 70

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist uploads a dataset to an Amazon S3 bucket protected with server-side encryption using AWS KMS.
How should the ML Specialist define the Amazon SageMaker notebook instance so it can read the same dataset from Amazon S3?

A. Define security group(s) to allow all HTTP inbound/outbound traffic and assign those security group(s) to the Amazon SageMaker notebook instance.

B. ¡onfigure the Amazon SageMaker notebook instance to have access to the VPC. Grant permission in the KMS key policy to the notebook's KMS role.

C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.

D. Assign the same KMS key used to encrypt data in Amazon S3 to the Amazon SageMaker notebook instance.

Show Suggested Answer

by ahquiceno at *Feb. 1, 2021, 7:38 p.m.*

Comments

seanLu Highly Voted 3 years, 3 months ago

Should be C.
"You don’t need to specify the AWS KMS key ID when you download an SSE-KMS-encrypted object from an S3 bucket. Instead, you need the permission to decrypt the AWS KMS key.

When a user sends a GET request, Amazon S3 checks if the AWS Identity and Access Management (IAM) user or role that sent the request is authorized to decrypt the key associated with the object. If the IAM user or role belongs to the same AWS account as the key, then the permission to decrypt must be granted on the AWS KMS key’s policy."
https://aws.amazon.com/premiumsupport/knowledge-center/decrypt-kms-encrypted-objects-s3/?nc1=h_ls

upvoted 29 times

askaron Highly Voted 3 years, 3 months ago

Should be C.
I think it is not possible to assign a key directly to a Sagemaker notebook instance like D suggests.
Normally in AWS in general, IAM roles are used to do so. So C.

upvoted 6 times

james2033 Most Recent 10 months, 1 week ago

Selected Answer: C

'IAM role' principle of least privilege (PoLP)

upvoted 1 times

VR10 10 months, 3 weeks ago

Selected Answer: C

IAM roles securely provide temporary AWS credentials that services (like SageMaker notebooks) can assume to access other resources. This avoids using long-lived access keys or directly embedding API keys into code.
KMS Key Policy: This policy controls access to your KMS key. Granting the notebook's role permission within this policy lets SageMaker decrypt the data when reading from S3.

upvoted 1 times

endeesa 1 year, 1 month ago

Selected Answer: C

Seems to follow the best cloud authorization practice

upvoted 1 times

sonoluminescence 1 year, 2 months ago

Selected Answer: C

IAM role associated with the SageMaker notebook instance must be given permissions in the KMS key policy to decrypt the data using the KMS key that was used for encryption.

upvoted 1 times

AmeeraM 1 year, 3 months ago

Selected Answer: C

answer is C

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: C

Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role. To read data from Amazon S3 that is encrypted with AWS KMS, the Amazon SageMaker notebook instance needs to have both S3 read access and KMS decrypt permissions. This can be achieved by assigning an IAM role to the notebook instance that has the necessary policies attached, and by granting permission in the KMS key policy to that role.

upvoted 1 times

ADVIT 1 year, 6 months ago

C only.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: C

Should be C. The reference doc provided did not have any information about assigning keys to the notebook. Doing so become very cumbersome as you can have 100's of notebooks and its not scalable. Someone needs to moderate these answers.

upvoted 1 times

oso0348 1 year, 9 months ago

Selected Answer: C

To allow an Amazon SageMaker notebook instance to read a dataset stored in an Amazon S3 bucket that is protected with server-side encryption using AWS KMS, the ML Specialist should assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. The IAM role should have permissions to access the S3 bucket and the KMS key that was used to encrypt the data. This role should be granted permission in the KMS key policy to allow it to decrypt the data.

upvoted 1 times

Nadia0012 1 year, 10 months ago

Selected Answer: D

To encrypt the machine learning (ML) storage volume that is attached to notebooks, processing jobs, training jobs, hyperparameter tuning jobs, batch transform jobs, and endpoints, you can pass a AWS KMS key to SageMaker. If you don't specify a KMS key, SageMaker encrypts storage volumes with a transient key and discards it immediately after encrypting the storage volume. For notebook instances, if you don't specify a KMS key, SageMaker encrypts both OS volumes and ML data volumes with a system-managed KMS key.

upvoted 1 times

Nadia0012 1 year, 10 months ago

I correct myself- Option C is correct:
Background
AWS Key Management Service (AWS KMS) enables Server-side encryption to protect your data at rest. Amazon SageMaker training works with KMS encrypted data if the IAM role used for S3 access has permissions to encrypt and decrypt data with the KMS key. Further, a KMS key can also be used to encrypt the model artifacts at rest using Amazon S3 server-side encryption. Additionally, a KMS key can also be used to encrypt the storage volume attached to training, endpoint, and transform instances. In this notebook, we demonstrate SageMaker encryption capabilities using KMS-managed keys.
resource: https://github.com/aws/amazon-sagemaker-examples/blob/main/advanced_functionality/handling_kms_encrypted_data/handling_kms_encrypted_data.ipynb
Option D is correct if sagemaker does the encryption, if you are dealing with encrypted data then C is 100% correct.

upvoted 3 times

AjoseO 1 year, 11 months ago

Selected Answer: C

C. Assign an IAM role to the Amazon SageMaker notebook with S3 read access to the dataset. Grant permission in the KMS key policy to that role.

To access the encrypted dataset in Amazon S3, the Amazon SageMaker notebook instance must have the appropriate permissions. This can be achieved by assigning an IAM role to the notebook with read access to the dataset in Amazon S3 and granting permission in the KMS key policy to that role. This ensures that the notebook has the necessary permissions to access the encrypted data in Amazon S3, while adhering to best practices for securing sensitive data.

upvoted 2 times

ystotest 2 years, 1 month ago

Selected Answer: C

agreed with C

upvoted 3 times

AmakamaxZanny 2 years, 10 months ago

Answer is C : Open the IAM console. Add a policy to the IAM user that grants the permissions to upload and download from the bucket. You can use a policy that's similar to the following:
https://aws.amazon.com/premiumsupport/knowledge-center/s3-bucket-access-default-encryption/
(number 2)

upvoted 1 times

Deepsachin 3 years, 2 months ago

Seems to be D
https://docs.aws.amazon.com/sagemaker/latest/dg/encryption-at-rest-nbi.html

upvoted 2 times

Madwyn 3 years, 2 months ago

Not D as if you assign the key in the notebook, that's not secure, it will make the encryption ineffective. Instead, you assign the access permission by using IAM.

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 69 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 69
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A large consumer goods manufacturer has the following products on sale:
* 34 different toothpaste variants
* 48 different toothbrush variants
* 43 different mouthwash variants
The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average
(ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched.
Which solution should a Machine Learning Specialist apply?

A. Train a custom ARIMA model to forecast demand for the new product.

B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product.

C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.

D. Train a custom XGBoost model to forecast demand for the new product.

Show Suggested Answer

by ac427 at *March 22, 2020, 11:19 p.m.*

Comments

HaiHN Highly Voted 3 years, 8 months ago

B
https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html
"...When your dataset contains hundreds of related time series, DeepAR outperforms the standard ARIMA and ETS methods. You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on."

upvoted 18 times

ninomfr64 Most Recent 1 year ago

Selected Answer: B

"You can also use the trained model to generate forecasts for new time series that are similar to the ones it has been trained on"
https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

upvoted 1 times

james2033 1 year, 4 months ago

Selected Answer: B

'autoregressive integrated moving average (ARIMA)' <--> DeepAR. https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

B - DeepAR is based on GluonTS, and can use multiple time series for learning

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

Option B

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

DeepAr for new products forever!

upvoted 4 times

AjoseO 2 years, 5 months ago

Selected Answer: B

The DeepAR algorithm is a powerful time series forecasting algorithm that is designed to handle multiple time series data and can handle irregularly spaced time series data and missing values, making it a good fit for this task.

Additionally, the large amount of sales history data available in Amazon S3 makes the use of a deep learning algorithm like DeepAR more appropriate.

upvoted 2 times

Shailendraa 2 years, 10 months ago

B https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

upvoted 1 times

hans1234 3 years, 9 months ago

It is B

upvoted 3 times

ac427 3 years, 9 months ago

This is the same question as Topic 2 Q4

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 68 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 68

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen.

Which combination of algorithms would provide the appropriate insights? (Choose two.)

A. The factorization machines (FM) algorithm

B. The Latent Dirichlet Allocation (LDA) algorithm

C. The principal component analysis (PCA) algorithm

D. The k-means algorithm

E. The Random Cut Forest (RCF) algorithm

Show Suggested Answer

by ac427 at *March 22, 2020, 11:20 p.m.*

Comments

HaiHN Highly Voted 3 years, 8 months ago

C: (OK) Use PCA for reducing number of variables. Each citizen's response should have answer for 500 questions, so it should have 500 variables
D: (OK) Use K-means clustering

A: (Not OK) Factorization Machines Algorithm is usually used for tasks dealing with high dimensional sparse datasets
B: (Not OK) The Latent Dirichlet Allocation (LDA) algorithm should be used for task dealing topic modeling in NLP
E: (Not OK) Random Cut Forest should be used for detecting anormal in data

upvoted 33 times

hans1234 Highly Voted 3 years, 8 months ago

https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/

Answer: C and D

upvoted 12 times

rodrick10 Most Recent 8 months ago

Selected Answer: BD

If the form contains free-text answers, it would be interesting to apply LDA to identify the most frequent/relevant topics in the answers

upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: CD

Option C and D

upvoted 1 times

Mickey321 1 year, 10 months ago

The answer depends on the type of question is it is open ended then would need LDA hence B and D but if the question is a feature then PCA should work

upvoted 1 times

kaike_reis 1 year, 11 months ago

Selected Answer: CD

C and D are the way

upvoted 1 times

ADVIT 2 years ago

CD,
C - for reduce number of columns.
D - for data clustering

upvoted 1 times

AjoseO 2 years, 5 months ago

Selected Answer: CD

C. The principal component analysis (PCA) algorithm
D. The k-means algorithm

PCA is a dimensionality reduction technique that can be used to identify the underlying structure of the census data. This algorithm can help to identify the most important questions and provide an overview of the relationship between the questions and the responses.

K-means is an unsupervised learning algorithm that can be used to segment the population into different groups based on their responses to the census questions. This algorithm can help to determine the healthcare and social program needs by province and city based on the responses collected from each citizen.

These algorithms can help to provide insights into the patterns and relationships within the census data, which can inform decision making for healthcare and social program planning.

upvoted 5 times

Peeking 2 years, 7 months ago

Selected Answer: CD

Reduce dimensionality and cluster subjects.

upvoted 2 times

ac427 3 years, 9 months ago

This is the same question as Topic 2 Q3

upvoted 1 times

muralee_xo 2 years, 5 months ago

how to reach Topic 2 every questions here seem to belong to topic 1

upvoted 1 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 67 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 67
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:
* Start the workflow as soon as data is uploaded to Amazon S3.
* When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon
S3.
* Store the results of joining datasets in Amazon S3.
* If one of the jobs fails, send a notification to the Administrator.
Which configuration will meet these requirements?

A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.

Show Suggested Answer

by Achievement at *July 18, 2020, 11:12 a.m.*

Comments

HaiHN Highly Voted 2 years, 9 months ago

A: Correct. S3 events can trigger AWS Lambda function.
B: Wrong. There's nothing to do with SageMaker in the provided context.
C: Wrong. AWS Batch cannot receive events from S3 directly.
D: Wrong. Will not meed the requirement: "When all the datasets are available in Amazon S3..."

https://docs.aws.amazon.com/step-functions/latest/dg/tutorial-cloudwatch-events-s3.html

upvoted 35 times

scuzzy2010 2 years, 9 months ago

I agree. Step Functions can be used to implement a workflow. In this case, wait for all the datasets to be loaded before triggering the glue job.

upvoted 3 times

cloud_trail 2 years, 9 months ago

Actually, I think that D does meet the requirement of waiting until all datasets are in S3, BUT you do need Glue to join the datasets. Answer is still A.

upvoted 4 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: A

Option A

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: A

Batch isn't event driven, answer is A.

upvoted 1 times

matteocal 1 year, 11 months ago

If EMR were present I would have chose that because of the size of dataset, else is Glue

upvoted 1 times

ZSun 1 year, 2 months ago

exactly, this is also where I got confused. Since Glue is not good at handling such large dataset, multiple terabyte-sized datasets + multiple ETL jobs + daily

upvoted 1 times

Huy 2 years, 8 months ago

A. The answer omits stuffs like Lambda functions and Event Bridge. https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/

upvoted 2 times

johnvik 2 years, 8 months ago

https://d1.awsstatic.com/r2018/a/product-page-diagram-aws-step-functions-use-case-aws-glue.bc69d97a332c2dd29abb724dd747fd82ae110352.png

upvoted 2 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 66 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 66

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only.

How should the Machine Learning Specialist transform the dataset to minimize query runtime?

A. Convert the records to Apache Parquet format.

B. Convert the records to JSON format.

C. Convert the records to GZIP CSV format.

D. Convert the records to XML format.

Show Suggested Answer

by Erso at *April 30, 2020, 2:03 p.m.*

Comments

Erso Highly Voted 2 years, 9 months ago

Answer A seems correct...

upvoted 12 times

Erso 2 years, 8 months ago

sorry, the link https://aws.amazon.com/blogs/big-data/prepare-data-for-model-training-and-invoke-machine-learning-models-with-amazon-athena/

upvoted 1 times

sonalev419 Highly Voted 2 years, 8 months ago

A (Most queries will span 5 to 10 columns only)

upvoted 5 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: A

Option A

upvoted 1 times

exam_prep 2 years, 1 month ago

clue is: most queries will span 5 to 10 column while there are 200 columns. Indicating Data Warehouse means columner storage. Option A is correct.

upvoted 2 times

edvardo 2 years, 7 months ago

Selected Answer: A

A. See https://aws.amazon.com/blogs/big-data/analyzing-data-in-s3-using-amazon-athena/

upvoted 4 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 65 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 65
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

Machine Learning Specialist is building a model to predict future employment rates based on a wide range of economic factors. While exploring the data, the
Specialist notices that the magnitude of the input features vary greatly. The Specialist does not want variables with a larger magnitude to dominate the model.
What should the Specialist do to prepare the data for model training?

A. Apply quantile binning to group the data into categorical bins to keep any relationships in the data by replacing the magnitude with distribution.

B. Apply the Cartesian product transformation to create new combinations of fields that are independent of the magnitude.

C. Apply normalization to ensure each field will have a mean of 0 and a variance of 1 to remove any significant magnitude.

D. Apply the orthogonal sparse bigram (OSB) transformation to apply a fixed-size sliding window to generate new features of a similar magnitude.

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 5:39 a.m.*

Comments

rsimham Highly Voted 2 years, 9 months ago
Ans: C; Normalization is correct
upvoted 34 times

gcpwhiz 2 years, 8 months ago
Ans is not C. What is listed there is the definition of STANDARDIZATION. Normalization just scales and is not useful for reducing the effect of outliers
upvoted 4 times

gcpwhiz 2 years, 8 months ago
nevermind ignore this
upvoted 4 times

Phong Highly Voted 2 years, 8 months ago
Guys, I passed the exam today. It is a tough one but there are many questions here. Good luck everyone! Thank examtopics

upvoted 14 times

haison8x 2 years, 8 months ago

Hi Phong!
Please add my skype: haison8x

upvoted 2 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: C

Ans: C; Normalization is correct

upvoted 2 times

kaike_reis 11 months, 2 weeks ago

C (Yep, STANDARDIZATION is the correct name)
That's an odd question for me

upvoted 1 times

OssamaAbdelatif 1 year, 7 months ago

Selected Answer: C

ans C is correct.

upvoted 1 times

Deepsachin 2 years, 8 months ago

ANS should be C as Normalization work best in case of amplitude diff

upvoted 1 times

grandgale 2 years, 9 months ago

Hi, guys,
First thanks this website for the information it provided.
However, the ML exam has updated most of the questions. only 20+ questions here are included in today's test. Anyway, it is still helpful.
GOOD LUCK EVERYONE!

upvoted 10 times

joker34 2 years, 8 months ago

So there are 40+ other questions on the exam that aren't included in Examtopics?

upvoted 2 times

nez15 2 years, 9 months ago

QUESTION 69
A large consumer goods manufacturer has the following products on sale:
• 34 different toothpaste variants
• 48 different toothbrush variants
• 43 different mouthwash variants
The entire sales history of all these products is available in Amazon S3. Currently, the company is using custom-built autoregressive integrated moving average (ARIMA) models to forecast demand for these products. The company wants to predict the demand for a new product that will soon be launched.
Which solution should a Machine Learning Specialist apply?
A. Train a custom ARIMA model to forecast demand for the new product.
B. Train an Amazon SageMaker DeepAR algorithm to forecast demand for the new product.
C. Train an Amazon SageMaker k-means clustering algorithm to forecast demand for the new product.
D. Train a custom XGBoost model to forecast demand for the new product.
Correct Answer: B

upvoted 4 times

VB 2 years, 8 months ago

https://aws.amazon.com/blogs/machine-learning/forecasting-time-series-with-dynamic-deep-learning-on-aws/

Answer: B

upvoted 1 times

nez15 2 years, 9 months ago

QUESTION 68
An agency collects census information within a country to determine healthcare and social program needs by province and city. The census form collects responses for approximately 500 questions from each citizen.
Which combination of algorithms would provide the appropriate insights? (Select TWO.)
A. The factorization machines (FM) algorithm
B. The Latent Dirichlet Allocation (LDA) algorithm
C. The principal component analysis (PCA) algorithm
D. The k-means algorithm
E. The Random Cut Forest (RCF) algorithm
Correct Answer: CD

upvoted 5 times

VB 2 years, 8 months ago

https://aws.amazon.com/blogs/machine-learning/analyze-us-census-data-for-population-segmentation-using-amazon-sagemaker/

Answer: C and D

upvoted 4 times

cybe001 2 years, 9 months ago

I think the answer is A and B.
The census question and answer will be in text. Use LDA (unsupervised algorithm) which takes the census question/answer and groups them into categories. Use the categorization to group the people and identify similar people.

Use the Factorization Machine to group the people. For each person identify if they answer a question or not. Find the total questions they answered and that will be the Target variable. Now the problem is similar to movie recommendation (consider each question a movie and the total number of questions answered will be the Rating). Based on the questions a Person answered, Factorization Machine groups the people.

Findings from both the algorithms can be used to compare and identify the people for the social programs.

upvoted 2 times

kaike_reis 11 months, 2 weeks ago

it's CD

upvoted 1 times

jasonsunbao 2 years, 9 months ago

FM is mainly used in recommendation system to find hidden variables between two known variables to find correlation between two variables.

upvoted 1 times

nez15 2 years, 9 months ago

QUESTION 67
A. Use AWS Lambda to trigger an AWS Step Functions workflow to wait for dataset uploads to complete in Amazon S3. Use AWS Glue to join the datasets. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
B. Develop the ETL workflow using AWS Lambda to start an Amazon SageMaker notebook instance. Use a lifecycle configuration script to join the datasets and persist the results in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
C. Develop the ETL workflow using AWS Batch to trigger the start of ETL jobs when data is uploaded to Amazon S3. Use AWS Glue to join the datasets in Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
D. Use AWS Lambda to chain other Lambda functions to read and join the datasets in Amazon S3 as soon as the data is uploaded to Amazon S3. Use an Amazon CloudWatch alarm to send an SNS notification to the Administrator in the case of a failure.
Correct Answer: A

upvoted 6 times

nez15 2 years, 9 months ago

QUESTION 67
A Machine Learning Specialist is developing a daily ETL workflow containing multiple ETL jobs. The workflow consists of the following processes:
• Start the workflow as soon as data is uploaded to Amazon S3.
• When all the datasets are available in Amazon S3, start an ETL job to join the uploaded datasets with multiple terabyte-sized datasets already stored in Amazon S3.
• Store the results of joining datasets in Amazon S3.
• If one of the jobs fails, send a notification to the Administrator.
Which configuration will meet these requirements?

upvoted 3 times

nez15 2 years, 9 months ago

QUESTION 66
A Machine Learning Specialist must build out a process to query a dataset on Amazon S3 using Amazon Athena. The dataset contains more than 800,000 records stored as plaintext CSV files. Each record contains 200 columns and is approximately 1.5 MB in size. Most queries will span 5 to 10 columns only.
How should the Machine Learning Specialist transform the dataset to minimize query runtime?
A. Convert the records to Apache Parquet format.
B. Convert the records to JSON format.
C. Convert the records to GZIP CSV format.
D. Convert the records to XML format.
Correct Answer: A

upvoted 11 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 64 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 64

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist is working on an application that performs sentiment analysis. The validation accuracy is poor, and the Data Scientist thinks that the cause may be a rich vocabulary and a low average frequency of words in the dataset.

Which tool should be used to improve the validation accuracy?

A. Amazon Comprehend syntax analysis and entity detection

B. Amazon SageMaker BlazingText cbow mode

C. Natural Language Toolkit (NLTK) stemming and stop word removal

D. Scikit-leam term frequency-inverse document frequency (TF-IDF) vectorizer

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 8:47 a.m.*

Comments

tap123 Highly Voted 3 years, 3 months ago

D is correct. Amazon Comprehend syntax analysis =/= Amazon Comprehend sentiment analysis. You need to read choices very carefully.

upvoted 35 times

mawsman 3 years, 3 months ago

We're looking only to improve the validation accuracy and Comprehend syntax analysis would help that because the word set is rich and the sentiment carying words infrequent. We're not looking to replace the sentiment analysis tool with Comprehend.

upvoted 4 times

DonaldCMLIN Highly Voted 3 years, 3 months ago

AWS COMPREHEND IS A NATURAL LANGUAGE PROCESSING (NLP) SERVICE THAT USES MACHINE LEARNING TO DISCOVER INSIGHTS FROM TEXT.
AMAZON COMPREHEND PROVIDES KEYPHRASE EXTRACTION, SENTIMENT ANALYSIS, ENTITY RECOGNITION, TOPIC MODELING, AND LANGUAGE DETECTION APIS SO YOU CAN EASILY INTEGRATE NATURAL LANGUAGE PROCESSING INTO YOUR APPLICATIONS.

HTTPS://AWS.AMAZON.COM/COMPREHEND/FEATURES/?NC1=H_LS

JUST THROUGH AMAZON COMPREHEND IS MUCH EASY THAN OTHER
THE MUCH MORE CONVENIENT ANSWER IS A.

upvoted 23 times

ComPah 3 years, 3 months ago

Agree Also Keyword is TOOL rest are frameworks

upvoted 2 times

VR10 Most Recent 10 months, 3 weeks ago

Selected Answer: A

Both Amazon Comprehend and the TF-IDF with a classifier solution are valid. If ease of use and pre-trained capabilities are high priorities, Comprehend is a solid option. If customization and dataset-specific nuances are crucial, building a custom model with TF-IDF may be needed. Since Comprehend is a tool, I am going with A.

upvoted 1 times

phdykd 1 year ago

D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer

Here's why:

TF-IDF Vectorizer: This tool from Scikit-learn is effective in handling issues of rich vocabularies and low frequency words. TF-IDF down-weights words that appear frequently across documents (thus might be less informative) and gives more weight to words that appear less frequently but might be more indicative of the sentiment. This approach can enhance the model's ability to focus on more relevant features, potentially improving validation accuracy.

upvoted 4 times

geoan13 1 year, 2 months ago

C I think c is correct. stemming involves reducing words to their root or base form, and stop word removal involves removing common words (e.g., "the," "and," "is") that may not contribute much to sentiment analysis. By using NLTK for stemming and stop word removal, you can simplify the vocabulary and potentially improve the model's ability to capture sentiment from the remaining meaningful words.
A - syntax and entity recognition wont solve the scenario
B - blaze text for words.
D - capturing the importance of words in a document collection. frequency of a word in a document.

upvoted 4 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: D

D is the correct guys

upvoted 1 times

wendaz 1 year, 2 months ago

Amazon Comprehend's syntax analysis and entity detection are more about understanding the structure of sentences and identifying entities within the text rather than tackling the problem of a rich vocabulary with low average frequency of words.

TF-IDF vectorization is a technique that can help reduce the impact of common, low-information words in the dataset while emphasizing the importance of more informative, less frequent words. This could potentially improve the validation accuracy by addressing the identified problem.

upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: A

A. YES - he works on an application and not a model, Amazon Comprehend is the ready-to-use tool he wants; TF-IDF is built-in
B. NO - word2vec will be challenged with low frequency terms; GloVe and FastText are better for that
C. NO - the vocabulary is righ, so stemming and stop word removal will not address the core issue
D. NO - right approach, but that is not "a tool"

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: D

Option D. This approach can help in reducing the impact of words that occur frequently in the dataset and increasing the impact of words that occur less frequently. This can help in improving the accuracy of the model.

upvoted 2 times

ashii007 1 year, 4 months ago

The anwer is B.
Blazing text can hadle OOV words as explained below. https://docs.aws.amazon.com/sagemaker/latest/dg/blazingtext.html

upvoted 2 times

jyrajan69 1 year, 5 months ago

This is an AWS exam, so why would you choose anything other than A or B, and based on the link, it looks like B most likely

upvoted 2 times

kaike_reis 1 year, 5 months ago

Selected Answer: D

The passage “low average frequency of words” points directly to the use of TF-IDF. Letter A deviates from what the question proposes and is discarded. Letter B proposes a radical change in my POV. Letter C does not solve the passage mentioned at the beginning. Letter D is correct.

upvoted 2 times

GOSD 1 year, 8 months ago

The Amazon SageMaker BlazingText algorithm provides highly optimized implementations of the Word2vec and text classification algorithms. The Word2vec algorithm is useful for many downstream natural language processing (NLP) tasks, such as *****sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

upvoted 1 times

vassof95 1 year, 8 months ago

Selected Answer: D

I would say since the buzzword "low average frequency" comes up, the safe choise would be the tfid vectorizer.

I go for D.

upvoted 2 times

ParkXD 1 year, 9 months ago

Selected Answer: D

The Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer is a widely used tool to mitigate the high dimensionality of text data.
Option A, Amazon Comprehend syntax analysis, and entity detection, can help in extracting useful features from the text, but it does not address the issue of high dimensionality.

Option B, Amazon SageMaker BlazingText cbow mode, is a tool for training word embeddings, which can help to represent words in a lower dimensional space. However, it does not directly address the issue of high dimensionality and low frequency of words.

Option C, Natural Language Toolkit (NLTK) stemming and stop word removal, can reduce the dimensionality of the feature space, but it does not address the issue of low-frequency words that are important for sentiment analysis.

upvoted 5 times

cpal012 1 year, 10 months ago

Selected Answer: C

Emphasis is on the rich words - so stemming can help reduce these to more common words. Blazing Text in cbow mode doesnt seem relevant is about providing words given a context. And TF-IDF I'm not sure would do anything except highlight the problem you are already having?

upvoted 1 times

bakarys 1 year, 10 months ago

Selected Answer: D

D. Scikit-learn term frequency-inverse document frequency (TF-IDF) vectorizer would be the best tool to use in this scenario. The TF-IDF vectorizer will give less weight to the less frequent words in the dataset, and allow the more informative and frequent words to have a greater impact on the sentiment analysis. This can help to improve the validation accuracy of the model.

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 63 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 63

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A retail company intends to use machine learning to categorize new products. A labeled dataset of current products was provided to the Data Science team. The dataset includes 1,200 products. The labeled dataset has 15 features for each product such as title dimensions, weight, and price. Each product is labeled as belonging to one of six categories such as books, games, electronics, and movies.
Which model should be used for categorizing new products using the provided dataset for training?

A. AnXGBoost model where the objective parameter is set to multi:softmax

B. A deep convolutional neural network (CNN) with a softmax activation function for the last layer

C. A regression forest where the number of trees is set equal to the number of product categories

D. A DeepAR forecasting model based on a recurrent neural network (RNN)

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 5:34 a.m.*

Comments

rsimham Highly Voted 2 years, 9 months ago

Ans: A XGBoost multi class classification. https://medium.com/@gabrielziegler3/multiclass-multilabel-classification-with-xgboost-66195e4d9f2d

CNN is used for image classificaiton problems

upvoted 34 times

JayK Highly Voted 2 years, 9 months ago

Answer is A. This a classification problem thus XGBoost and the fact that there are six categories SOFTMAX is the right activation function

upvoted 14 times

sonoluminescence Most Recent 8 months, 2 weeks ago

Selected Answer: A

Deep convolutional neural networks (CNNs) are primarily used for image processing tasks. Given that the dataset provided is structured/tabular in nature (with features like dimensions, weight, and price) and does not mention image data, a CNN is not the most appropriate choice.

upvoted 2 times

loict 10 months ago

Selected Answer: A

A. YES - perfect fit, multi:softmax the highest probability class is assigned
B. NO - CNN is for imaging
C. NO - regression forest is for continuuous variables, we can discrete classification
D. NO - it is classification, not forecasting

upvoted 3 times

Mickey321 10 months, 2 weeks ago

Selected Answer: A

Option A XGBoost multi class classification

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: A

A is the answer.

upvoted 1 times

oso0348 1 year, 4 months ago

Selected Answer: A

The XGBoost algorithm is a popular and effective technique for multi-class classification. The objective parameter can be set to multi:softmax, which uses a softmax objective function for multi-class classification. This will train the model to predict the probability of each product belonging to each category, and the most probable category will be chosen as the final prediction.

A deep convolutional neural network (CNN) (B) is a powerful technique commonly used for image recognition tasks. However, it is less appropriate for tabular data like the dataset provided.

upvoted 1 times

Konga98 1 year, 5 months ago

Selected Answer: A

A, CNN is used for image classification. It would be suitable if we were classifying products using pictures of them.

upvoted 1 times

yemauricio 1 year, 6 months ago

Selected Answer: A

https://xgboost.readthedocs.io/en/stable/parameter.html

upvoted 2 times

GiyeonShin 1 year, 6 months ago

Selected Answer: A

B - CNN is used for dataset that have "local intermediate features" ex) images, or textCNN, etc
C - We need classfication model, not regression model
D - RNN is used for dataset that have sequential features
A is correct

upvoted 3 times

Peeking 1 year, 7 months ago

Selected Answer: A

A is the best option here. Only 1200 items and 6 classes are not enough data to involve a deep neural architecture for classification.

upvoted 1 times

Shailendraa 1 year, 10 months ago

Ans- A … For multiclassification - multi: SoftMax

upvoted 2 times

Morsa 1 year, 12 months ago

Selected Answer: A

That is a classification problem so A is the answer

upvoted 1 times

apprehensive_scar 2 years, 5 months ago

Selected Answer: A

Easy one. A is correct

upvoted 2 times

Kevinkoo 2 years, 7 months ago

Selected Answer: A

A is correct

upvoted 3 times

stardustWu 2 years, 8 months ago

Definitely A.

upvoted 1 times

syu31svc 2 years, 8 months ago

100% is A; the the others are clearly wrong

Convolutional Neural Network (ConvNet or CNN) is a special type of Neural Network used effectively for image recognition and classification
Recurrent neural networks (RNN) are a class of neural networks that is powerful for modeling sequence data such as time series or natural language

upvoted 5 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 62 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 62
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Data Scientist wants to gain real-time insights into a data stream of GZIP files.
Which solution would allow the use of SQL to query the stream with the LEAST latency?

A. Amazon Kinesis Data Analytics with an AWS Lambda function to transform the data.

B. AWS Glue with a custom ETL script to transform the data.

C. An Amazon Kinesis Client Library to transform the data and save it to an Amazon ES cluster.

D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket.

Show Suggested Answer

by JayK at *Jan. 4, 2020, 4:03 p.m.*

Comments

cybe001 Highly Voted 3 years, 3 months ago
A is correct. Kinesis Data Analytics can use lamda to convert GZIP and can run SQL on the converted data.
https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/
upvoted 44 times

VB Highly Voted 3 years, 2 months ago
A is correct:

https://aws.amazon.com/about-aws/whats-new/2017/10/amazon-kinesis-analytics-can-now-pre-process-data-prior-to-running-sql-queries/

"To get started, simply select an AWS Lambda function from the Kinesis Analytics application source page in the AWS Management console. Your Kinesis Analytics application will automatically process your raw data records using the Lambda function, and send transformed data to your SQL code for further processing.
Kinesis Analytics provides Lambda blueprints for common use cases like converting GZIP
..."
upvoted 17 times

ef12052 Most Recent 3 months, 1 week ago
Selected Answer: A
Use Amazon Kinesis Data Analytics if you need SQL-based processing and advanced analytics capabilities for streaming data.
Use Amazon Kinesis Data Firehose if your primary requirement is to deliver, transform, and load streaming data into various AWS destinations with simplified configurations, but not for SQL-based processing.

upvoted 1 times

Denise123 10 months, 3 weeks ago

Selected Answer: D

If gaining real-time insights involves complex analytics or custom processing, Amazon Kinesis Data Analytics with AWS Lambda is likely a more suitable choice. If the requirements can be met with simpler data transformations, Amazon Kinesis Data Firehose might provide a more straightforward and potentially lower-latency solution.

In other words, if this data is in GZIP files and the processing requirements are relatively simple, Amazon Kinesis Data Firehose might be a more straightforward and efficient choice. GZIP files typically contain compressed data, and if our primary objective is to ingest, transform, and load this data into other AWS services for real-time insights, Kinesis Data Firehose provides a managed and streamlined solution that can handle GZIP compression.

upvoted 1 times

Denise123 10 months, 3 weeks ago

The answer can be A , please comment if you have more clarity. After searching more, I also found out the following:
(I have missed the SQL requirement in the question)
Use Amazon Kinesis Data Analytics if you need SQL-based processing and advanced analytics capabilities for streaming data.
Use Amazon Kinesis Data Firehose if your primary requirement is to deliver, transform, and load streaming data into various AWS destinations with simplified configurations, but not for SQL-based processing.

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: A

A is correct, why D xiyarsan sen?

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

A is correct

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: A

"allow the use ohttps://www.examtopics.com/exams/amazon/aws-certified-machine-learning-specialty/view/13/#f SQL to query the stream with the LEAST latency?"
Well, the only solution that presents SQL query is (A). It's a description of KDA.

upvoted 2 times

Nadia0012 1 year, 10 months ago

Selected Answer: A

the term "lease latency" is the the hidden point. with Glue we can have near real-time but Kinesis data analytics will give you real-time transformation with internal lambda

upvoted 3 times

Valcilio 1 year, 10 months ago

Selected Answer: A

A is correct, with KDA you can run sql queries in the data during the streaming (real-time SQL queries).

upvoted 2 times

bakarys 1 year, 10 months ago

Selected Answer: D

D. Amazon Kinesis Data Firehose to transform the data and put it into an Amazon S3 bucket would be the best solution for allowing the use of SQL to query the stream with the least latency. Amazon Kinesis Data Firehose can be configured to transform the data before writing it to Amazon S3 in real-time. Once the data is in S3, it can be queried using SQL with Amazon Athena, which is a serverless query service that allows running standard SQL queries against data stored in Amazon S3. This approach provides the lowest latency compared to other options and requires minimal setup and maintenance.

upvoted 3 times

akgarg00 1 year, 1 month ago

Query has to be run on stream so firehose not possible.

upvoted 1 times

OssamaAbdelatif 2 years, 1 month ago

Selected Answer: A

A is correct.

upvoted 1 times

AddiWei 2 years, 11 months ago

And somehow "transformation" is added to the answer as a requirement when it clearly was not part of the requirement from the question.

upvoted 2 times

apprehensive_scar 2 years, 11 months ago

AAAAAAA

upvoted 1 times

HalloSpencer 3 years, 2 months ago

what about "LEAST latency"?

upvoted 4 times

Erso 3 years, 2 months ago

A is correct. you can pre-process data prior to running SQL queries with Kinesis Data Analytics and Lambda (more or less) is always a best practice :)

upvoted 3 times

JayK 3 years, 3 months ago

Answer is B. Kinesis Data Analytics does not do any transformation, it is only for querying. Glue ETL can have scripts that can transform the data

upvoted 2 times

SophieSu 3 years, 2 months ago

so you need lambda

upvoted 1 times

am7 3 years, 3 months ago

But we need to run SQL on real time stream data.

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 61 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 61
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is working with a large cybersecurity company that manages security events in real time for companies around the world. The cybersecurity company wants to design a solution that will allow it to use machine learning to score malicious events as anomalies on the data as it is being ingested. The company also wants be able to save the results in its data lake for later processing and analysis.
What is the MOST efficient way to accomplish these tasks?

A. Ingest the data using Amazon Kinesis Data Firehose, and use Amazon Kinesis Data Analytics Random Cut Forest (RCF) for anomaly detection. Then use Kinesis Data Firehose to stream the results to Amazon S3.

B. Ingest the data into Apache Spark Streaming using Amazon EMR, and use Spark MLlib with k-means to perform anomaly detection. Then store the results in an Apache Hadoop Distributed File System (HDFS) using Amazon EMR with a replication factor of three as the data lake.

C. Ingest the data and store it in Amazon S3. Use AWS Batch along with the AWS Deep Learning AMIs to train a k-means model using TensorFlow on the data in Amazon S3.

D. Ingest the data and store it in Amazon S3. Have an AWS Glue job that is triggered on demand transform the new data. Then use the built-in Random Cut Forest (RCF) model within Amazon SageMaker to detect anomalies in the data.

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 7:49 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago
I WOULD LIKE TO CHOOSE ANSWER A.

https://aws.amazon.com/tw/blogs/machine-learning/use-the-built-in-amazon-sagemaker-random-cut-forest-algorithm-for-anomaly-detection/
upvoted 60 times

hamimelon 2 years ago
Donald, do you know your CAPS LOCK has been on the whole time?
upvoted 15 times

Nadia0012 1 year, 10 months ago
I know why his caps lock has been on :D to enter the "I am not robot" code easier :D
upvoted 4 times

ccpmad 1 year, 5 months ago

yes, but it works with minus also...

upvoted 1 times

JayK Highly Voted 3 years, 3 months ago

Answer is A. As the word anamoly talks about Random Cut Forest in the exam and that can be done in a cost effective manner using Kinesis Data Analytics

upvoted 15 times

Shakespeare 7 months ago

I think it would have been more accurate if the options were kinetic data stream -> kinesis data analytics -> kinesis firehose -> S3

upvoted 2 times

saclim Most Recent 9 months ago

The question says REAL TIME events doesn't that eliminate Data Firehose as it is technically NEAR real time but not real time like Data Stream? Though Random Cut Forest seems like the best option for anomaly detection. I'm torn between A and B

upvoted 1 times

vkbajoria 9 months, 2 weeks ago

Selected Answer: A

Kinesis Firehose and Data Analytics with random cut forest should do it.

upvoted 1 times

phdykd 1 year ago

A.
Based on these considerations, Option A is the most efficient way to accomplish the tasks. It provides a seamless, real-time data ingestion and processing pipeline, leverages machine learning for anomaly detection, and efficiently stores data in a data lake, meeting all the key requirements of the cybersecurity company.

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 2 months ago

Selected Answer: A

ONLY A

upvoted 1 times

sonoluminescence 1 year, 2 months ago

Selected Answer: A

B not as efficient for real-time processing and storing results as using Kinesis services.

upvoted 2 times

DimLam 1 year, 2 months ago

Selected Answer: B

At least B is a possible solution, but A will not work as KDF doesn't support KDA as a destination service https://docs.aws.amazon.com/firehose/latest/dev/create-name.html . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

upvoted 1 times

Dun6 1 year, 1 month ago

KDF does support KDA as destination

upvoted 1 times

AmeeraM 1 year, 3 months ago

Selected Answer: A

A has all the required steps

upvoted 1 times

loict 1 year, 4 months ago

Selected Answer: A

A. YES - Firehose can pipe into KDA, and KDA supports RCF
B. NO - RCF best for anomality detection
C. NO - no need for intermediary S3 storage
D. NO - no need for intermediary S3 storage

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

option A

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: A

A is the correct. One tip for the exam: When you see Data Streaming, possibly the solution should contains a Kinesis Service. B is too much

complex!

upvoted 3 times

nilmans 1 year, 6 months ago

Selected Answer: A

Makes sense to select A here.

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: A

I strongly believe A is the right answer. At a minimum there should be some justification provided for your answer.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: A

Amazon Kinesis Data Firehose is a fully managed service for streaming real-time data to Amazon S3 and can handle the ingestion of large amounts of data in real time. Kinesis Data Analytics Random Cut Forest (RCF) is a fully managed service that can be used to perform anomaly detection on streaming data, making it well suited for this use case. The results of the anomaly detection can then be streamed to Amazon S3 using Kinesis Data Firehose, providing a scalable and cost-effective data lake for later processing and analysis.

upvoted 2 times

DimLam 1 year, 2 months ago

The problem with A, is that there is that KDF doesn't support KDA as a destination service https://docs.aws.amazon.com/firehose/latest/dev/create-name.html . In my opinion, KDF should always be the latest Kinesis Service in a streaming pipeline

upvoted 1 times

OssamaAbdelatif 2 years, 1 month ago

I would select A

upvoted 1 times

ovokpus 2 years, 6 months ago

Selected Answer: A

B is too resource intensive for that use case. I choose A, but I think the data should be better ingested using Kinesis streams

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 60 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 60

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is assigned a TensorFlow project using Amazon SageMaker for training, and needs to continue working for an extended period with no Wi-Fi access.

Which approach should the Specialist use to continue working?

A. Install Python 3 and boto3 on their laptop and continue the code development using that environment.

B. Download the TensorFlow Docker container used in Amazon SageMaker from GitHub to their local environment, and use the Amazon SageMaker Python SDK to test the code.

C. Download TensorFlow from tensorflow.org to emulate the TensorFlow kernel in the SageMaker environment.

D. Download the SageMaker notebook to their local environment, then install Jupyter Notebooks on their laptop and continue the development in a local notebook.

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 7:32 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

ANSWER B.

YOU COULD INSTALL DOCKER-COMPOSE (AND NVIDIA-DOCKER IF TRAINING WITH A GPU) FOR LOCAL TRAINING

HTTPS://SAGEMAKER.READTHEDOCS.IO/EN/STABLE/OVERVIEW.HTML#LOCAL-MODE
HTTPS://GITHUB.COM/AWSLABS/AMAZON-SAGEMAKER-EXAMPLES/BLOB/MASTER/SAGEMAKER-PYTHON-SDK/TENSORFLOW_DISTRIBUTED_MNIST/TENSORFLOW_LOCAL_MODE_MNIST.IPYNB

upvoted 42 times

sqavi 1 year, 11 months ago

None of these links are working

upvoted 3 times

VB Highly Voted 3 years, 2 months ago

https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/

B

upvoted 10 times

ef12052 Most Recent 3 months, 1 week ago

Selected Answer: B

https://aws.amazon.com/blogs/machine-learning/use-the-amazon-sagemaker-local-mode-to-train-on-your-notebook-instance/

stop using gpt....

upvoted 1 times

sfwewv 5 months ago

Selected Answer: D

GPT said SageMaker Python SDK is less suitable for offline

upvoted 1 times

rookiee1111 8 months, 2 weeks ago

Selected Answer: B

Correction it will be B, while D is possible, it cannot exactly mimic the sagemaker env, with docker all the configuration and libs will be available to the user which would be an ideal working setup for the DS to work with.

upvoted 1 times

rookiee1111 8 months, 2 weeks ago

Selected Answer: D

You can easily download the notebook instance, and work locally using jupyter notebook configured on your laptop which is one the advantages of using sagemaker, and that is what Amazon also promotes imo.

upvoted 2 times

ArchMelody 10 months, 3 weeks ago

Selected Answer: D

Both Amazon Q (AWS Expert) and ChatGPT insist on D. Plus all the links that I see here about Docker/Git and stuff, they either not working or deprecated so far. Not to mention their complexity to my eyes.
Thus, I will go for D.

upvoted 3 times

rav009 1 year, 3 months ago

Selected Answer: B

the local mode of sagemaker SDK:
https://sagemaker.readthedocs.io/en/stable/overview.html#local-mode
B

upvoted 2 times

Mickey321 1 year, 4 months ago

Selected Answer: B

Option B

upvoted 1 times

ADVIT 1 year, 6 months ago

B,
https://github.com/aws/sagemaker-tensorflow-serving-container

upvoted 1 times

Valcilio 1 year, 10 months ago

Selected Answer: B

It's B

upvoted 1 times

SriAkula 2 years, 10 months ago

Answer : D

upvoted 2 times

noblate 3 years ago

why not D?

upvoted 1 times

AddiWei 2 years, 10 months ago

My assumption is that D there is no way to test the code. You need the Sagemaker SDK in order to utilize dockerized container of Tensorflow from Sagemaker is my best guess.

upvoted 2 times

Tomatoteacher 1 year, 12 months ago

Cannot be D. If you used Jupyter notebook, you are unable to use it without internet access.

upvoted 1 times

rookiee1111 8 months, 2 weeks ago

That is incorrect, once jupyter notebook is configured you can use it offline.

upvoted 1 times

CMMC 3 years, 2 months ago

Agreed for B

upvoted 4 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 59 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 59

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A data scientist has explored and sanitized a dataset in preparation for the modeling phase of a supervised learning task. The statistical dispersion can vary widely between features, sometimes by several orders of magnitude. Before moving on to the modeling phase, the data scientist wants to ensure that the prediction performance on the production data is as accurate as possible.
Which sequence of steps should the data scientist take to meet these requirements?

A. Apply random sampling to the dataset. Then split the dataset into training, validation, and test sets.

B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.

C. Rescale the dataset. Then split the dataset into training, validation, and test sets.

D. Split the dataset into training, validation, and test sets. Then rescale the training set, the validation set, and the test set independently.

Show Suggested Answer

by cron0001 at *April 24, 2022, 1:44 a.m.*

Comments

cron0001 Highly Voted 3 years, 2 months ago

Selected Answer: C

C would be my answer here. Rescaling each set independently could lead to strange skews. Training set, Test set and Evaluation set should be on the same scale

upvoted 17 times

GiyeonShin 2 years, 6 months ago

You're right. test set and val set should be rescaled on the same scale.
But the scale value should be extracted by only statistical value from training data.
I think C means that the rescaling stage is affected by the values from the whole data (with val, test set)
So, I think B is correct

upvoted 9 times

masoa3b Highly Voted 2 years, 8 months ago

Selected Answer: B

https://stackoverflow.com/questions/49444262/normalize-data-before-or-after-split-of-training-and-testing-data

C also leads to data leakage. You are using the test data to scale everything. So part of the data in the test set is used to scale for when you build the model on the training and check against the validation set.

upvoted 17 times

ML_2 Most Recent 11 months ago

Selected Answer: B

If you Rescale all the data first you are going to do data leakage by showing all the variance of data with in training. The rescaling needs to be after splitting the data and not before it

upvoted 1 times

Denise123 1 year, 4 months ago

Selected Answer: B

The best practice is --> to split the dataset into training, validation, and test sets first, and then rescale the training set and apply the SAME scaling to the validation and test sets. This ensures that the scaling parameters (e.g., mean and standard deviation for standardization or min and max values for min-max scaling) are calculated only based on the training set to prevent data leakage and maintain the integrity of the evaluation process.

By following this approach, you prevent information from the validation and test sets from influencing the scaling parameters, which could lead to data leakage and overestimation of model performance. Keeping the scaling consistent across all subsets ensures a fair evaluation of the model's generalization performance on new, unseen data.

upvoted 5 times

phdykd 1 year, 6 months ago

Answer is B.
The other options have shortcomings:

A: Random sampling is a good practice, but it doesn't address the issue of feature scaling. Also, rescaling should occur after splitting the data.
C: Rescaling the entire dataset before splitting could lead to data leakage, where information from the validation/test sets inadvertently influences the training process.
D: Rescaling the sets independently would lead to inconsistencies in scale across the training, validation, and test sets, which could negatively impact model performance and evaluation.

upvoted 2 times

Sukhi4fornet 1 year, 6 months ago

OPTION C. Rescale the dataset. Then split the dataset into training, validation, and test sets.

Explanation:

Rescaling the dataset:

This is the first step to address the varying statistical dispersion among features. By rescaling, you ensure that all features are on a similar scale, which is important for many machine learning algorithms.
Splitting into training, validation, and test sets:

After rescaling, the dataset is split into training, validation, and test sets. This ensures that the model is trained on one set, validated on another set, and tested on a third set. This separation helps evaluate the model's performance on unseen data.
Option C ensures that the rescaling is applied before splitting the data, ensuring consistency in the scaling across different sets. This approach prevents data leakage and provides a more accurate representation of how the model will perform on new, unseen data.

upvoted 1 times

akgarg00 1 year, 7 months ago

Selected Answer: B

Validation and test set should be scaled as per parameters used for scaling of training set. Independent scaling of test set would mean that drift of model in production will be way quicker and is not recommended in data science

upvoted 1 times

elvin_ml_qayiran25091992razor 1 year, 8 months ago

Selected Answer: B

B is correct, scale on train and apply the others. prevent to data leakage

upvoted 1 times

akgarg00 1 year, 8 months ago

Selected Answer: B

Answer B, C is not a good data science practise.

upvoted 1 times

DimLam 1 year, 8 months ago

Selected Answer: B

We need firstly split the data to avoid data leakage from test/eval sets, then rescale data in all sets using statistics from training set

upvoted 1 times

DavidRou 1 year, 10 months ago

Selected Answer: B

I think the right answer here is B. We need to split the dataset into Training, Validation and Test set. Then we can only scale (by using some

technique) data contained in the Training set. Data that belong to Validation and Test set must be scaled by using the parameters used on the training.

For example, if we want to apply a standardization, we can do that only on the Training set as we should not be allowed to use mean and standard deviation computed on Validation/Test set. We must act as we don't own those data!

upvoted 2 times

Mickey321 1 year, 10 months ago

Selected Answer: B

option B

upvoted 1 times

kaike_reis 1 year, 11 months ago

Data Science 101:
(A) Given the question, doesn't solve the magnitude problem.
(B) Correct
(C) Data Leakage
(D) It's not correct, still data leakage.

upvoted 1 times

gusta_dantas 1 year, 11 months ago

Tricky question, but, D, definitely!

B: You can't apply the same scaling to the validation and test sets 'cause you may suffer data leakage!

C: You shouldn't rescale the whole dataset then split into training, validation and test, it's not a good practice and may suffer data leakage as well.

D: You're first splitting the whole dataset and applying rescaling individually, preventing any data leakage and each set is rescaled based in your own statistics.

upvoted 1 times

DavidRou 1 year, 10 months ago

Theoretically, you should not have Test set data at Training time (when you're doing the scaling), so how do you think to do that?
What if you will not have an entire Test set, but you will receive each new row at a time?

upvoted 1 times

kaike_reis 1 year, 11 months ago

but you are leaking information from validation samples between themselves.

upvoted 1 times

JK1977 2 years, 1 month ago

Selected Answer: B

From Bing chat (and it makes complete sense)
"Based on the search results, I think the best sequence of steps for the data scientist to take is B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.

This sequence of steps ensures that the data scientist can evaluate the model performance on different subsets of data that have not been used for training or tuning. It also ensures that the data scientist can rescale the features to have a common scale without introducing any data leakage from the validation or test sets. Rescaling the features can help improve the accuracy of some machine learning algorithms that are sensitive to the magnitude or distribution of the data, such as distance-based methods or gradient-based methods 1.

upvoted 3 times

tommct 2 years, 1 month ago

Selected Answer: B

You want to measure how the model performs on new data. Scaling with the test set is a no-no.

upvoted 1 times

GOSD 2 years, 2 months ago

B or D, I dont understand the semantics of "independently" and the effect it would have. It's most def not done before because of data leakage.
https://www.linkedin.com/pulse/feature-scaling-dataset-spliting-arnab-mukherjee/

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 58 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 58

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist needs to be able to ingest streaming data and store it in Apache Parquet files for exploration and analysis.

Which of the following services would both ingest and store this data in the correct format?

A. AWS DMS

B. Amazon Kinesis Data Streams

C. Amazon Kinesis Data Firehose

D. Amazon Kinesis Data Analytics

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 3:30 a.m.*

Comments

JayK Highly Voted 3 years, 9 months ago

the answer is C. as the main point of the question is data transformation to Parquet format which is done by Kinesis Data Firehose not Data Stream. Coming to the data store the data store in Kinesis Data Stream is only for couple of days so it does not serve the purpose here

upvoted 52 times

shammous 11 months ago

The storage part will be taken care of by S3 anyway. Firehose would just transform to Parquet on the fly.

upvoted 1 times

eganilovic Highly Voted 3 years, 8 months ago

Firehose

upvoted 5 times

earthMover Most Recent 2 years, 1 month ago

Not sure Firehose can store the data .... Data Stream can store the data. Someone please explain the answer

upvoted 1 times

kaike_reis 1 year, 11 months ago

Firehose is to Store the data. Stream requires other service to do that.

upvoted 1 times

GOSD 2 years, 2 months ago

Kinesis Data Streams can Store for up to 365 days, While Firehouse sends it to S3. Which is correct?

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: C

Firehose can do it if the data is in JSON or ORC format initially!

upvoted 2 times

DS2021 2 years, 4 months ago

It should be KDS

upvoted 1 times

AjoseO 2 years, 5 months ago

Selected Answer: C

Amazon Kinesis Data Firehose is a fully managed service that can automatically load streaming data into data stores and analytics tools.

It can ingest real-time streaming data such as application logs, website clickstreams, and IoT telemetry data, and then store it in the correct format, such as Apache Parquet files, for exploration and analysis.

This makes it a suitable option for the requirement described in the question.

upvoted 1 times

Thai_Xuan 3 years, 8 months ago

B
https://github.com/ravsau/aws-exam-prep/issues/10

upvoted 2 times

weslleylc 3 years, 8 months ago

B) Only Amazon Kinesis Data Streams can store and Ingest data. We don't need to apply any transformation; the question asks to ingest and store data in Apache Parquet format, There is no assumption that the data coming in a different format than parquet.

upvoted 3 times

joe3232 2 years, 5 months ago

KDS cant store to s3
https://stackoverflow.com/questions/66097886/writing-to-s3-via-kinesis-stream-or-firehose

upvoted 1 times

In 3 years, 8 months ago

It is C with no doubt
https://aws.amazon.com/about-aws/whats-new/2018/05/stream_real_time_data_in_apache_parquet_or_orc_format_using_firehose/

upvoted 5 times

GeeBeeEl 3 years, 9 months ago

It appears all agree that the answer is between Firehose and Analytics. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html shows firehose can store parquet to S3

upvoted 3 times

GeeBeeEl 3 years, 8 months ago

It appears all agree that the answer is between Firehose and Analytics. Data Streams handle stuff like event data, clickstream etc. Its not interested in special format, the focus is speed. The question did not talk of transformation, only ingestion. Kinesis Firehose is used for ingestion. Both firehose and analytics can store, only firehose can ingest. https://docs.aws.amazon.com/firehose/latest/dev/record-format-conversion.html shows firehose can store parquet to S3

upvoted 1 times

Urban_Life 3 years, 9 months ago

Think just like this -- batch process Glue ETL and Streaming process Firehose ETL .....covert to parquet or any other format.

upvoted 1 times

CMMC 3 years, 9 months ago

C for Firehose

upvoted 2 times

Erso 3 years, 9 months ago

Just in case https://acloud.guru/forums/aws-certified-big-data-specialty/discussion/-KhI3MgPEo-FY5rfgl3J/what_is_difference_between_kin

upvoted 2 times

BigEv 3 years, 9 months ago

Amazon Kinesis Data Firehose can convert the format of your input data from JSON to Apache Parquet or Apache ORC before storing the data in Amazon S3.
https://github.com/awsdocs/amazon-kinesis-data-firehose-developer-guide/blob/master/doc_source/record-format-conversion.md

upvoted 3 times

rsimham 3 years, 9 months ago

I would go with B. Kinesis data streams stores data, while Firehose not.

upvoted 3 times

cloud_trail 3 years, 8 months ago

It's the other way around. Firehouses stores data; data streams does not.

upvoted 1 times

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 57 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 57
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is required to build a supervised image-recognition model to identify a cat. The ML Specialist performs some tests and records the following results for a neural network-based image classifier:
Total number of images available = 1,000
Test set images = 100 (constant test set)
The ML Specialist notices that, in over 75% of the misclassified images, the cats were held upside down by their owners.
Which techniques can be used by the ML Specialist to improve this specific test error?

A. Increase the training data by adding variation in rotation for training images.

B. Increase the number of epochs for model training

C. Increase the number of layers for the neural network.

D. Increase the dropout rate for the second-to-last layer.

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 4:38 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

NO CORRECT TRAINING DATA, MORE WORKS JUST WASTE TIME.

ONE OF THE REASONS FOR POOR ACCURACY COULD BE INSUFFICIENT DATA. THIS CAN BE OVERCOME BY IMAGE AUGMENTATION. IMAGE AUGMENTATION IS A TECHNIQUE OF INCREASING THE DATASET SIZE BY PROCESSING (MIRRORING, FLIPPING, ROTATING, INCREASING/DECREASING BRIGHTNESS, CONTRAST, COLOR) THE IMAGES.

HTTPS://MEDIUM.COM/DATADRIVENINVESTOR/AUTO-MODEL-TUNING-FOR-KERAS-ON-AMAZON-SAGEMAKER-PLANT-SEEDLING-DATASET-7B591334501E

ANSWER A. ADD MORE TRAINING DATA FOR ROTATION IMAGES COULD BE A WAY TO DEAL WITH ISSUE

upvoted 63 times

Jeremy1 2 years, 1 month ago

Donald, your caps lock is on.

upvoted 13 times

kaike_reis 1 year, 5 months ago

Okay, was funny

upvoted 1 times

Nadia0012 1 year, 10 months ago

LOL :D

upvoted 1 times

ccpmad 1 year, 5 months ago

is it possible no using MAYUS? it is annoying

upvoted 1 times

tap123 3 years, 3 months ago

The key phrase might be "constant test set", so you can't increase training set by shrinking the size of test set. Thus the only feasible choice is to increase training time by increasing the number of epochs => answer B.

upvoted 2 times

mawsman 3 years, 2 months ago

The problem is images are upside down and misclassified. If right side up then the model would classify correctly. This can only be fixed ba rotating not by trying to recognise upside down cat more times.

upvoted 3 times

Urban_Life 3 years, 2 months ago

What's your answer B?

upvoted 1 times

VB 3 years, 3 months ago

A . Increase the training data by adding variation in rotation for training images.

It never says to move the images from Test data set (because it is constant)... only variations are added to the images..so, A is correct.

upvoted 1 times

rsimham 3 years, 3 months ago

agree with A

upvoted 9 times

phdykd Most Recent 1 year ago

A is answer

upvoted 1 times

Kensev 1 year ago

Selected Answer: A

Data Augmentation would fix the missing conditional data

upvoted 2 times

cgsoft 1 year, 1 month ago

Selected Answer: A

ChatGPT says the answer is A. Trust a model to answer an ML question correctly! ;)

upvoted 1 times

AmeeraM 1 year, 3 months ago

Selected Answer: A

how come more epochs it better than augmentation?

upvoted 1 times

Mickey321 1 year, 4 months ago

Selected Answer: A

option A

upvoted 1 times

kaike_reis 1 year, 5 months ago

Selected Answer: A

The question is clear and the answer is clear as well

upvoted 1 times

nilmans 1 year, 6 months ago

Selected Answer: A

should be A

upvoted 1 times

earthMover 1 year, 7 months ago

Selected Answer: A

More epochs is not a good approach to fundamental data issues

upvoted 2 times

oso0348 1 year, 10 months ago

Selected Answer: A

the Specialist can apply data augmentation techniques to increase the training data by adding variation in rotation for training images. This technique will allow the model to learn to recognize cats in various orientations, including upside down.

upvoted 1 times

AjoseO 1 year, 11 months ago

Selected Answer: A

Adding more variation in rotation to the training data can help the model to learn how to classify cats in different orientations, including when they are held upside down. This can improve the model's ability to identify cats in this position and reduce the misclassification rate for images in which the cats are upside down.

By adding more rotation to the training data, the model can be trained to generalize better to new images, including those with cats in different orientations. This can help to reduce overfitting and improve the model's overall performance.

upvoted 1 times

Tomatoteacher 1 year, 12 months ago

Selected Answer: A

Only logical answer 100% A.

upvoted 1 times

Jeremy1 2 years, 1 month ago

Selected Answer: A

More data is a good answer. A

upvoted 1 times

ryuhei 2 years, 3 months ago

Selected Answer: A

Answer is ”A””

upvoted 1 times

Morsa 2 years, 6 months ago

Answer is A

upvoted 1 times

ovokpus 2 years, 6 months ago

Selected Answer: A

This is a clear case of Data Augumentation solution.

upvoted 1 times

yc1005 2 years, 7 months ago

Selected Answer: A

Common step in CNN, Image augmentation. A.

upvoted 1 times

EXAMTOPICS

Amazon Discussions

Exam AWS Certified Machine Learning - Specialty All Questions

View all questions & answers for the AWS Certified Machine Learning - Specialty exam

Go to Exam

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 56 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 56
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A Machine Learning Specialist is preparing data for training on Amazon SageMaker. The Specialist is using one of the SageMaker built-in algorithms for the training. The dataset is stored in .CSV format and is transformed into a numpy.array, which appears to be negatively affecting the speed of the training.
What should the Specialist do to optimize the data for training on SageMaker?

A. Use the SageMaker batch transform feature to transform the training data into a DataFrame.

B. Use AWS Glue to compress the data into the Apache Parquet format.

C. Transform the dataset into the RecordIO protobuf format.

D. Use the SageMaker hyperparameter optimization feature to automatically optimize the data.

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 3:28 a.m.*

Comments

rsimham Highly Voted 2 years, 9 months ago

C is okay

upvoted 19 times

stamarpadar Highly Voted 2 years, 9 months ago

Anwer is C.
Most Amazon SageMaker algorithms work best when you use the optimized protobuf recordIO format for the training data.
https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html

upvoted 16 times

Mickey321 Most Recent 10 months, 2 weeks ago

Selected Answer: C

option C

upvoted 1 times

AjoseO 1 year, 5 months ago

Selected Answer: C

The Specialist should transform the dataset into the RecordIO protobuf format. This format is optimized for use with SageMaker and has been

shown to improve the speed and efficiency of training algorithms.

Using the RecordIO protobuf format is a best practice for preparing data for use with Amazon SageMaker, and it is specifically recommended for use with the built-in algorithms.

upvoted 1 times

Jeremy1 1 year, 7 months ago

Selected Answer: C

I would assume the issue is the transformation. It can be nasty slow between pandas / csv / numpy. Go to protobuf.

upvoted 1 times

C10ud9 2 years, 8 months ago

C is the best

upvoted 5 times

PRC 2 years, 8 months ago

Agree with C

upvoted 6 times

Amazon Discussions

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 55 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 55
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company is running a machine learning prediction service that generates 100 TB of predictions every day. A Machine Learning Specialist must generate a visualization of the daily precision-recall curve from the predictions, and forward a read-only version to the Business team.
Which solution requires the LEAST coding effort?

A. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Give the Business team read-only access to S3.

B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.

C. Run a daily Amazon EMR workflow to generate precision-recall data, and save the results in Amazon S3. Visualize the arrays in Amazon QuickSight, and publish them in a dashboard shared with the Business team.

D. Generate daily precision-recall data in Amazon ES, and publish the results in a dashboard shared with the Business team.

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 3:26 a.m.*

Comments

rsimham Highly Voted 3 years, 9 months ago
Ans C is reasonable
upvoted 28 times

cloud_trail Highly Voted 3 years, 8 months ago
Agree with C. Quicksight cannot handle 100TB each day.
upvoted 6 times

MultiCloudIronMan Most Recent 8 months, 3 weeks ago
Selected Answer: C
Amazon QuickSight, particularly when using its SPICE (Super-fast, Parallel, In-memory Calculation Engine) feature, has specific data capacity limits. For the Enterprise Edition, SPICE can handle up to 1 billion rows or 1 TB per dataset1. This means that while QuickSight is highly capable, handling 100 TB of data per day would exceed its current capacity limits.
upvoted 1 times

AMEJack 9 months, 1 week ago
Selected Answer: B

The limit of QuickSight for 1TB is soft limit which can be increased to unlimited number of TBs.
upvoted 1 times

Ali_Redha 1 year, 3 months ago
Ans C Because Quicksight Can't handle

100 TB even in Entiripise

Quotas for SPICE are as follows:

2,047 Unicode characters for each field

127 Unicode characters for each column name

2,000 columns for each file

1,000 files for each manifest

For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset

For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset

https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html
upvoted 2 times

VR10 1 year, 4 months ago
QuickSight can handle large volumes of data for analytics and visualizations. Some key points:

QuickSight scales seamlessly from hundreds of megabytes to many terabytes of data without needing to manage infrastructure.

It uses an in-memory engine called SPICE to enable high performance analytics on large datasets.
so the choice is B
upvoted 1 times

kyuhuck 1 year, 5 months ago
Selected Answer: B
B. Generate daily precision-recall data in Amazon QuickSight, and publish the results in a dashboard shared with the Business team.

This solution leverages QuickSight's managed service capabilities for both data processing and visualization, which should minimize the coding effort required to provide the Business team with the necessary insights. However, it's important to note that QuickSight's ability to calculate the precision-recall data depends on its support for the necessary statistical functions or the availability of such calculations in the dataset. If QuickSight cannot perform these calculations directly, option C might be necessary, despite the increased effort.
upvoted 1 times

Topg4u 1 year, 5 months ago
The question does not ask for processing of 1Tb data. it asks for visuals/predications of that data. So B
upvoted 2 times

phdykd 1 year, 6 months ago
C.
Considering the large volume of data (100 TB daily), Option C seems to be the most appropriate solution
upvoted 1 times

iskorini 1 year, 7 months ago
Selected Answer: C
B it's not correct because of 100tb data size.
C is the answer: https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html
upvoted 2 times

Snape 1 year, 8 months ago
Selected Answer: C
ANs c is correct
upvoted 1 times

loict 1 year, 10 months ago
Selected Answer: C
A. NO - we want a dashboard for business
B. NO - 100TB is very large, it will not fit in memory (1TB max for SPICE dataset) or return within the 2min limit if delegated to a DB (https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html)
C. YES - best combination; EMR can distribute the computation of precision-recall for each slice of data
D. NO - ES cannot help to generate precision-recall
upvoted 1 times

Mickey321 1 year, 10 months ago

Selected Answer: B

although C is tempting but goes with B due to less effort

upvoted 1 times

teka112233 1 year, 10 months ago

it is not about the least effort only, since the least effort solution here will not get your job done, look at the quick sight max data it can deal with when it compared to EMR which is built to deal with Big data.

upvoted 1 times

teka112233 1 year, 10 months ago

Selected Answer: C

using quick sight for creation of the precision recall with 100 TB every day cann't be done since the max size for quick sight to deal with is :
For Standard edition, 25 million (25,000,000) rows or 25 GB for each dataset
For Enterprise edition, 1 billion (1,000,000,000) rows or 1 TB for each dataset
acc to AWS documentation :
https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html
but we can do it with EMR and latterly use quick sight to visualize the results

upvoted 2 times

kaike_reis 1 year, 11 months ago

Selected Answer: C

Looking at the QuickSight documentation: it has a limit of 1 TB per dataset. So it's necessary a previous layer. Letter C is the correct one.

upvoted 1 times

ADVIT 2 years ago

It's 100TB daily, need EMR to reduce, option C is correct.

upvoted 1 times

petervu 2 years ago

Selected Answer: C

Quicksight can handle maximum 1TB data set only. We have 100TB data set so we need EMR.
https://docs.aws.amazon.com/quicksight/latest/user/data-source-limits.html

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 54 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty

Question #: 54

Topic #: 1

[All AWS Certified Machine Learning - Specialty Questions]

Which of the following metrics should a Machine Learning Specialist generally use to compare/evaluate machine learning classification models against each other?

A. Recall

B. Misclassification rate

C. Mean absolute percentage error (MAPE)

D. Area Under the ROC Curve (AUC)

Show Suggested Answer

by DonaldCMLIN at *Nov. 17, 2019, 4:07 a.m.*

Comments

DonaldCMLIN Highly Voted 3 years, 3 months ago

RECALL IS ONE OF FACTOR IN CLASSIFY,

AUC IS MORE FACTORS TO COMPREHENSIVE JUDGEMENT
https://docs.aws.amazon.com/zh_tw/machine-learning/latest/dg/cross-validation.html

ANSWER MIGHT BE D.

upvoted 38 times

devsean 3 years, 3 months ago

AUC is to determine hyperparams in a single model, not compare different models.

upvoted 6 times

DScode 3 years, 2 months ago

Not might be, but should be D

upvoted 5 times

AjoseO Highly Voted 1 year, 11 months ago

Selected Answer: D

Area Under the ROC Curve (AUC) is a commonly used metric to compare and evaluate machine learning classification models against each other. The AUC measures the model's ability to distinguish between positive and negative classes, and its performance across different classification thresholds. The AUC ranges from 0 to 1, with a score of 1 representing a perfect classifier and a score of 0.5 representing a classifier that is no

better than random.

While recall is an important evaluation metric for classification models, it alone is not sufficient to compare and evaluate different models against each other. Recall measures the proportion of actual positive cases that are correctly identified as positive, but does not take into account the false positive rate.

upvoted 5 times

ccpmad 1 year, 5 months ago

chatgpt answers, all your answers are from chatgpt

upvoted 2 times

AsusTuf Most Recent 1 year, 3 months ago

why not C?

upvoted 1 times

Scrook 8 months ago

it's a classification problem, mape is for regression

upvoted 2 times

Mickey321 1 year, 4 months ago

Selected Answer: D

option D

upvoted 1 times

Valcilio 1 year, 10 months ago

Selected Answer: D

AUC is the best metric.

upvoted 1 times

cloud_trail 3 years, 2 months ago

D. AUC is always used to compare ML classification models. The others can all be misleading. Consider the cases where classes are highly imbalanced. In those cases accuracy, misclassification rate and the like are useless. Recall is only useful if used in combination with precision or specificity, which what AUC does.

upvoted 4 times

harmanbirstudy 3 years, 2 months ago

AUC/ROC work well with special case of Binary Classification not in general

upvoted 5 times

MohamedSharaf 3 years, 2 months ago

AUC is to compare different models in terms of their separation power. 0.5 is useless as it's the diagonal line. 1 is perfect. I would go with F1 Score if it was an option. However, taking Recall only as a metric for comparing between models, would be misleading.

upvoted 4 times

harmanbirstudy 3 years, 2 months ago

Its Accuracy,Precision,Recall and F1 score , there is no metion of AUC/ROC for comparing models in many articles , so ANSWER is A

upvoted 1 times

DavidRou 1 year, 4 months ago

When you draw the ROC graph, you're considering True and False Positive Rate. The first one is also called Recall ;)

upvoted 1 times

Thai_Xuan 3 years, 2 months ago

D. AUC is scale- and threshold-invariant, enabling it compare models.
https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1

upvoted 1 times

johnny_chick 3 years, 2 months ago

Actually A, B and D seem to be correct

upvoted 1 times

deep_n 3 years, 2 months ago

Probably D
https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5

upvoted 2 times

hughhughhugh 3 years, 3 months ago

why not B?

upvoted 1 times

PRC 3 years, 3 months ago

Answer should be D..ROC is used to determine the diagnostic capability of classification model varying on threshold

upvoted 3 times

Hypermasterd 3 years, 3 months ago

Should be A. A is the only one that generally works for classifcation.
AUC only works with binary classification.

upvoted 4 times

oMARKOo 3 years, 2 months ago

Actually AUC could be generalized for multi-class problem.
https://www.datascienceblog.net/post/machine-learning/performance-measures-multi-class-problems/

upvoted 1 times

sebas10 3 years, 2 months ago

Could be, you mean in a multiclass clasification problem. But in that con context recall directly can't be compare because first you have to decide recall of what of the classes, in a 3 classes problem we have 3 recalls or you suppose a weighted recall or average recall ?. Do you think in that ?

upvoted 2 times

mrsimoes 3 years, 2 months ago

Also in multi-class classification, if you follow an One-vs_Rest strategy you can still use AUC.
https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py

upvoted 1 times

stamarpadar 3 years, 3 months ago

Correct Answer is D. Another benefit of using AUC is that it is classification-threshold-invariant like log loss.
https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226

upvoted 3 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 53 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 53
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A company's Machine Learning Specialist needs to improve the training speed of a time-series forecasting model using TensorFlow. The training is currently implemented on a single-GPU machine and takes approximately 23 hours to complete. The training needs to be run daily.
The model accuracy is acceptable, but the company anticipates a continuous increase in the size of the training data and a need to update the model on an hourly, rather than a daily, basis. The company also wants to minimize coding effort and infrastructure changes.
What should the Machine Learning Specialist do to the training solution to allow it to scale for future demand?

A. Do not change the TensorFlow code. Change the machine to one with a more powerful GPU to speed up the training.

B. Change the TensorFlow code to implement a Horovod distributed framework supported by Amazon SageMaker. Parallelize the training to as many machines as needed to achieve the business goals.

C. Switch to using a built-in AWS SageMaker DeepAR model. Parallelize the training to as many machines as needed to achieve the business goals.

D. Move the training to Amazon EMR and distribute the workload to as many machines as needed to achieve the business goals.

Show Suggested Answer

by rsimham at *Dec. 10, 2019, 3:24 a.m.*

Comments

JayK Highly Voted 3 years, 9 months ago

the answer is B. using Hovord distribution results in less coding effort

upvoted 38 times

cybe001 Highly Voted 3 years, 9 months ago

Answer is B. "minimize coding effort and infrastructure changes" If we use DeepAR then the code and infra has to be changed to work with DeepAR.

upvoted 15 times

ninomfr64 Most Recent 1 year ago

Selected Answer: C

A. NO, this will not address training dataset continuous increase
B. NO, this will require code effort and infrastructure change
C. YES, a built-in model ensure low code effort, so only infrastructure change needed*
D. This will not work

* they say current model accuracy is acceptable, we doo expect good results with DeepAR as it allows to automatically pick among 5 different models what works best for the customer

upvoted 4 times

ninomfr64 1 year ago

DeepAR doesn't pick among 5 models. However, I still think that switching to DeepAR can assure accuracy and minimize coding effort as the model is built-in

upvoted 2 times

VR10 1 year, 4 months ago

A comes with minimum changes, but it wont scale.
B code changes are minimum but infrastructure still needs to be changed to achieve a distributed solution.
C. Is even more significant infra and code change.
D. wont work.
It is really subjective and tricky.
Could be A or B, depending on what change is considered "SMALL".
For scalability, B seems better. for quick win A could work.
I keep going back and forth.

upvoted 1 times

loict 1 year, 10 months ago

Selected Answer: B

A. NO - one time shot and not scalable
B. YES - best practice
C. NO - DeepAR is for forecasting
D. NO - code will not benefit from parallelization without change

upvoted 4 times

Mickey321 1 year, 10 months ago

Selected Answer: B

option B

upvoted 2 times

kaike_reis 1 year, 11 months ago

Selected Answer: B

Note that we want to increase training speed, minimize code and infrastructure modification effort on AWS. Letter A would only delay the problem and increase costs too much. The solution that best translates the problem would be Letter B: we would keep the code in tensorflow and use Horovod to make our training faster through parallelization. Letter D is too complex and would change the execution infrastructure a lot and Letter C would be too abrupt a turn as we would throw our model away.

upvoted 2 times

ZSun 2 years, 2 months ago

A is better option even though B helps. Firstly, you only have One GPU, in this case distributed training Horovod doesn't help much;
Secondly, the question is about minimize "coding effort" not minimize budget. adding distributed framework require much more coding, but increase gpu instance only require single click.

upvoted 1 times

Valcilio 2 years, 4 months ago

Selected Answer: B

Horovod distribution is accepted by sagemaker, making easy to implement!

upvoted 1 times

AjoseO 2 years, 5 months ago

Selected Answer: B

Hovord distribution will allow the Machine Learning Specialist to take advantage of Amazon SageMaker's built-in support for Horovod, which is a popular, open-source distributed deep learning framework.

Implementing Horovod in TensorFlow will allow the Specialist to parallelize the training across multiple GPUs or instances, which can significantly reduce the time it takes to train the model.

This will allow the company to meet its requirement to update the model on an hourly basis, and minimize coding effort and infrastructure changes as it leverages the existing TensorFlow code and infrastructure, along with the scalability and ease of use of Amazon SageMaker.

upvoted 4 times

joe3232 2 years, 5 months ago

Are is there a 23X differential between the weakest and strongest GPU in AWS? (and allow for future growth). I don' tthink so.

upvoted 1 times

vbal 2 years, 6 months ago

Answer:C- built-in sagemaker DeepAR model. minimize coding & infra changes.

upvoted 1 times

cpal012 2 years, 4 months ago

But they are happy with it - just want it to go faster. Not throw the whole thing out.

upvoted 1 times

KingGuo 3 years ago

Selected Answer: B

the answer is B. using Hovord distribution results in less coding effort

upvoted 2 times

John_Pongthorn 3 years, 4 months ago

Selected Answer: A

Most likely , it is A because it is based on AWS teachnoloy, why we have to use open source

we exam AWS ML , the answer should be relevant to AWS technology inevitably
https://aws.amazon.com/sagemaker/distributed-training/

upvoted 1 times

cloud_trail 3 years, 8 months ago

This one reminds me of an old saying by Yogi Berra: "When you come to a fork in the road, take it." If you see Horovod as an option in a question about scaling TF, take it. Answer is B.

upvoted 9 times

RaniaSayed 3 years, 8 months ago

I Think it's B
https://aws.amazon.com/blogs/machine-learning/launching-tensorflow-distributed-training-easily-with-horovod-or-parameter-servers-in-amazon-sagemaker/
&
https://aws.amazon.com/blogs/machine-learning/multi-gpu-and-distributed-training-using-horovod-in-amazon-sagemaker-pipe-mode/

upvoted 5 times

harmanbirstudy 3 years, 8 months ago

Seen similar question on udemy/whizlab , its always Horvord when Tensorflow needs scaling. ANWSER is B

upvoted 5 times

EXAM AWS CERTIFIED MACHINE LEARNING - SPECIALTY TOPIC 1 QUESTION 52 DISCUSSION

Exam question from Amazon's AWS Certified Machine Learning - Specialty
Question #: 52
Topic #: 1
[All AWS Certified Machine Learning - Specialty Questions]

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:
⇨ Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
⇨ Support event-driven ETL pipelines
⇨ Provide a quick and easy way to understand metadata
Which approach meets these requirements?

A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.

B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.

C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.

D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

Show Suggested Answer

---

by DonaldCMLIN at *Nov. 17, 2019, 3:26 a.m.*

Comments

DonaldCMLIN Highly Voted 2 years, 9 months ago

BOTH A AND B ARE ANSWERS.

BUT external Apache Hive MIGHT BE NOT SERVERLESS SOLUTION.

The AWS Glue Data Catalog is your persistent metadata store. It is a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore.
The Data Catalog is a drop-in replacement for the Apache Hive Metastore

https://docs.aws.amazon.com/zh_tw/glue/latest/dg/components-overview.html

BEAUTIFUL ANSWER IS A.

upvoted 45 times

rsimham 2 years, 9 months ago

I am thinking about Answer C, because events can be triggered by cloudwatch w/Glue metastore

upvoted 1 times

qwerty456 2 years, 8 months ago

you can't schedule AWS Batch with CloudWatch

upvoted 4 times

kalyanvarma 2 years, 8 months ago

We can schedule batch with cloud watch events.

upvoted 1 times

qwerty456 2 years, 8 months ago

srr, looks like you can apart from Cron, the argument should be AWS Batch aren't SERVERLESS

upvoted 2 times

ComPah 2 years, 9 months ago

if we use Flexible as key word ..Using Lambda might be a constraint

upvoted 4 times

cybe001 Highly Voted 2 years, 9 months ago

Answer is A. Lamda is the preferred way of implementing event-driven ETL job with S3, when new data arrives in S3, it notifies lamda which can start the ETL job.

upvoted 18 times

rb39 1 year, 10 months ago

agree, event-driven means Lambda, CloudWatch alarms are just to trigger alarms based on log analysis.

upvoted 3 times

loict Most Recent 10 months ago

Selected Answer: A

A. YES - all integrated components
B. NO - missing a component to invoke the Lambda
C. NO - CloudWatch will not trigger when there is a new file to process
D. NO - CloudWatch will not trigger when there is a new file to process

upvoted 2 times

Mickey321 10 months, 2 weeks ago

Selected Answer: A

A for me

upvoted 1 times

kaike_reis 11 months, 2 weeks ago

Selected Answer: A

Note that the question asks for a serverless system. In this case, the letters B, C and D are wrong, as they bring options that are managed: AWS Batch (managed) and external Apache Hive (even more managed). For event-driven AWS ETL solutions that are serverless, activation through the Lambda function is recommended, so the correct alternative is Letter A. Note that CloudWatch Alarms only activates from log evaluation, which is not mentioned in the question.

upvoted 1 times

jackzhao 1 year, 4 months ago

I will chose A, I think C & D is wrong, you can use Amazon CloudWatch Event to trigger lambda but not CloudWatch alarm.

upvoted 1 times

Valcilio 1 year, 4 months ago

Selected Answer: A

Batch is more for configurations and other kinds of things by scheduling than event driven and batch data processing with ETL, the answer is A.

upvoted 1 times

Jeremy1 1 year, 7 months ago

Selected Answer: A

Found this supporting A - Lambda used to trigger ETL job after crawler completes. The crawler starts on schedules or events (files arriving).

upvoted 1 times

Skychaser 2 years ago

Selected Answer: A

Based on Majority discussion

upvoted 2 times

exam887 2 years, 1 month ago

Selected Answer: C

Quite confused between A&C since they all workable solution. In below AWS Blog, even mix the CloudWatch + Lambda to use the Glue. For key word event trigger, prefer CloudWatch
https://aws.amazon.com/blogs/big-data/build-and-automate-a-serverless-data-lake-using-an-aws-glue-trigger-for-the-data-catalog-and-etl-jobs/
https://docs.aws.amazon.com/glue/latest/dg/automating-awsglue-with-cloudwatch-events.html

upvoted 2 times

ZSun 1 year, 2 months ago

cloudwatch and lambda function can work together to trigger event. But AWS batch cannot independently conduct ETL and require other service. when it comes to ETL, glue is much easier choice than Batch

upvoted 1 times

VinceCar 1 year, 7 months ago

Agreed. CloudWatch could trigger event to launch Lambda. Refer to: https://docs.aws.amazon.com/lambda/latest/dg/services-cloudwatchevents.html

upvoted 1 times

syu31svc 2 years, 8 months ago

Answer is A 100%

upvoted 2 times

halfway 2 years, 8 months ago

A is preferred. Lambda can trigger ETL pipelines: https://aws.amazon.com/glue/

upvoted 3 times

PRC 2 years, 9 months ago

A is correct...Lambda is event driven and Glue is serverless as opposed to Hive

upvoted 4 times