

 Google Discussions



## Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

[Go to Exam](#)



## EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 128 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 128

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

[Show Suggested Answer](#)

by [deleted] at *March 22, 2020, 11:11 a.m.*

## Comments

Type your comment...

[Submit](#)

  **Callumr** Highly Voted 4 years, 10 months ago

This is a case of underfitting - not overfitting (for over fitting the model will have extremely low training error but a high testing error) - so we need to make the model more complex - answer is D

   upvoted 72 times

  **hellofrnds** 3 years, 7 months ago

@callumr , "root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set." clearly means testing error is twice of training error. So, it is clearly overfitting. Isn't it?

   upvoted 4 times

  **hellofrnds** 3 years, 6 months ago

So, answer should be C

   upvoted 1 times

  **tavva\_prudhvi** 3 years ago

If you training RMSE=0.2. and testing RMSE = 0.4, and we want the RMSE to be low as its the error, now is it overfitting or underfitting? think wisely!

   upvoted 2 times

  **alecuba16** 2 years, 8 months ago

It's overfitting.

Overfitting->low rmse in train / high accuracy-f1 score in train for classification.

Underfitting -> high rmse / low f1score or accuracy in train, you don't have to look into test set if there is an underfitting problem.

   upvoted 1 times

  **jfab** 1 year, 10 months ago

But the question clearly states we have higher RMSE on the train than the test. So how would it be overfitting?

  upvoted 1 times

[Load full discussion...](#)

  **velliger** 3 years, 5 months ago



High rmse: The model is underfitting the train data. To reduce overfitting, we increase the number of layers in the model or we change the type of layer.

   upvoted 1 times

  **velliger** 3 years, 5 months ago

\*underfitting

   upvoted 2 times

  **odacir** 2 years, 4 months ago

NO, its underfitting.

   upvoted 3 times

[Load full discussion...](#)

  **NeoNitin** 1 year, 9 months ago

Based on the given information, this scenario indicates a case of overfitting.


Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

   upvoted 1 times

  **ckanaar** 1 year, 7 months ago

Wrong! This scenario indicates a case of underfitting. The RSME is twice as high on the training dataset compared to the test dataset, so the model is underfitting.

   upvoted 2 times

  **[Removed]** Highly Voted 5 years, 1 month ago

should be D

   upvoted 20 times

  **NeoNitin** 1 year, 9 months ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

   upvoted 4 times

  **samtestking** Most Recent 4 months ago

**Selected Answer: B**

Could be B (data requirement for task is vague), but let's assume 100 million data points is enough and rule that out.


Indication of overfitting is significantly better performance on training data compared to unseen data. Here we are told that the unseen data is performing significantly better which is the opposite of what we should see if it were overfitting. Rule out C.

Symptoms of model underfitting is poor performance in BOTH training AND unseen data. While underfitting might be the issue, the more pressing concern is that the test set is clearly not representative of the overall data and could be skewed. This is further supported by the 90/10 split (academic/industry standard is 80/20 or 75/25 based on the Pareto principle: [https://en.wikipedia.org/wiki/Pareto\\_principle](https://en.wikipedia.org/wiki/Pareto_principle)). A 90/10 split would be useful if we were doing k-fold cross validation (<https://machinelearningmastery.com/k-fold-cross-validation/>), however there is no indication of such in the prompt.

Note: The question does not explicitly say that the model is performing poorly/errors are significantly bad, just that the error is twice as high in the training set (they could both have low error values).

So whilst it could be a case of underfitting (D), the first step taken should be addressing the obviously problematic data representation by adjusting the train-test split (option A).

   upvoted 1 times

  **SamuelTsch** 6 months, 1 week ago

**Selected Answer: D**

It is underfitting problem, which means that the used models is too easy.

   upvoted 2 times

  **baimus** 7 months, 1 week ago

This is A. The key is that 90/10 is a weirdly small test set, that stood out to me straight away (I work professionally as a machine learning engineer and have the cert). Next tip, that everyone seems to be ignoring - this is not underfit OR overfit. The model outperforms on the TEST set, this is not a miswording. Test scores higher than train. The time you might expect to see this is if your test set is too small to be a representative sample, leading to unrepresentative results. Seeing as the question already set up this conclusion with the 90/10 thing, it's definitely A. None of the others (or indeed anything else) can address Test outperforming Train, and the conclusion of others below that this is due to a poorly worded question is a bizarre conclusion.

   upvoted 1 times

  **cuadradobertolinisebastiancami** 1 year, 2 months ago

**Selected Answer: D**

Underfitting scenario

   upvoted 2 times

  **Sofia98** 1 year, 3 months ago

**Selected Answer: D**

It is an underfitting situation - D

   upvoted 2 times

  **Kimich** 1 year, 5 months ago

**Selected Answer: C**

Should be C

C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting:

This is a reasonable approach. Regularization techniques can help prevent overfitting, especially when the model shows a significantly higher error on the training set compared to the test set.

D. Increase the complexity of your model (e.g., introducing an additional layer or increasing the size of vocabularies or n-grams):

This could potentially exacerbate the overfitting issue. Increasing model complexity without addressing overfitting concerns may lead to poor generalization on new data.

   upvoted 2 times

  **Kimich** 1 year, 5 months ago

<https://dooinnkim.medium.com/what-are-overfitting-and-underfitting-855d5952c0b6>

   upvoted 1 times

🗨️ **hallo** 1 year, 5 months ago

Are the questions in this relevant for the new exam or are these all now outdated?

👍 ↩️ 🚩 upvoted 3 times

🗨️ **pss111423** 1 year, 5 months ago

<https://stats.stackexchange.com/questions/497050/how-big-a-difference-for-test-train-rmse-is-considered-as-overfit#:~:text=RMSE%20of%20test%20%3C%20RMSE%20of,is%20always%20overfit%20or%20underfit.>

RMSE of test > RMSE of train => OVER FITTING of the data.

RMSE of test < RMSE of train => UNDER FITTING of the data.

so for answer is D

👍 ↩️ 🚩 upvoted 1 times

🗨️ **steghe** 1 year, 5 months ago

Underfitting models: In general High Train RMSE, High Test RMSE.

Overfitting models: In general Low Train RMSE, High Test RMSE.

<https://davidalpiroz.github.io/r4sl/regression-for-statistical-learning.html>

👍 ↩️ 🚩 upvoted 1 times

🗨️ **ha1p** 1 year, 7 months ago

I passed the exam today. I am pretty sure it is overfitting. Answer must be c

👍 ↩️ 🚩 upvoted 2 times

🗨️ **MULTITASKER** 1 year, 7 months ago

**Selected Answer: D**

RMSE is more on training. That means, model is not performing well on training dataset but performing well on testing dataset. This happens in the case of underfitting. So D.

👍 ↩️ 🚩 upvoted 3 times

🗨️ **[Removed]** 1 year, 7 months ago

**Selected Answer: D**

RMSE training = 2 x testing

When training > testing, it is a case of underfitting

Hence D

👍 ↩️ 🚩 upvoted 2 times

🗨️ **pulse008** 1 year, 8 months ago

chatGPT says option C

👍 ↩️ 🚩 upvoted 1 times

🗨️ **stonefl** 1 year, 8 months ago

**Selected Answer: D**

"root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set." means the RMSE of training set is two time of RMSE of test set, which indicates the training is not as good as test, then underfitting, so D.

👍 ↩️ 🚩 upvoted 2 times

🗨️ **NeoNitin** 1 year, 9 months ago

Based on the given information, this scenario indicates a case of overfitting.

Overfitting occurs when a machine learning model performs well on the training data but poorly on unseen data (test data). In this case, the root-mean-squared error (RMSE) of the model is twice as high on the train set (the data used for training) compared to the test set (the data used for evaluation). This suggests that the model is fitting the training data too closely and is not generalizing well to new, unseen data.

So with dropout method we can overcome the overfitting so C is correct

👍 ↩️ 🚩 upvoted 1 times

[Load full discussion...](#)

> Home

> All Exams

> Examtopics PRO

> Training Courses

