Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 53 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 53

Topic #: 1

[All Professional Machine Learning Engineer Questions]

Your team is working on an NLP research project to predict political affiliation of authors based on articles they have written. You have a large training dataset that is structured like this:

```
AuthorA:Political Party A
    TextA1: [SentenceA11, SentenceA12, SentenceA13, ...]
    TextA2: [SentenceA21, SentenceA22, SentenceA23, ...]
    ...
AuthorB:Political Party B
    TextB1: [SentenceB11, SentenceB12, SentenceB13, ...]
    TextB2: [SentenceB21, SentenceB22, SentenceB23, ...]
    ...
AuthorC:Political Party B
    TextC1: [SentenceC11, SentenceC12, SentenceC13, ...]
    TextC2: [SentenceC21, SentenceC22, SentenceC23, ...]
    ...
AuthorD:Political Party A
    TextD1: [SentenceD11, SentenceD12, SentenceD13, ...]
    TextD2: [SentenceD21, SentenceD22, SentenceD23, ...]
    ...
...
```

You followed the standard 80%-10%-10% data distribution across the training, testing, and evaluation subsets. How should you distribute the training examples across the train-test-eval subsets while maintaining the 80-10-10 proportion?

A. Distribute texts randomly across the train-test-eval subsets: Train set: [TextA1, TextB2, ...] Test set: [TextA2, TextC1, TextD2, ...] Eval set: [TextB1, TextC2, TextD1, ...]

B. Distribute authors randomly across the train-test-eval subsets: (*) Train set: [TextA1, TextA2, TextD1, TextD2, ...] Test set:

[TextB1, TextB2, ...] Eval set: [TexC1,TextC2 ...]

C. Distribute sentences randomly across the train-test-eval subsets: Train set: [SentenceA11, SentenceA21, SentenceB11, SentenceB21, SentenceC11, SentenceD21 ...] Test set: [SentenceA12, SentenceA22, SentenceB12, SentenceC22, SentenceC12, SentenceD22 ...] Eval set: [SentenceA13, SentenceA23, SentenceB13, SentenceC23, SentenceC13, SentenceD31 ...]

D. Distribute paragraphs of texts (i.e., chunks of consecutive sentences) across the train-test-eval subsets: Train set: [SentenceA11, SentenceA12, SentenceD11, SentenceD12 ...] Test set: [SentenceA13, SentenceB13, SentenceB21, SentenceD23, SentenceC12, SentenceD13 ...] Eval set: [SentenceA11, SentenceA22, SentenceB13, SentenceD22, SentenceC23, SentenceD11 ...]

Show Suggested Answer

by inder0007 at *July 6, 2021, 6:49 a.m.*

Comments

Type your comment...

Submit

⊟ **rc380** Highly Voted 3 years, 2 months ago

I think since we are predicting political leaning of authors, perhaps distributing authors make more sense? (B)

upvoted 19 times

⊟ **sensev** 3 years, 2 months ago

Agree it should be B. Since every author has his/her distinct style, splitting different text from the same author across different set could result in data label leakage.

upvoted 7 times

⊟ **dxxdd7** 3 years, 2 months ago

I don't agree as we want to know the political affiliation from a text and not based on an author. I think A is better

upvoted 1 times

⊟ **jk73** 3 years, 1 month ago

it is the political affiliation from a text, but to whom belong that text?
The statement clearly says ... Predict political affiliation of authors based on articles they have written. Hence the political affiliation is for each author according to the text he wrote.

upvoted 2 times

⊟ **jk73** 3 years, 1 month ago

Exactly! I also consider is B
Check this out!
If we just put inside the Training set , Validation set and Test set , randomly Text, Paragraph or sentences the model will have the ability to learn specific qualities about The Author's use of language beyond just his own articles. Therefore the model will mixed up different opinions.
Rather if we divided things up a the author level, so that given authors were only on the training data, or only in the test data or only in the validation data. The model will find more difficult to get a high accuracy on the test validation (What is correct and have more sense!). Because it will need to really focus in author by author articles rather than get a single political affiliation based on a bunch of mixed articles from different authors.

https://developers.google.com/machine-learning/crash-course/18th-century-literature

upvoted 12 times

⊟ **inder0007** Highly Voted 3 years, 4 months ago

Should be A, we are trying to get a label on the entire text so only A makes sense

upvoted 8 times

⊟ **GogoG** 3 years ago

Correct answer is B - https://developers.google.com/machine-learning/crash-course/18th-century-literature

upvoted 5 times

Dunnoth 1 year, 8 months ago

This is a known study. if you use A, the moment a new author is given in a test set the accuracy is waay low than what your metrics might suggest. To have realistic evaluation results it should be B. Also note that the label is for the "authour" not a text.

upvoted 1 times

PhilipKoku Most Recent 5 months ago

Selected Answer: B

B) Authors

upvoted 1 times

girgu 5 months, 1 week ago

Selected Answer: B

We have divide / split at author level. Other wise model will used text to author relationship but we want to find text to political affiliation relation ship. While prediction we already know text to author relation but we want to find text to political relation (and therefore author to political relation is implied.

upvoted 1 times

tavva_prudhvi 1 year, 4 months ago

Selected Answer: B

This is the best approach as it ensures that the data is distributed in a way that is representative of the overall population. By randomly distributing authors across the subsets, we ensure that each subset has a similar distribution of political affiliations. This helps to minimize bias and increases the likelihood that our model will generalize well to new data.

Distributing texts randomly or by sentences or paragraphs may result in subsets that are biased towards a particular political affiliation. This could lead to overfitting and poor generalization performance. Therefore, it is important to distribute the data in a way that maintains the overall distribution of political affiliations across the subsets.

upvoted 3 times

M25 1 year, 6 months ago

Selected Answer: B

Went with B

upvoted 1 times

John_Pongthorn 1 year, 8 months ago

Selected Answer: B

https://cloud.google.com/automl-tables/docs/prepare#split
https://developers.google.com/machine-learning/crash-course/18th-century-literature

upvoted 1 times

enghabeth 1 year, 9 months ago

Selected Answer: B

Ans B
The model is to predict which political party the author belongs to, not which political party the text belongs to... You do not have the information of the political party of each text, you are assuming that the texts are associated with the political party of the author.

upvoted 1 times

bL357A 2 years, 2 months ago

Selected Answer: A

label is party, feature is text

upvoted 1 times

suresh_vn 2 years, 2 months ago

IMO, B is correct
A,C,D label leakaged

upvoted 1 times

ggorzki 2 years, 9 months ago

Selected Answer: B

https://developers.google.com/machine-learning/crash-course/18th-century-literature
Split by authors, otherwise there will be data leakage - the model will get the ability to learn author specific use of language

upvoted 6 times

NamitSehgal 2 years, 10 months ago

B I agree

upvoted 1 times

JobQ 2 years, 10 months ago

I already saw the video in: https://developers.google.com/machine-learning/crash-course/18th-century-literature

Based on this video I concluded that the answer is A. What answer B is saying is that you will have Author B's texts in the training set, Author A's texts in the testing set and Author C's texts in the validation set. According to the video B is incorrect.

We want to have texts from author A in the training, testing and validation set. So A is correct. I think most people are choosing B because the word "author" but let's be careful.

upvoted 2 times

giaZ 2 years, 8 months ago

I though the same initially, but no..We'd want texts from author A in the training, testing and validation set if the task was to predict the author from a text (meaning, if the label was the author..right? You train the model to learn the style of text and connect it to an author. You'd need new texts from the same author in the test and validation sets, to see if the model is able to recognize him/her). HERE, the task is to predict political affiliation from a text of an author. The author is given. In the test and validation sets you need new authors, to see wether the model is able to guess their political affiliation. So you would do 80 authors (and corresponding texts) for training, 10 different authors for validation, and 10 different ones for test.

upvoted 5 times

pddddd 3 years, 1 month ago

Partition by author - there is an actual example in Coursera 'Production ML systems' course

upvoted 1 times

Macgogo 3 years, 1 month ago

I think it is B.
--
Your test data includes data from populations that will not be represented in production.

For example, suppose you are training a model with purchase data from a number of stores. You know, however, that the model will be used primarily to make predictions for stores that are not in the training data. To ensure that the model can generalize to unseen stores, you should segregate your data sets by stores. In other words, your test set should include only stores different from the evaluation set, and the evaluation set should include only stores different from the training set.
https://cloud.google.com/automl-tables/docs/prepare#ml-use

upvoted 4 times

Danny2021 3 years, 1 month ago

Should be D. Please see the dataset provided, it is based on the text / paragraphs.

upvoted 1 times

george_ognyanov 3 years ago

Have a look at the link the other have already provided twice. Splitting sentence by sentence is literally mentioned in said video as a bad example and something we should not do in this case.

upvoted 1 times

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.