Questions & Answers

COMPLETE

Instructor: Applied AI Course **Duration:** 30 mins

You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)
(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
Is rotation necessary in PCA? If yes, Why? https://google-interview-hacks.blogspot.com/2017/04/is-rotation-necessary-in-pca-if-yes-why.html
You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?(https://www.linkedin.com/pulse/questions-machine-learning-statistics-can-you-answer-saraswat/)

If you face any new Interview questions please put in comments ,we will work it out

Prev

Revision Questions

Next

Exploratory Data Analysis

285 Comment(s)

search comments Search

Leave a response

Format

Submit

kjghjghfhgdgfdhdtdasdarewtrsdfdfhdjh *11 Votes*

https://www.analyticsvidhya.com/blog/2017/03/questions-dimensionality-reduction-data-scientist/?utm_medium=social&utm_source=linkedin.com&utm_campaign=buffer

Have a look on this.

Reply *Jun 21, 2018 14:11 PM*

AppliedAICourse

nice one :)

Reply

Jun 21, 2018 14:38 PM

Anchit

Suppose we are using dimensionality reduction as pre-processing technique, i.e, instead of using all the features, we reduce the data to k dimensions with PCA. And then use these PCA projections as our features. Which of the following statement is correct?
A. Higher ‘k’ means more regularization
B. Higher ‘k’ means less regularization
C. Can’t Say

Solution: (B)
Higher k would lead to less smoothening as we would be able to preserve more characteristics in data, hence less regularization.

I am not able to understand this question. What is regularization and how is it related to PCA ?

Reply

Dec 19, 2018 11:41 AM

Applied AI Course

This is an interesting question.
Audio reply: https://soundcloud.com/applied-ai-course/comment-pca-regularization/s-srZV3

Reply

Dec 20, 2018 06:16 AM

Dinesh_G

sir ,what is mean by overfitting and what is the actual meaning of regularization ??

Reply

Mar 20, 2019 14:03 PM

AppliedAI Course

Overfitting means the model performs very well on the training data but it is unable to perform well on unseen data. Regularization is a technique to reduce overfitting.
Please go through the remaining videos. You will understand it better after studying Logistic regression videos.

Reply

Mar 20, 2019 18:52 PM

Sunney Sood

one question in this link is: 2) [True or False] It is not necessary to have a target variable for applying dimensionality reduction algorithms.

A. TRUE

B. FALSE

Solution: (A)

LDA is an example of supervised dimensionality reduction algorithm.

Are we covering LDA ?

Reply

Apr 17, 2019 13:04 PM

Applied AI Tech Admin

As of now we are not covering LDA. You can refer about it through online resources and if you ave any queries, you can mail us at **team@appliedaicourse.com**

Reply *Apr 17, 2019 15:44 PM*

Jibin V

Why the above answer is true.? Do we need a target variable to apply PCA? I think we don't need. Then why TRUE

Reply *Sep 19, 2020 00:03 AM*

team aaic

here the question was "it is 'not' necessary to have a target variable" so it is true only right we do not need a target variable

Reply *Sep 19, 2020 05:13 AM*

Jibin V

Sorry team. I read it the other way

Reply *Sep 19, 2020 11:27 AM*

prasad4ever

Sir, in the above question it says after reducing to k-dimension and then PCA projects are used as features.
However, in our earlier videos, we have learned that we create PCA to reduce dimensions to visualize only.
But, we never used PCA components as features in our exercise.
Please advice, what exactly to
Either PCA to only visualize data and model on original data?
or we can use PCA projects as new features for our model?
please advice.

Reply *Jul 28, 2019 00:10 AM*

AppliedAI

That's the same meaning reducing the k-dimensions means , we are reducing the features from d to k and after that we can use that k linearly independent features to apply in a model.

Reply *Jul 29, 2019 00:42 AM*

prasad4ever

Thank you sir for the answer.
How about t-sne? Is it the same way that we can use the reduced dimensions as features of new model?

Reply *Jul 30, 2019 02:19 AM*

Applied AI Tech Admin

Yes, after reducing the features we can use the new set of features in building a model.

Reply *Jul 30, 2019 10:48 AM*

anuj shrivastav

But t-SNE is just for *visualization*, right?

Reply *Oct 05, 2019 20:37 PM*

Applied AI Tech Admin

T-SNe is used for dimensionality reduction and also for visualization prpose.

Reply *Oct 05, 2019 23:11 PM*

Venkat

Which means only those 2 dimensions will be used for ML modeling?

Reply *Nov 27, 2019 21:57 PM*

AppliedAI Course

No, dimensionality reduction means we can use d' dimensions from d dimensions. d-original number of dimension, d'-number of dimensions after tsne. d'<d. For visualization, d' can be 2 or 3.

Reply *Nov 28, 2019 07:02 AM*

Sanjay235

How is it possible for dimensionality reduction are there any other dimension vectors being stored for doing transform on new test data that are got using the train data?
I don't find any **transform** method in SKLearn API's TSNE reference. Please clarify on this?

It would be great if you could show a snippet of how to use T-SNE on train data and test data for dimensionality reduction(not visualization) and train models on top of it.

Reply *Feb 03, 2020 14:17 PM*

Applied AI Tech Admin

```
from sklearn.manifold import TSNE
tnse_instance = TSNE(n_components =
100, perplexity = 40, metric = 'euclidean')
tnse_instance.fit(X_train)
X_dimensionality_reduced_train =
tnse_instance.transform(X_train)
X_dimensionality_reduced_test =
tnse_instance.transform(X_test)
```

The above mentioned code sample is an example of how to use TSNE in transforming the data and reducing the dimensionality. Please refer this documentation

Reply *Feb 05, 2020 00:32 AM*

Sanjay235

I know that how to use it.

Thanks, but that doesn't answer my question because sklearn's TSNE API doesn't has a *transform* method. Check the documentation provided by you carefully.

Reply *Feb 05, 2020 17:15 PM*

AppliedAI Course

Yes, you are correct. Please look at this link.

Reply

Feb 07, 2020 09:13 AM

Sanjay235

Then how can I use TSNE for incoming new data dimensionality reduction as you said previously in this thread? What does TSNE actually learn during the fitting process?

Reply

Feb 07, 2020 10:31 AM

AppliedAI Course

t-SNE makes a projection that tries to keep pairwise distances between the samples that you fit.
Please refer to this link.

Reply

Feb 08, 2020 07:25 AM

Naveen Kumar

this link is not working even I have tried in multiple browsers and have refreshed many times. Can you pls check once

Reply *Jan 15, 2020 22:21 PM*

team aaic

Hey Naveen, both of the links (one for the article and one for audio reply) are working at my end. Can you please check them again?

Article link: https://www.analyticsvidhya.com/blog/2017/03/questions-dimensionality-reduction-data-scientist/?utm_medium=social&utm_source=linkedin.com&utm_campaign=buffer

Audio reply: https://soundcloud.com/applied-ai-course/comment-pca-regularization/s-srZV3

If anyone of them doesn't work this time too, please revert back.

Reply *Jan 15, 2020 22:44 PM*

Kuruva Ramanjaneyulu *5 Votes*

You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do?
for this question, They told to remove correlated variables in answer. But if removing correlated variables is not recommended because variance explained by principal components is inflated by correlated variables. so what we have to do in this case?

Reply *Sep 26, 2018 13:43 PM*

AppliedAI

We can calculate the pearson correlation coefficient[range 0-1] between the columns and filter out columns that are correlated above threshold lets say (.90).

Reply *Sep 26, 2018 16:52 PM*

pavankumar2978

This is just to save computation time right? if I have enough resources , not removing some correlated columns will give better result. is my understanding correct?

Reply *Oct 17, 2018 01:30 AM*

AppliedAI

Not exactly, more correlated variables will over-emphasize particular eigenvectors, (directions), and if there are many correlated variables, then there would be so many more overemphasized 'fake' directions, that drown out an 'original' eigenvector/direction that would have otherwise been easily seen i.e., The reality is that a set of correlated variables might "load" onto several principal components (eigenvectors), so including many variables from such a set will differentially weight several eigenvectors and thereby change the directions of all eigenvectors, too. further reading:
https://stats.stackexchange.com/questions/50537/should-one-remove-highly-correlated-variables-before-doing-pca

Reply *Oct 17, 2018 03:51 AM*

Subrahmanyam Kesani

Should we "calculate the pearson correlation coefficient[range 0-1] between the columns and filter out columns that are correlated above threshold lets say (.90)" as a 1st step before applying PCA ... before column standardization?

Reply *May 11, 2020 17:49 PM*

AppliedAI

yes, Because by definition the Pearson correlation coefficient is independent of the change of origin and scale. As such standardization will not alter the value of correlation.

Reply *May 12, 2020 18:56 PM*

saurabh singh — *3 Votes*

How are PCA and LDA linear and TSNE a non linear technique??

Reply *Jun 09, 2019 21:54 PM*

AppliedAI

The linearity in PCA refers to the fact that, to perform the dimensionality reduction, you are projecting vectors into a lower-dimensionality space through means of a linear transformation i.e. the data is projected onto a lower dimensional linear subspace (hyperplane), as opposed to a nonlinear manifold. This hyperplane is the linear subspace generated by the eigenvectors corresponding to the largest eigenvalues of the covariance matrix.

t-SNE, unlike PCA, is not a linear projection. It uses the local relationships between points to create a low-dimensional mapping. This allows it to capture non-linear structure.

Reply *Jun 09, 2019 22:50 PM*

Sagar Verma — *2 Votes*

Hello Sir,
Could you please explain why can't we use t-sne as dimensionality reduction technique?
https://stats.stackexchange.com/questions/340175/why-is-t-sne-not-used-as-a-dimensionality-reduction-technique-for-clustering-or

Reply *Feb 27, 2020 20:38 PM*

team aaic

This answer to the same question provides the correct details about why t-SNE is not used as dim. reduction technique. Refer to this.

Reply *Feb 27, 2020 22:27 PM*

Uttam Dey

I didn't get the answer what is being posted for t-SNE's comparison with a clustering why it is not used or considered a method in clustering?

Reply *May 13, 2020 22:27 PM*

team aaic

It's because t-SNE doesn't learn any pattern from the data. It just tries to encode the high-dimensional data onto low-dim. space by looking at the distance between the points. So, t-SNE will fail on new unseen data points.

The same thing has been said in this answer of this thread on SO.

Reply *May 13, 2020 22:43 PM*

raviraj shinde — *2 Votes*

Hi,
Suppose if we have correlated features and we use pca it will give false results , so should we use different technique to remove correlated features and then use PCA.

Reply *Nov 20, 2019 10:26 AM*

Applied_AI

Yes, more correlated variables will over-emphasize particular eigenvectors, and if there are many correlated variables, then the original eigenvector would be overlooked. It is

therefore recommended to remove correlated features.

Reply *Nov 20, 2019 10:46 AM*

raviraj shinde

ok ,
then the step would we
1) by spearman correlation coefficeint identify the correlated colums
2) remove the corrlated colums with some threshold
3) use pca to further reduce columns
is this understanding ok

Reply *Nov 20, 2019 11:18 AM*

Applied_AI

Yes, right.

Reply *Nov 20, 2019 11:18 AM*

Athmuri

How do I identify that threshold?

Reply *Apr 02, 2020 11:12 AM*

team aaic

That's majorly dependent on the problem and the dataset we've at the end.

We usually try different values for threshold and select the one that gives the best result out of many.

Reply *Apr 02, 2020 22:33 PM*

Tushar Verma *2 Votes*

sir after going through the link u mentioned for question 3 i still have a doubt. can u explain it little bit
i not clear with the line " the variance explained by a particular component gets inflated" i mean how?

Reply *Jul 24, 2018 17:41 PM*

Applied AI Course

Audio reply: https://soundcloud.com/applied-ai-course/variance-inflation-pca/s-JB1tt

Reply *Jul 26, 2018 05:36 AM*

Mohit Kumar

Sir,, you have mentioned something called Principal Components. Does it means the eigen Values?

Reply *Mar 03, 2019 17:31 PM*

Applied AI Course Team1

These are eigen vectors here corresponding to each eigen values.

Reply *Mar 04, 2019 05:46 AM*

Mohit Angrish

Can you please tell about the intuition behind the inflation i.e. why does that inflation happens and how co- related features effects pca?

Reply

Sep 08, 2019 18:14 PM

Applied AI Course Team1

can you please what you mean by inflation here.Correlated features are removed and a new transformed feature is created using PCA.

Reply

Sep 09, 2019 00:19 AM

Mohit Angrish

"in presence of correlated variables, the variance explained by a particular component gets inflated"
i want to know why this happens and whats the intuition behind it?

Reply

Sep 09, 2019 01:34 AM

Applied AI Course Team1

What happes when we have correlated variable is PCA try to combine the same information/variance by combining into one principal components.Hence a single principal component would be expressing the multiple variance hence variance gets inflated.
For example lets say we have 3 variables with us and two of them are correlated then,first principal component would exhibit twice the variance present in first principal component.

Reply

Sep 09, 2019 21:42 PM

Mohit Angrish

can you please tell me what do you mean by this statement, i am unable to relate to it:
"first principal component would exhibit twice the variance present in first principal component."

Reply

Sep 09, 2019 22:40 PM

Applied AI Tech Admin

Could you please help us with the timestamp where you've encountered this statement in the video?

Reply

Sep 09, 2019 23:08 PM

Mohit Angrish

this statement is taken from the above mention comment by "Applied AI Course Team1 ".

Reply

Sep 10, 2019 22:56 PM

Applied AI Tech Admin

Yes when there exists correlation among the features, after applying PCA, the variance among the corelated features will be preserved in a single component. Hence we say that after dimensionality reduction, we do not see collinearity existing among the features.

Reply *Sep 11, 2019 15:59 PM*

abhishek

I did not get why variance gets increased (doubled) for PC1 when we have 3 features out of which 2 are correlated compared to 3 features in which no variables are correlated
Case 1: When we have 3 features and 2 are correlated and lets say variance is maximum towards these two correlated variables which means lambda1 and lambda2 will be same lets say 3 and lets say lambda3 is 1, so percentage of variance explained by lambda1 is 3/3+3+1 = 3/7

Case 2: When out of 3 features, no features are highly correlated. lets say lambda1 is 3 , lambda2 is 0 and lambda3 is 1, so in this case percentage of variance explained by lambda1 is 3/3+0+1 = 3/4

clearly, in case 2 we have higher variance for PC1

Reply *Nov 11, 2019 06:56 AM*

Applied AI Tech Admin

If there is no correlation among the features, then the amount of variance retained is the highest an the maximum.
When there exists correlation among the features, then there could be redundancy in the data and the amount of variance retained is less than the amount of variance retained in the above case(with no correlation).

Please refer to this blog

Reply

Nov 11, 2019 13:11 PM

Uttam Dey

It can be thought as an additional weightage is given in the presence of correlated variables when compared to data which dont have correlated features.Due to this a wrong direction can be followed in presence of correlated. One more reason is if we have data which dont have correlated variables we wont miss the feature which gives maximum information but if we have data which has correlated features in it suppose a feature in this data is correlated with another feature PCA would mislead the 1st component as it might be of less information but it seems to be more because of correlated feature presence in the dataset.

Reply *May 13, 2020 22:00 PM*

Applied AI Course Team1

You can take that way,but in the end PCA wipe out the correlation and just gives the correct transformed features with zero correlation

Reply *May 14, 2020 21:03 PM*

Venkat

Can we see those newly transformed feature in data frame, if yes can you place code snippet here?

Reply *Nov 27, 2019 21:59 PM*

Applied AI Course Team1

Yes we can surely print those transformed features.When you do fit_transform() the train data you get with proposed principal components can be easily printed.I hope for this small operation you do not need the code.As with the flow you will get to see the code.

```
from sklearn.decomposition import PCApca =
PCA(n_components=2)
principalComponents = pca.fit_transform(x)
principalDf = pd.DataFrame(data = principalComponents
               , columns = ['principal component 1',
'principal component 2'])
```

print(principalDF)

Reply *Nov 27, 2019 22:36 PM*

Pranjul Mittal *1 Votes*

You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?

Reply *May 15, 2020 21:33 PM*

team aaic

yes because the more different the variables are the better it is. Variables are correlated means the information is kind of same so we prefer to remove correlated first.

Reply *May 15, 2020 21:37 PM*

---

abhishekrawat *1 Votes*

Hello Sir,
For removing correlated variables(categorical) we use Chi-Squared test. Can you please explain what the test basically is?

Reply *Jun 12, 2018 09:59 AM*

Applied AI Course

Yes, we can use chi-squared test to see if two categorical variables are correlated. It is also referred to as pearosn's chi-square test. Here is a very nice example which is very easy to follow: https://en.wikipedia.org/wiki/Chi-squared_test#Example_chi-squared_test_for_categorical_data Here is a very good video explaining the concept with a good example: https://www.khanacademy.org/math/statistics-probability/inference-categorical-data

good example: https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/pearson-s-chi-square-test-goodness-of-fit

Reply *Jul 04, 2018 05:32 AM*

Sai Kumar

Will you explain the concept of the chi-squared test in the later section of the course?

Reply *Feb 11, 2019 21:28 PM*

AppliedAI Team

We didn't explain the chi-squared test in this course. we are updating some of the course content within a couple of months. in this, we are planning on adding the chi-squared test. please go through those link if you want to learn. if you have any doubt regarding that please feel free to contact us.

Reply *Feb 12, 2019 11:36 AM*

Sai Kumar

Can you please provide code explaining the chi square test and how it helps in choosing categorical values??. I have found multiple articles but nothing's been helpful. I really couldn't understand it. Please provide code for it. You can share the .ipynb file. Please it will be really helpful not just me for anyone else looking for it.

If possible please explain it for titanic dataset.

df = pd.read_csv('http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv')

Reply *Feb 15, 2019 21:06 PM*

AppliedAI Team

Check this blog. it was explained with titanic dataset only.

Reply *Feb 16, 2019 16:37 PM*

shravs

Sir,

From where is this Chi-Squared test coming from? is this related to PCA/t-sNE? I never heard of this being explained anywhere until lesson 15.x.

May be I would have missed. Can you please point me to where this test has been covered?

Reply *Sep 03, 2018 23:40 PM*

---

kuldeep singh *0 Votes*

9) Suppose we are using dimensionality reduction as pre-processing technique, i.e, instead of using all the features, we reduce the data to k dimensions with PCA. And then use these PCA projections as our features. Which of the following statement is correct?

A. Higher ‘k’ means more regularization

B. Higher ‘k’ means less regularization

C. Can’t Say

could you explain me this question?

Reply

Feb 25, 2021 15:35 PM

team aaic

suppose when we have high dimensional data and suppose it may cause curse of dimensionality then we apply pca and reduce the dimensions and give this reduced dimensional data to modeling. we use reduced dimensional projections as our features

Reply

Feb 25, 2021 17:50 PM

Srijan Banerjee *0 Votes*

Why we should remove the correlated variables first in the 3rd question? Why Chi Sq test for categorical variables?

Reply

Feb 20, 2021 12:22 PM

team aaic

Please take a look at this comment. The Chi-square test of independence determines whether there is a statistically significant relationship between categorical variables. It is a hypothesis test that tells whethe the values of one categorical variable depend on the value of other categorical variables or not.

Reply

Feb 20, 2021 12:37 PM

Srijan Banerjee

Thanks a lot for team for your response.Is Chi Squared test covered in course videos??

Reply

Feb 21, 2021 10:45 AM

team aaic

No, we have not covered it in the course.

Reply

Feb 21, 2021 10:55 AM

Previous 1 2 3 4 5 6 Next