Questions and Answers for PCA

Bingzhang Zhu

December 2, 2021

1 Explain the Curse of Dimensionality?

The curse of dimensionality refers to all the problems that arise working with data in the higher dimensions. As the number of features increase, the number of samples increases, hence, the model becomes more complex. The more the number of features, the more the chances of overfitting. A machine learning model that is trained on a large number of features, gets increasingly dependent on the data it was trained on and in turn overfitted, resulting in poor performance on real data, beating the purpose. The fewer features our training data has, the lesser assumptions our model makes and the simpler it will be.

2 Can PCA be used to reduce the dimensionality of a highly nonlinear dataset?

PCA can be used to significantly reduce the dimensionality of most datasets, even if they are highly nonlinear because it can at least get rid of useless dimensions. However, if there are no useless dimensions, reducing dimensionality with PCA will lose too much information.

3 How can you evaluate the performance of a dimensionality reduction algorithm on your dataset?

Intuitively, a dimensionality reduction algorithm performs well if it eliminates a lot of dimensions from the dataset without losing too much information. Alternatively, if you are using dimensionality reduction as a preprocessing step before another Machine Learning algorithm (e.g., a Random Forest classifier), then you can simply measure the performance of that second algorithm; if dimensionality reduction did not lose too much information, then the algorithm should perform just as well as when using the original dataset.

Reference: https://alekhyo.medium.com/interview-questions-on-pca-9cdc96ddaa9f