Pivotal

Manish Tripathi <mtripathi@pivotal.io>

Thursday Thunder: Week 6, Learning 6: PCA % Var

Manish Tripathi <mtripathi@pivotal.io> Thu, Jan 21, 2016 at 8:22 AM
To: PDS-ALL <PDS-ALL@pivotal.io>

Hey

I don't have a great topic to talk about this week, something I hope to have next week after few readings. However this week wanted to share a small trivia on PCA ,courtesy Ambarish. He helped me learn this during one of our IM conversation a week back. So this week "Thunder" is by Joshi :D

How to get % Variance Captured by Principal Components in PCA:

Given a situation which he described.

12K features, wants to reduce to 20 dimensions. So instead of getting an eigen decomposition matrix for all eigenvectors and eigenvalues, he kept it only for 20.And now wants to check what is the variation captured by these 20 PC's.

Suppose you don't have access to any scree plot function or any other function which tells you % variance captured, then how will you find the % variance captured. Like using madlib and don't have access to such functions.

Hmm. Why can't I have whole matrix of eigenvectors.

You can. Just that running PCA using some library, he kept the no. of components as 20. So it gives him a matrix with 20 components only and removes others.

Ok. Joshi doesn't keep anything which is not required. Parsimonious :P

Yeah. Except that he is parsimonious on sweets too :-(. Anyway, coming to the point. So I have 20 eigenvectors and eigenvalues only now. How do I find the % variance captured by them when I discarded other PC's?.

Sounds like a problem :/

Naaaah!!!. Linear Algebra to rescue. If you see, eigen values capture the variance of each Principal Component. If I want to find total variance captured by all eigenvalues, it would be equal to total variance of the dataset. Right?. And in which place can you get total variance in the dataset? It's the Variance-Covariance Matrix.

Ok. How do we get total variance from that?

Variance components in Variance-Covariance Matrix are the diagonal values. You need a sum of these diagonal values for all 12k cells. And what's the sum of diagonal values for a square matrix called in matrix jargon?. It's the ***trace*** of the matrix.

So if you just do trace(Cov(X)), you would get the total variance in the dataset. And then you don't need to have all eigen values to know the variance captured. :-)

Aah.. that's nothing. Don't act smart. This week piece is very lame. Get off!