Pivotal

Manish Tripathi <mtripathi@pivotal.io>

Thursday Thunder: Week 3: Learning 1 & Question 3

Manish Tripathi <mtripathi@pivotal.io> Fri, Oct 23, 2015 at 12:52 PM
To: Pivotal Data Science - All <PDS-ALL@pivotal.io>

Hi All

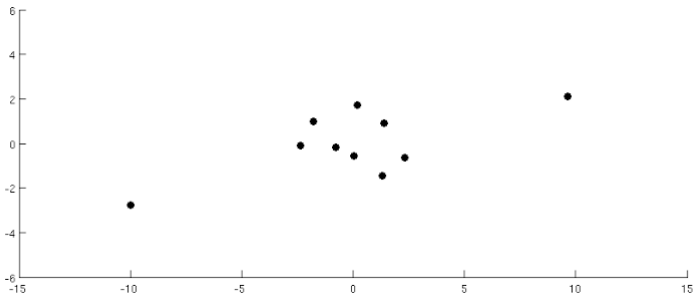
Sorry for getting late on this for two weeks. I have decided that one week I would share some learning or knowledge and one week would pose a question as per Gautam's last email. This week I am keeping things simple, so both learning and question are going to be very easy. Expect tough ones from next week. :)

So here we go:

What is one problem with PCA as a visualization technique:

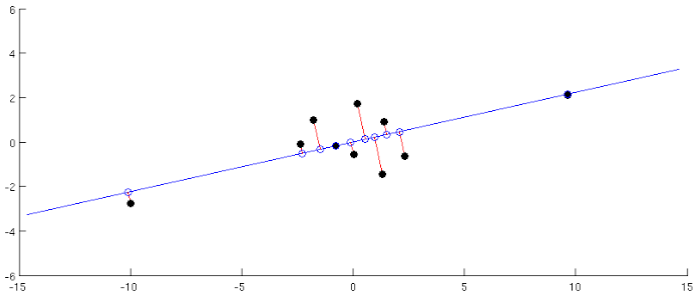
We know to visualize data from high dimension to low dimension we use PCA as one of the technique. But there is one problem with PCA. I*t only preserves large pairwise distances better.* So what does that mean?

Consider a dataset below:



Now if we want to reduce the dimension the best direction is the line cutting the above set of points diagonally from lower left to upper right since this is the eigenvector which has the highest eigenvalue. Aah.. what the hell is this eigen-vector and eigenvalue? We use it a lot in machine learning but what do they stand for?. I would talk about that in a separate post.

So coming back to above points. Below is the first PC loading vector(eigenvector) for the above points.



The red lines are the projection of these points on the loading vector to get the corresponding principal component. Now if you look at the projections. The two extreme points when projected on this loading vector are still about the same distance as they were originally. Even some medium separated points are able to preserve the distance a bit.

But look the points right in the center. In the original 2 dimension space they were pretty different (on each side of the line). But when projected they just become neighbours.

If one does clustering on the projections, these points in the middle which could have been part of different clusters would now become part of same cluster as they are right next to each other. This is a problem with PCA. It only **preserves large pairwise distances** and the final result of PCA would not give so good results for follow-on clustering procedures.

Multi-dimensional scaling helps to some extent with this problem.

Now having talked about PCA a bit, a simple question now. Yeah, I won't spare you guys this week from a question :D

Question:

1). If we are working with say high dimensional data like genomics,gene expression or even on text data with bag of words features, where the usual situation is n<d; where n is the no. of observations and d is no. of dimensions. Lets say 100 observations with 5000 features.

If one does PCA to reduce the dimensionality (which one would have to for any machine learning) and wants to lets say have lesser no. of dimensions than original d dimensions, how many max Principal Components you would get in your eigen decomposition matrix?. And why?.

No googling the answer.

Again, person responding first with the right reason would get a coffee from me :)

Thanks
Manish