The Analytics Edge FALL 2020

PCA concepts

In class exercise: Week 3

1. Indicate which of the following (if any) are true about principal component analysis.

   (a) The output of PCA is a new representation of the data that is always of dimensionality equal to or lower than the original feature representation.

   (b) The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

   Explanation:
   (a) PCA performs a rotation of the data into orthogonal directions and orders them according to direction of maximum variability. The maximum number of such directions is bounded by the dimension of the data p. If less than p directions are chosen or a direction has zero variability (a zero eigenvalue), then the dimension reduces, otherwise it remains the same.
   (b) PCA relates to unsupervised learning, hence, there isn't necessarily an output variable.

2. In principal component analysis, a smaller eigenvalue indicates that

   (a) A given variable in the original data set, say X_j, is more important.

   (b) A given variable in the original data set, say X_j, is less important.

   (c) A given principal component, say Y_j,is more important.

   (d) A given principal component, say Y_j, is less important.

   Explanation: A principal component with higher eigenvalue means the data along that principal component has higher variability. Hence smaller eigenvalue implies smaller variance, meaning less important principal component.

3. Indicate which of the following (if any) are true about principal component analysis.

   (a) The principal components are eigenvectors of the centred data matrix.

   (b) The principal components are eigenvectors of the sample covariance matrix.

   (c) Subsequent principal components are always orthogonal to each other.

   Explanation:
   (a) and (b) The principal components are found by finding eigenvectors of the covariance matrix.
   (c) The eigenvectors of the covariance matrix are orthogonal to each other, and hence so are the principal components.

4. Indicate which of the following (if any) are true about principal component analysis. Assume that no two eigenvectors of the sample covariance matrix have the same eigenvalue.

   (a) Appending a 1 to the end of every sample point doesn't change the results of performing PCA (except that the useful principal component vectors have an extra 0 at the end, and there is one extra useless component with eigenvalue zero).

   (b) If you use PCA to project d-dimensional points down to j principal coordinates, and then you run PCA again to project those j-dimensional coordinates down to k-principal coordinates, with $d > j > k$ you always get the same result as if you had just used PCA to project the d-dimensional points directly down to k-principal coordinates.

   Explanation: Appending an extra dimension with the same values introduces no variance in the extra dimension, so PCA will ignore that dimension. PCA discards the eigenvector directions associated with the largest eigenvalues; as the eigenvectors are mutually orthogonal, this does not affect the variance in the surviving dimensions, so your results depend solely on how many directions you discard.