Dimensionality Reduction

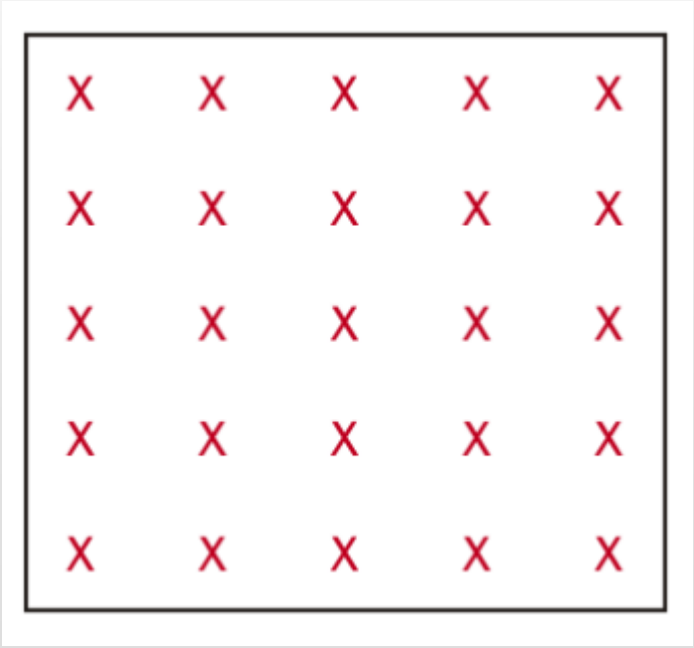
Principal Component Analysis (PCA)

Contents

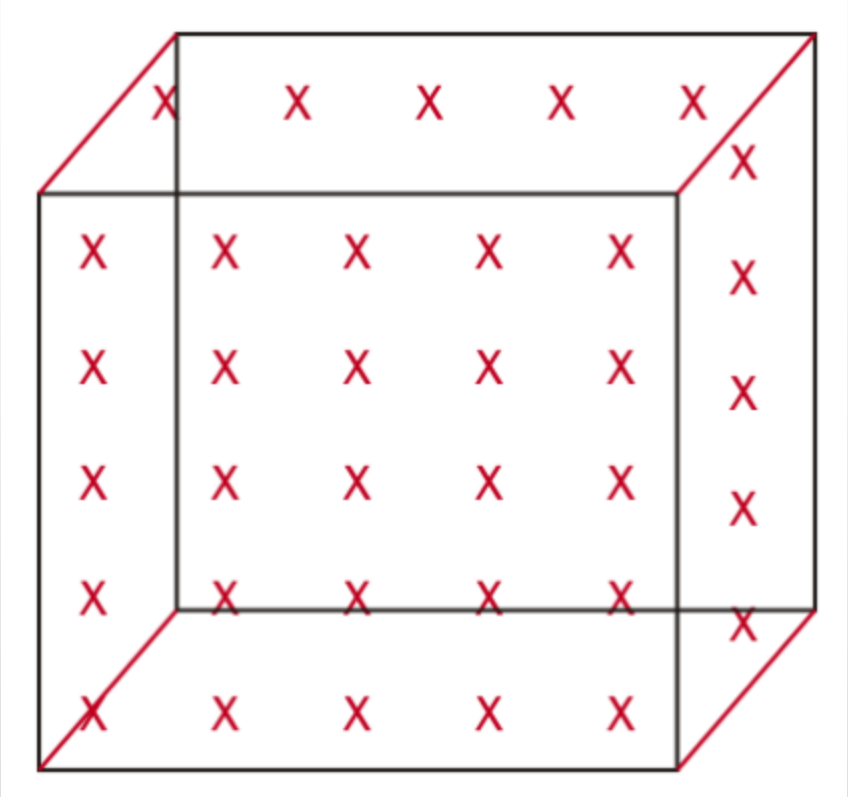
- *Dimensionality Reduction*
  - *Curse of Dimensionality*
- *Principal Component Analysis*
  - *Intution behind PCA*
  - *Uses*
  - *Limitations*
- *Example*

Dimensionality Reduction

- *1-d, 5 data points*



- *2-d, 25 data points*

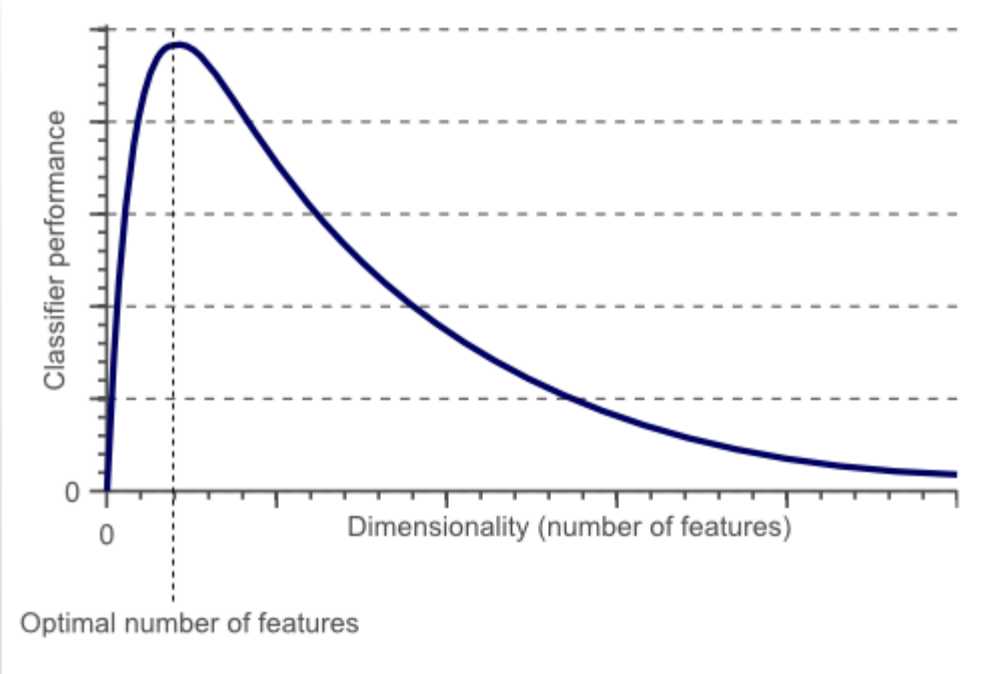


- *3-d, 25 data points*

Dimensionality Reduction

Hughes Phenomenon

- As the number of features increases, the classifier's performance increases as well until we reach the optimal number of features.
- Adding more features based on the same size as the training set will then degrade the classifier's performance.

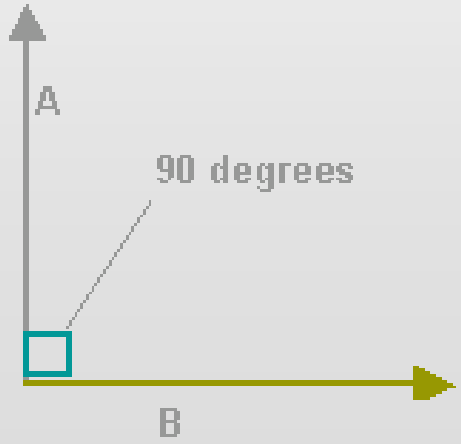


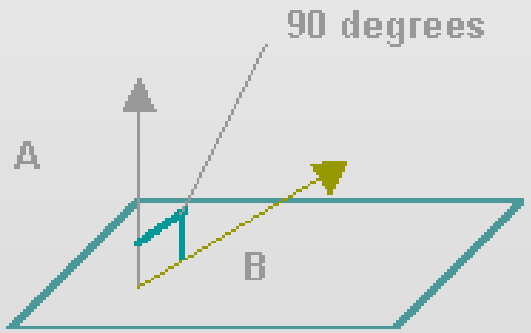
Dimensionality Reduction

- ***Dimensionality*** - in statistics refers to how many attributes a dataset has.
- Need for reduction → '***Curse of dimensionality***'.
- Curse of dimensionality refers to an exponential increase in the size of data caused by a large number of dimensions.
- As the number of dimensions of a data increases, it becomes more and more difficult to process it.
- ***Dimensionality Reduction*** is a solution - reduce the size of data by extracting relevant information and disposing rest of data as noise.

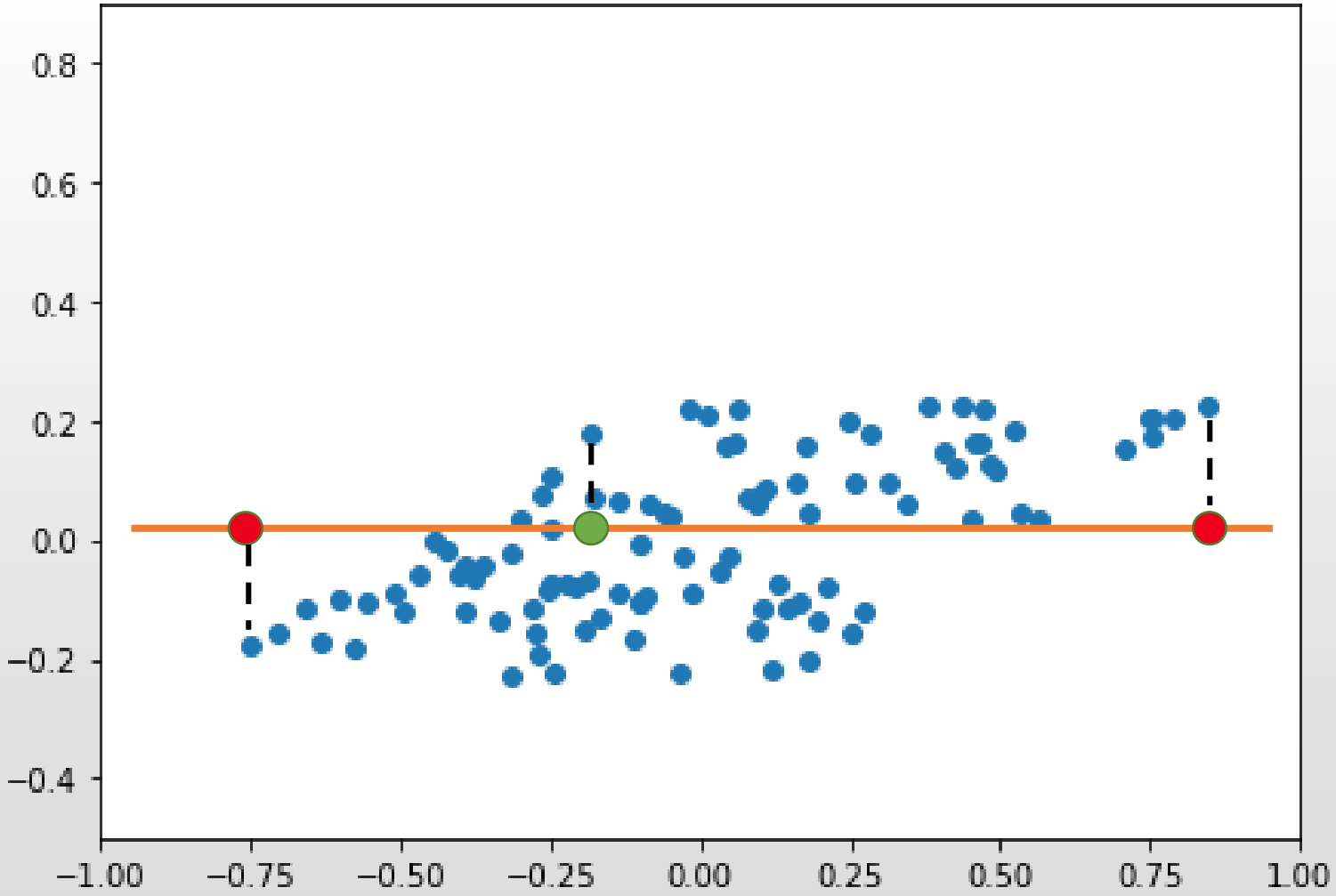
Principal Component Analysis

- PCA is one of the most popular linear dimension reduction.
- It's a projection based method.
- Transforms the data by projecting it onto a set of orthogonal (involving in right angles, perpendicular) axes.
- PCA creates new variables from old ones.





Principal Component Analysis



Principal Component Analysis

- Understanding PCA through animation.

- Each blue dot on the plot represents a point from data given by its x & y coordinate.
- A line P (red line) is drawn from the center of the dataset i.e. from the mean of x & y.
- Every point on the graph is projected on this line shown by two sets of points red & green.
- The spread or variance of data along line p is given by the distance between the two big red points.
- As the line p rotates the distance between the two red points changes according to the angle created by line p with the x-axis.
- The purple lines which join a point and its projection represent the error which arises when we approximate a point by its projection.

Principal Component Analysis

- The approximation error should be small, when the new variables closely approximate the old variables.
- The squared sum of the lengths of all purple lines gives the total error in approximation.
- The angle which minimizes the squared sum of errors also maximizes the distance between the red points.
- The direction of maximum spread is called the ***principal axis***.
- We apply the same procedure to find the next principal axis, which must be orthogonal to the other principal axes.
- Once, we get all the principal axes, the dataset is projected onto these axes. The columns in the projected or transformed dataset are called ***principal components***.

When should you use PCA?

- Reducing the dimensionality of the dataset reduces the size.
- If your learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up.
-

Limitations of PCA

- If the number of variables is large, it becomes hard to interpret the principal components.
- PCA is most suitable when variables have a linear relationship among them.
- PCA is influenced to big outliers.