PCA

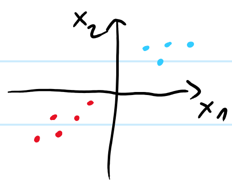
Thursday, April 25, 2024 8:13 PM

PCA - Principal Component Analysis

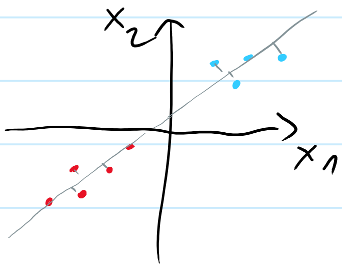
If input has too many dimensions:
- need more data to train
- more computationally difficult
- some ML Algos don't work well in high dimensions

Intuitive solution:

Assume we have data for a classification problem



We can project on a line



Now our new data is only

We got rid of 1 dimension.

<u>Formally</u>:

Instead of inputs $(x_1, \dots, x_D)$, use $(y_1, \dots, y_k)$, where $y_i = \sum a_j x_j + b_i$

(new input feature is just a linear combination of the original ones.

(linear with bias).

(linear with bias).

y_i's are called principal components

y_1 such that projecting on y_1 gives maximal variance.
To get y_2, project onto $\langle y_1 \rangle^\perp$ and again, take y_2 in $\langle y_1 \rangle^\perp$ that maximizes Variance.

Algorithm

Step 1 Standardizing

transform each input $x \in \mathbb{R}^D$

$x \rightsquigarrow \frac{x-\mu}{\sigma}$, where μ is the mean, σ the standard deviation (componentwise)

Step 2 Let $A := \left(Cov(x_i, x_j)\right)_{i,j \in \{1,..,D\}^2}$

Where $Cov(x_i,x_j) = \frac{1}{N-1}\sum_{\ell=1}^{N}(x_i^\ell - \overline{x}_i)(x_j^\ell - \overline{x}_j)$.

One can show, that the first k principal components are the k eigenvectors corresponding to the k largest eigenvalues of A (without abs. val.).