PCA Algorithm

Frederik Mallmann-Trenn
6CCS3AIN

PCA Algorithm - Data

- Let's say this is our data matrix (say our houses), where each data point is an d-dimensional row vector.
-
$$\mathbf{X} = \begin{pmatrix} — & \mathbf{x}_1^T & — \\ — & \mathbf{x}_2^T & — \\ & \vdots & \\ — & \mathbf{x}_n^T & — \end{pmatrix}$$

  The dimensions are $n \times d$

PCA Algorithm

- Step 1: Compute the mean row vector $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Step 2: Compute the mean row matrix $\bar{\mathbf{X}} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \bar{\mathbf{x}}^T = \begin{pmatrix} — & \bar{\mathbf{x}}^T & — \\ — & \bar{\mathbf{x}}^T & — \\ & \vdots & \\ — & \bar{\mathbf{x}}^T & — \end{pmatrix}$

  The dimensions are $n \times d$

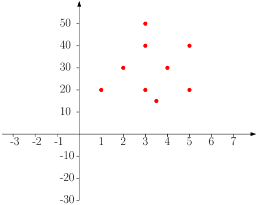
PCA Algorithm

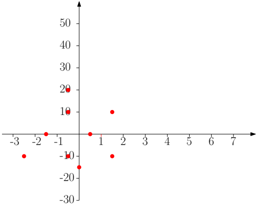
- Step 3: Subtract mean (obtain mean centred data)

$$B = \mathbf{X} - \bar{\mathbf{X}}$$

  The dimensions are $n \times d$

- Example:





PCA Algorithm

- Step 4: Compute the covariance matrix of rows of B

$$\mathbf{C} = \mathbf{B}^T\mathbf{B}$$

The dimensions are $(n \times d)^T \times (n \times d) = (d \times n) \times (n \times d) = d \times d$

PCA Algorithm

- Step 5: Compute the k largest eigenvectors $\mathbf{v}_1, \mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{C}$ (not covered how to do this in this module. You use Python or WolframAlpha).
  Each eigenvector has dimensions $1 \times d$

  Pro tip: Python doesn't sort the eigenvectors for you. Sort eigenvectors by decreasing order of eigenvalues.

- Step 6: Compute matrix $\mathbf{W}$ of k-largest eigenvectors

$$\mathbf{W} = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_k \\ | & | & & | \end{pmatrix}$$

  Dimensions of $\mathbf{W}$ are $(d \times k)$.

PCA Algorithm

- Step 7: Multiply each datapoint $\mathbf{x}_i$ for $i \in \{1, 2, \ldots, n\}$ with $\mathbf{W}^T$

$$\mathbf{y}_i = \mathbf{W}^T \cdot \mathbf{x}_i$$

Dimensions of $\mathbf{y}_i$ are $(k \times d) \times (d \times 1) = k \times 1$

PCA Algorithm

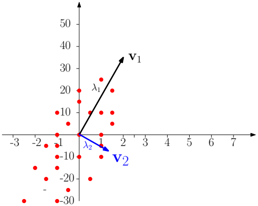
- Step 7: Multiply each datapoint $\mathbf{x}_i$ for $i \in \{1, 2, \ldots, n\}$ with $\mathbf{W}^T$

$$\mathbf{y}_i = \mathbf{W}^T \cdot \mathbf{x}_i$$

  Dimensions of $\mathbf{y}_i$ are $(k \times d) \times (d \times 1) = k \times 1$

- Congratulations! You've reduced the number of dimensions from d to k!

Why do we compute the covariance matrix?

- Example illustration: 

- The covariance matrix measures the correlation between pairs of features.
- Finding the largest eigenvectors allows us to explain most of the variance in data
- The more variance is explained by the eigenvectors, the more important they are

Why do we compute the covariance matrix?

- We can measure the explained variance by considering the quantity

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

- Example: