Lecture 9 Mar 21, 2021

Unsupervised Learning

3 Aspects:

1) Density estimation
 - image restoration
 - anomaly detection

2) Clustering
 - group data that is similar

3) Dimensionality reduction
 - remove redundant information
 - less parameters
 - deal's with curse of dimensionality
 - feature selection
 - visualization
 - computationally cheaper

Dimensionality Reduction //

Given observation y, find a low representation of y x and the mapping $y = f(x)$.

Principal Component Analysis //

Essentially assuming f is linear.

Given data $\{y_i\}_{i=1}^{N}$, $y_i \in \mathbb{R}^d$, we want to learn $W \in \mathbb{R}^{d \times c}$ and $x_i, b \in \mathbb{R}^c$ such that $y_i \approx W x_i + b$. $c << d$.

Orthagonality

Suppose w^* is a solution for $y_i = w^* x_i$

We can also have $y_i = (a w^*) \frac{x_i}{a}$, so we can have infinite solutions $(a w^*)$

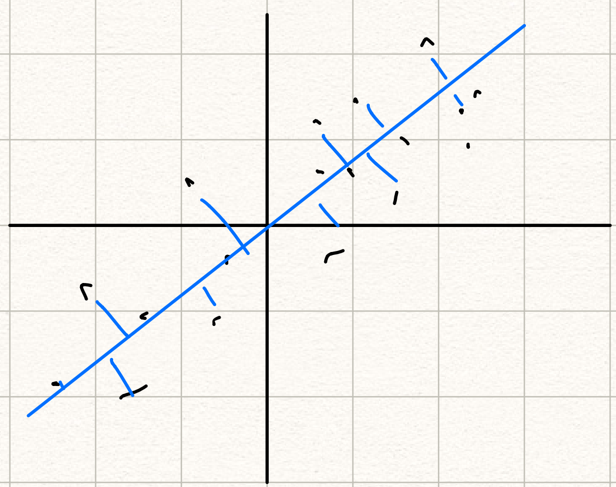
To counteract this (find unique W), we find orthonormal W. Thus $W = \begin{bmatrix} | & & | \\ w_1 & \cdots & w_c \\ | & & | \end{bmatrix}$ s.t. $\|w_i\|^2 = 1$ $\forall i \in \{1, \ldots, c\}$

With orthonormal W, not only is it unique, but the orthagonality of column vectors restricts the components from sharing information.

Zero Mean

Assume $\{y_i\}_{i=1}^{N}$ is zero mean. If mean not zero, center it to zero mean.

Visual of PCA



Projects data onto hyperplane.

Objective Func

$$E(w) = \sum_{i}^{N} \min_{x_i} \|y_i - W x_i\|^2$$

Find W for minimized E.

1) Given W, we have regression problem to find optimal x_is that minimize $E(w)$.

$x^* = (W'W)^{-1}W'y$ which is the LS solution.
$= W'y$ since w is orthonormal

2) Find W that minimizes $E(W)$.

a) suppose y is already on a line, then residual error $= 0$

b) suppose y has noise, now find W to minimize projection error.

We fit a guassian elipsoid around the data. Then W is projection of the c major axis with the highest variance.

$$\text{Elipsoid}: \vec{y}^T K^{-1} \vec{y} = 1$$

$$K = \frac{1}{N} \sum y_i y_i^T$$

The data is projeted on a $\mathbb{R}^c$ system consisting of the c major axis.

To find which axis have the highest variance, define K, find eigenvectors/values.

$$\text{Let } U = [u_1 \cdots u_d]$$
$$S = \text{diag}(\lambda_1, \ldots, \lambda_d)$$
$$\text{where } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Take the first c axis of V as major axis.

Summary Algorithm//

1) $b = \frac{1}{N} \sum y_i$ calculate mean
2) $\hat{y}_i = y_i - b$ center

3) $K = \frac{1}{N} \sum y_i y_i^T$ find K

4) Compute U, S, eigenvectors/values of K

5) Assume eigenvalues are sorted desc in S

6) Let $W = [u_1, \dots u_c]$

7) Let $x_i = W^T \bar{y}_i$

For reconstruction

$$y_i = Wx_i + b$$

Properties //

1) Zero mean

2) Cov of x_i is diagonal with eigenvalues in S

3) The total variance of data is

$$tr(S) = \sum_i^d \lambda_i$$

The total subspace variance is

$$\sum_i^c \lambda_i$$

The total out-of-subspace variance is

$$\sum_{i=c+1}^d \lambda_i$$

Choose c that captures 95% of total var.