Lecture Notes on ML.

付星宇. 中山大学数学学院

2018.1.31. 广东·广州

Section 2: Principal Components Analysis.

When faced with a large set of correlated predictor variables, PCA allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original dataset. PCA is an unsupervised approach, since it involves only a set of features (X_1 X_2 $\cdots$ X_p) and no associated response Y.

Say we have a following dataset:

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{pmatrix} = X.$$

where each feature's mean is normalized to zero.

we give each observation a new feature, say:

$$z_{i1} = \sum_{j=1}^{p} \phi_{j1} X_{ij}, \quad \text{where} \quad \sum_{j=1}^{p} |\phi_{j1}|^2 = 1$$

we want to maximize the variance of (Z_{11} Z_{21} $\cdots$ Z_{n1})

i.e. $$\max_{\phi_1 \in R^{1\times p}} \frac{1}{n} \sum_{i=1}^{n} (z_{i1} - 0)^2, \quad \text{where} \quad \|\phi_1\|_2 = 1.$$
$(\phi_1 = (\phi_{11} \vdots)^T)$.

while :

$$\frac{1}{n}\sum_{i=1}^{n}(z_{i1})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\phi_1 \cdot X_i^T\right)^2 \quad (X_i = (X_{i1}\ X_{i2} \cdots X_{ip}))$$

$$= \frac{1}{n}\sum_{i=1}^{n} \phi_1 \cdot (X_i^T \cdot X_i)\, \phi_1^T$$

$$= \phi_1 \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T \cdot X_i\right) \phi_1^T$$

we define $\left(\Sigma \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)$.

Our problem now converted into :

$$\max_{\substack{\phi \in R^{1\times p} \\ \|\phi_1\|=1}} Q(\phi_1) = \max_{\substack{\phi_1 \in R^{1\times p} \\ \|\phi_1\|=1}} \phi_1 \Sigma \phi_1^T$$

use Lagrange Multiplier Method, it's easy to show that $Q(\phi_1)$ is maximized if ϕ_1 is the principal eigen-vector of Σ ~~and~~ and $\max Q = \max \lambda_i$ where λ_i is the eigen-value of Σ. (HW.)

After Finding the way to construct a highly variant feature for each observation, a natural question is how to construct another variant feature?

We just need to choose top K eigenvectors of Σ

Then, each observation is mapped into a k-dimensional feature space where we can choose $k \ll p$.

Hence PCA is referred as a dimension reduction algorithm.

- Application 1.

If we reduce the dimension of observation to $k=2$ or 3, then we can visualize the data.

- Application 2.

Preprocess dataset to reduce its dimension before running a supervised Learning algorithm. Apart from computational benefits, reducing dimension can reduce the complexity of the superivised model and therefore ~~hold~~ help avoid over-fitting.

- Application 3.

$$\underset{\substack{\uparrow \\ \text{Face image}}}{x_i} \xrightarrow{PCA} y_i \quad \text{then similar}(x_i, x_j) \triangleq \| y_i - y_j \|_2$$

it turns out that this is a surprisingly good face matching algorithm..