Principal Component Analysis (PCA)

Note Title 6/22/2021

Feature selection: Start with as many features as you can collect, and then find a good subset of features

PCA (Main idea):
Project given data onto a lower dimensional subspace such that
(i) Reconstruction error is minimized
(ii) Variance of the projected data is maximized

Problem formulation:

Given: Dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$

Goal: Project $\mathcal{D}$ onto a m-dimensional subspace "m: input parameter"

Arriving at the PCA algorithm!

Let $\mathcal{B} = \{u_1, \ldots, u_m\}$ be a orthonormal basis for a m-dimensional subspace.

↳ Fix a subspace & then find the best projection

→ Later, we will understand how to choose the subspace $\mathcal{B}$ optimally.

Extend $\mathcal{B}$ to an orthonormal basis for $\mathbb{R}^d$. Let this extended basis be $\underbrace{\{u_1, \dots u_m, u_{m+1}, \dots u_d\}}$

Denote this as $\mathcal{B}'$

Any vector $x \in \mathbb{R}^d$ can be written using $\mathcal{B}'$ as follows:

$$x = \alpha_1 u_1 + \dots + \alpha_d u_d \text{, where } \alpha_j = x^T u_j \text{ for } j = 1, \dots, d.$$

Expressing each data point x_i in $\mathcal{B} = \{x_1 \dots x_n\}$ using $\mathcal{B}'$, we have

$$x_i = \sum_{j=1}^{d} (x_i^T u_j) u_j$$ ← original data point

Approximate x_i by $\tilde{x}_i$ as follows:

$$\tilde{x}_i = \sum_{j=1}^{m} z_{ij} u_j + \sum_{j=m+1}^{d} \beta_j u_j$$ ← projected data point (i.e., belonging to a m-dimensional subspace)

Next step: Find optimal z_{ij}, β_j to minimize square error:

Want to minimize J over z_{ij}, β_j →

$$J = \frac{1}{n} \sum_{i=1}^{n} \| x_i - \tilde{x}_i \|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \Big\| \underbrace{\sum_{j=1}^{m} (x_i^T u_j - z_{ij}) u_j}_{(I)} + \underbrace{\sum_{j=m+1}^{d} (x_i^T u_j - \beta_j) u_j}_{(II)} \Big\|^2 \quad \text{— } (*)$$

Why? $= \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_i^T u_j - z_{ij})^2 + \sum_{j=m+1}^{d} (x_i^T u_j - \beta_j)^2 \right]$ —— (**)

A simple case

$$\|c_1 u_1 + c_2 u_2\|^2 = c_1^2 \|u_1\|^2 + c_2^2 \|u_2\|^2 + 2 c_1 c_2 u_1^T u_2$$
$$= c_1^2 + c_2^2$$

if $\{u_1, u_2\}$ is orthonormal then $\|u_i\|^2 = 1$, $i = 1, 2$
$u_1^T u_2 = 0$

What you have in (*) is $\|(c_1 u_1 + \cdots + c_m u_m) + (c_{m+1} u_{m+1} + \cdots + c_d u_d)\|^2$

$= c_1^2 + \cdots + c_m^2 + c_{m+1}^2 + \cdots + c_d^2$

→ then in the form of (**)

So, $J = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_i^T u_j - z_{ij})^2 + \sum_{j=m+1}^{d} (x_i^T u_j - \beta_j)^2 \right]$

To minimize J over z_{ij}, β_j, we find the partial derivatives & equate them to zero

$$\frac{\partial J}{\partial z_{ij}} = 0 \Rightarrow 2(x_i^T u_j - z_{ij}) = 0 \Rightarrow \boxed{z_{ij} = x_i^T u_j}$$

$$\frac{\partial J}{\partial \beta_j} = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(x_i^T u_j - \beta_j) = 0 \Rightarrow \boxed{\beta_j = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^T u_j = \bar{x}^T u_j,}$$

where $\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$

So, for a given m-dimensional subspace spanned by $\mathcal{B} = \{u_1, \dots, u_m\}$, the best projected data is

$$\boxed{\tilde{x}_i = \sum_{j=1}^{m}(x_i^T u_j)u_j + \sum_{j=m+1}^{d}(\bar{x}^T u_j)u_j}$$

If the data be already centered, then $\bar{x} = 0$ & the second term vanishes.

From the foregoing,

$$x_i - \tilde{x}_i = \sum_{j=m+1}^{d}(x_i^T u_j - \bar{x}^T u_j)u_j$$

$$\underline{\|x_i - \tilde{x}_i\|^2 = \sum_{j=m+1}^{d}(x_i^T u_j - \bar{x}^T u_j)^2} = \sum_{j=m+1}^{d}\left((x_i - \bar{x})^T u_j\right)^2$$

With optimal z_{ij}, β_j, the square error becomes

$$J^* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=m+1}^{d} \left((x_i - \bar{x})^T u_j \right)^2$$

$$= \frac{1}{n} \sum_{j=m+1}^{d} \sum_{i=1}^{n} \left((x_i - \bar{x})^T u_j \right)^T \left((x_i - \bar{x})^T u_j \right)$$

$$= \frac{1}{n} \sum_{j=m+1}^{d} \sum_{i=1}^{n} u_j^T (x_i - \bar{x})(x_i - \bar{x})^T u_j$$

$$= \sum_{j=m+1}^{d} u_j^T \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right] u_j$$

$$\boxed{J^* = \sum_{j=m+1}^{d} u_j^T C u_j}, \text{ where } C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$