PCA 推导

Soul Walker

2020 年 7 月 22 日

给定数据矩阵 $\mathbf{X} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n) \in R^{n\times m}$，内含有 n 个样本，特征维度为 m, 拟降至维度 d, 此外，样本均值为 $\boldsymbol{\mu} \in R^m$, 可在减维前保证样本已经去中心化 ($\boldsymbol{\mu} = 0$).

0.1 最大方差

0.1.1 数据变换

构建变换函数, 针对任意样本 $\boldsymbol{x}_i$, 定义变换函数 $f(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{x}_i$, 其中 $\boldsymbol{w} \in R^n$ 为变换系数.

0.1.2 方差最大化

变换后的方差计算为

$$
\begin{aligned}
var(f(\boldsymbol{x})) &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{w}^T\boldsymbol{\mu})^2 \qquad (0.1)\\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i)^2 \quad \boldsymbol{\mu} = 0\\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{w})\\
&= \frac{1}{n}\boldsymbol{w}^T\sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{w}\\
&= \frac{1}{n}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}
\end{aligned}
$$

0.1.3 拉格朗日求解

通过以上对方差的计算，可以得到目标函数为

$$\max_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{w} \quad s.t. \quad \boldsymbol{w}^T \boldsymbol{w} = 1 \tag{0.2}$$

为了解上式我们构建拉格朗日函数如下

$$L(\boldsymbol{w}, \lambda) = \frac{1}{n} \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{w} - \lambda (\boldsymbol{w}^T \boldsymbol{w} - 1) \tag{0.3}$$

对拉格朗日函数对 $\boldsymbol{w}$ 求偏导如下

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - 2\lambda \boldsymbol{w} \tag{0.4}$$

对于求最大值，令上式为 0 得

$$\begin{aligned} 2\boldsymbol{X}^T \boldsymbol{X} w - 2\lambda \boldsymbol{w} = 0 \\ \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} = \lambda \boldsymbol{w} \end{aligned} \tag{0.5}$$

此处 λ 是特征值，$\boldsymbol{w}$ 是特征向量，那现在就是寻找最大特征值。由式 (1.1.5) 可以推广如下，对于每一个 $\boldsymbol{w}_i$（第 i 维）均有

$$\Sigma \boldsymbol{w}_i = \lambda \boldsymbol{w}_i \tag{0.6}$$

那么我们就对特征值可以由大到小排序，取最大的前 d 维。