主成分分析教程

2021 年 2 月 3 日

1 随机变量与其数字特征

1.1 概率三元组

有了概率三元组 $(\Omega, \mathcal{F}, \mathcal{P})$ 我们就可以定义概率空间，这里 Ω 是样本空间，也就是所有基本得不能再基本的事件，而 $\mathcal{F}$ 是事件空间，一个事件是一个 Ω 的子集，所以，注意 $\mathcal{F}$ 是元素的集合，也就是说 $\mathcal{F}$ 是集合的集合，$\mathcal{P}$ 是一个函数，也叫概率函数或者叫分布，它就是一个 $\mathcal{F}$ 到 $\mathcal{P}$ 的对应法则.

一枚骰子可否用一个概率空间来描述？可以的．骰子有六个面，那么样本空间可以定为

$$\Omega = \{1 \text{ 朝上}, 2 \text{ 朝上}, \cdots, 6 \text{ 朝上}\} \tag{1.1}$$

事件空间 $\mathcal{F}$ 呢，其实就是 Ω 所有的子集，这里我们写不下．概率嘛，它不是人为规定的，它取决于骰子，也就是说它跟骰子的材质有关，不过我们一般都默认掷出每个面的概率是「接近相等的」.

1.2 随机变量

概率三元组可以用来对自然界的随机现象建模，同时它还蕴含了一个分布．随机变量 X 假如说它服从某某分布，那么它就是那个分布所处在的概率空间下的样本空间 Ω 到实数集 $\mathbb{R}$ 的一个映射．也就是说，随机变量把普通话和数字对应起来，怎么对应呢？是这样子的:

$$X = \begin{cases} 1, & 1 \text{ 号面朝上}; \\ 2, & 2 \text{ 号面朝上}; \\ 3, & 3 \text{ 号面朝上}; \\ 4, & 4 \text{ 号面朝上}; \\ 5, & 5 \text{ 号面朝上}; \\ 6, & 6 \text{ 号面朝上}. \end{cases} \tag{1.2}$$

从而我们可以把自然现象当做「数」来处理.

1.3 数字特征

1.3.1 均值

我们就拿离散型随机变量的均值来举例:

$$\mathrm{E}[X] = \sum_{x \in \mathcal{F}} X(x) P(x) \tag{1.3}$$

考虑个具体的例子：骰子掷出 1,2,3,4,5,6 各个面朝上的概率分别都是 1/6，我用 X 表示在一次掷骰子试验中，朝上的那一面的点数，求 X 的数学期望（均值）：

$$\mathrm{E}[X] = 1 \cdot 1/6 + 2 \cdot 1/6 + \cdots + 6 \cdot 1/6 = 3.5 \tag{1.4}$$

请读者自行悟出连续型随机变量的均值表示方法.

当我对一个系统进行观察，在有限长的一段时间内，我记录了 N 次它的状态，或者也可以说，做了 N 次试验，那么我得到一组数字 $x_1, x_2, \cdots, x_N$，这些数字称为样本，我们可以假设这些数字的概率性质，比如说说它们服从某某分布，当然前提是我已经找好了一个随机变量把口头描述的事件转化为实数，然后我也可以计算这些样本，或者不如说这个随机变量的均值（数学期望）：

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1.5}$$

就是这么简单．它们也叫做 X 的样本均值.

1.3.2 方差

确定了要研究的对象，再找到了一个随机变量把从研究对象身上得到的试验结果映射到实数集上，那么接下来的事情就纯粹是数学的了，或者干脆说，就可以用这个随机变量 X 来在语境上下文中直接替换原来的具体的研究对象了，比如说用 X 表示掷一次骰子朝上的面的点数，那么以后我们就不说研究骰子了，我们只说研究随机变量 X，这是一回事．既然用 X 指代骰子，我们可以把 X 看成是一个能不断蹦出随机数的随机数生成器来，那么它蹦出来的这些随机数就会有一个离散的程度的度量．比如说 X_1 和 X_2 是两个不同的随机变量（它们不一定是上面我们说的骰子），那么 X_1 的样本可能是

$$1, 100, 1000, 10000, 100000, \cdots \tag{1.6}$$

而 X_2 的样本可能是

$$1, 2, 3, 4, 5, 6, 7, \cdots \tag{1.7}$$

显然前者散得更开一些，后者更密，所以前者方差就大，而具体地，如果是用期望来定义方差，就是

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \tag{1.8}$$

如果我们已经知道了 $\mathrm{E}[X^2]$ 以及 $\mathrm{E}[X]$，那么方差也可以这样算：

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \tag{1.9}$$

如果我们知道了从 X 蹦出的 N 个数字（也叫样本），那么还可以这样算：

$$\hat{\mathrm{Var}}[X] = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{1.10}$$

式 1.10的计算结果只是 $\mathrm{Var}[X]$ 的一个估计，N 越大，估计得越准.

1.3.3 协方差

我们知道，身高与体重有一定的正相关关系，收入与支出也有一定的正的相关关系，年龄与健康程度又一定的相关关系，虽然协方差并不能够准确描述这种相关关系的程度的，但是它的符号可以说明一些问题，用随机变量 X, Y 分别表示身高与体重，那么协方差 $\mathrm{cov}[X, Y]$ 的符号就是正的，用 X, Y 分别表示收入和支出，那么协方差 $\mathrm{cov}[X, Y]$ 的符号则是正的，用 X, Y 表示年龄与健康程度，那 $\mathrm{cov}[X, Y]$ 就是负的符号．而协方差的大小还受到度量影响.

用均值来表示协方差:

$$\operatorname{cov}[X,Y] = \operatorname{E}[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])] \tag{1.11}$$

也可以用样本来估计协方差:

$$\hat{\operatorname{cov}}[x,y] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}). \tag{1.12}$$

2 主成分分析

对于一个 $n \times m$ 资料矩阵 X，我们假设它的每一行对应一条数据，每一列对应一个属性，并且假设 X 的每个元素都是连续型的，那么我们想找到这样一个线性算子 P：通过对 X 的每一个数据点施加这个线性算子，使得得到的 $n \times m$ 资料矩阵 Y 具备这样的性质：任意 Y 中的两个变量 $\boldsymbol{y}_i, \boldsymbol{y}_j$ 的协方差满足

$$\operatorname{cov}\left[\boldsymbol{y}_i, \boldsymbol{y}_j\right] = \begin{cases} 0, & i \neq j; \\ \operatorname{var}\left[\boldsymbol{y}_i\right], & i = j. \end{cases} \tag{2.1}$$

注意观察式 1.12所描述的样本协方差的计算方法，和矩阵的乘法联系起来，我们有：对任意 $1 \leq i \leq n$，对任意 $1 \leq j \leq m$，都有

$$X_c^{\mathsf{T}} X_c[i,j] = \operatorname{cov}[\boldsymbol{x}_i, \boldsymbol{x}_j] \tag{2.2}$$

这里 X_c 可以是任意的一个 $n \times m$ 的经过中心化的资料矩阵（中心化就是让每一列减去它的均值），并且假设它的每一行对应一个数据点（从而每一列对应一个变量）. 那么这个问题变为了：找到矩阵 P，令

$$Y_c = X_c P \tag{2.3}$$

并且使得

$$Y_c^{\mathsf{T}} Y_c = P^{\mathsf{T}} X_c^{\mathsf{T}} X_c P = \operatorname{diag}\{\lambda_1, \cdots, \lambda_r\} \tag{2.4}$$

其中：$r = \min\{n, m\}$.

首先，由于 $X_c^{\mathsf{T}} X_c$ 是一个实对称矩阵，所以，一定可以对它做特征值分解，也就是说，一定存在正交矩阵 Q，对角矩阵 D，使得

$$X_c^{\mathsf{T}} X_c = Q\, D\, Q^{\mathsf{T}} \tag{2.5}$$

再观察式 2.4，我们发现：当 P 取 Q 时，有

$$Y_c^{\mathsf{T}} Y_c = Q^{\mathsf{T}} Q D Q^{\mathsf{T}} Q = D \tag{2.6}$$

所以说 P 就是 Q，而 Q 就是 $X^{\mathsf{T}} X$ 的特征向量作为列向量组成的矩阵. 值得一提的是，通过对调 Q 中列向量的次序，我们总是可以使得对角矩阵 D 的对角元依次递减.

2.1 上机实验

我们使用 NumPy 和 SciPy 来演示一遍主成分分析.

首先生成样本数据:

```
1 import numpy as np
2 from scipy.stats import multivariate_normal
3 from scipy.stats import norm
4
5 n_samples = 1000
6 n_vars = 2
```

```
mat_x = np.zeros((n_samples, n_vars,), dtype=np.float)
mat_x[:, 0] = norm.rvs(loc=5, size=1000)
y_means = 1.2 * mat_x[:, 0] + 1
mat_x[:, 1] = multivariate_normal.rvs(
    mean = y_means
)
```

我们生成的是一个有 1000 行数据，2 个变量的数据集（图 2-1）

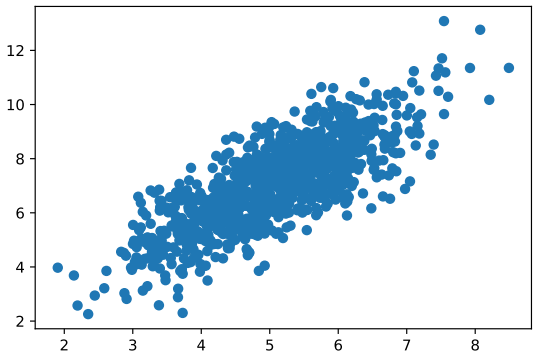

图 2-1: 随机生成的数据点

接下来我们对它中心化，也就是算出那个 X_c:

```
mat_x_c = mat_x - np.mean(mat_x, axis=0).reshape(1, n_vars)
```

接下来就是求协方差矩阵了:

```
cov_mat = (mat_x_c.T @ mat_x_c) / (n_samples-1)

# 还可以这样求：
# cov_mat = np.cov(mat_x.T)
# 这里 mat_x 的每一行对应一个样本，所以用了转置
```

接下来做特征值分解:

```
from scipy import linalg
w, vl, vr = linalg.eig(cov_mat, left=True, right=True)
w = np.real(w)

# 作为验证，读者可以运行：
# print((cov_mat @ vl) - (vl @ np.diag(w)))
# print(cov_mat - ( vl @ np.diag(w) @ vl.T ))
# 并观察输出结果
```

所以这里算出的 `vl` 就是那个矩阵 Q, 它的每一列就是 `cov_mat` 的一个特征向量, 而 `np.diag(w)` 则是那个对角矩阵 D. 现在我们令 $P = Q$，看是否满足条件:

```
mat_p = vl
mat_y_c = mat_x_c @ mat_p
print(np.cov(mat_y_c.T))
```

将会看到，非对角元的元素的大小都非常接近 0. 我们还发现:

```
lhs = np.sum(np.diag(np.cov(mat_y_c.T)))
rhs = np.sum(np.diag(np.cov(mat_x_c.T)))
print(lhs - rhs)
```

非常接近零，这也进一步地验证了计算结果的正确性.

我们可以通过将 Y_c 的每一个样本点画出来，看这样使得 Y_c 的变量两两不线性先关的线性变换 P 到底对 X_c 的每一个样本点都做了些什么：

```
plt.scatter(mat_y_c[:, 0], mat_y_c[:, 1])
```

输出见图 2-2：

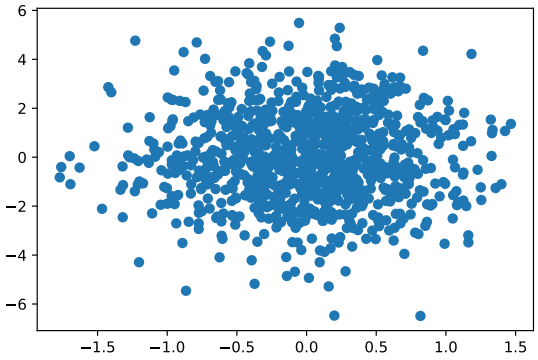

图 2-2: 对 X_c 每个样本点施加线性变换 P 得到的

那么我们就能够理解：其实主成分分析就是把样本的椭圆轮廓做了一个旋转，让长轴短轴分别与坐标轴平行.

2.2 与奇异值分解的关系

对任何一个数域 K 上的 $n \times m$ 矩阵 X，都存在酉矩阵 (unitary matrix)U，对角矩阵 Σ，酉矩阵 V，使得

$$X = U \Sigma V^{\star} \tag{2.7}$$

如果 V 是实的那么 $V^{\star}$ 相当于 V^{T}. 现在，根据酉矩阵的性质：自身与自身的共轭转置的乘积为单位阵，我们得

$$\begin{aligned} X^{\mathsf{T}} X &= \left(U \Sigma V^{\mathsf{T}}\right)^{\mathsf{T}} \left(U \Sigma V^{\mathsf{T}}\right) \\ &= V \Sigma U^{\mathsf{T}} U \Sigma V^{\mathsf{T}} \\ &= V \Sigma^2 V^{\mathsf{T}} \end{aligned} \tag{2.8}$$

所以说通过奇异值分解得到的 Σ^2 就可以当做 D，而 V 就可以当做 Q.