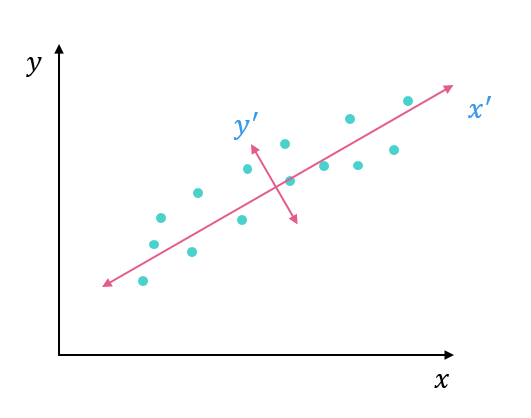
主成分分析与回归

1. PCA

PCA是主成分分析的缩写。假设一个有m列属性的数据，切某些列的属性具有相关性，其每一个属性对应一个坐标轴，那么每一行数据可以对应坐标轴上的一个点。PCA的作用在于重新计算坐标系，然后计算这些点在新坐标系上的坐标。新的坐标系下，数据点在某些坐标轴方向上的分布范围相对其他轴方向会小很多。这样，可以分析出主要的属性，忽略次要属性，达到降维和去除属性相关性的效果。

如下图所示，有坐标轴x、y构成的坐标系中，数据主要分布在45度方向，呈现一定的相关性。如果用x’、y’来建立新的坐标系，可以观察到，数据在x’方向上分布较广，在y’方向上分布范围很小。新坐标系下，y’坐标对应的属性值可以忽略，从而达到降维或者属性去耦合的目的。



从以上描述可以看出，PCA的主要工作在于计算新的坐标系，以及新坐标系对应的坐标。下面介绍PCA算法的主要步骤。