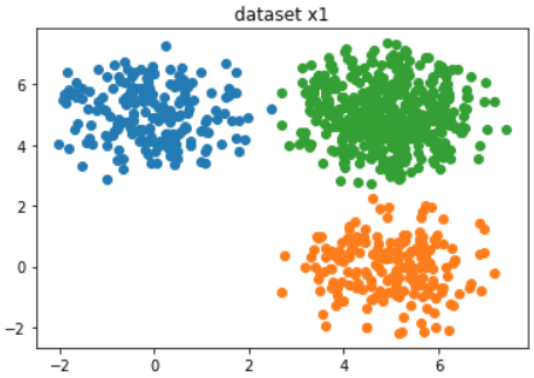
Ex5 record and report

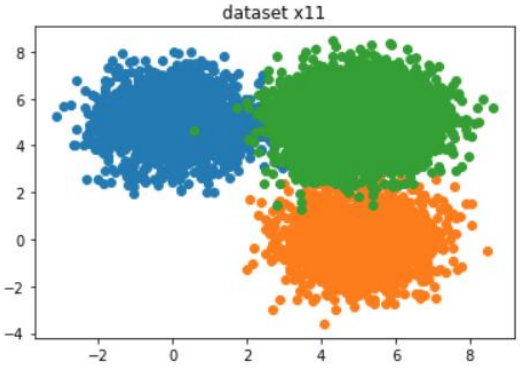
王敏行 id：2018012386 wangmx18@mails.tsinghua.edu.cn

Part 12-1: data generation

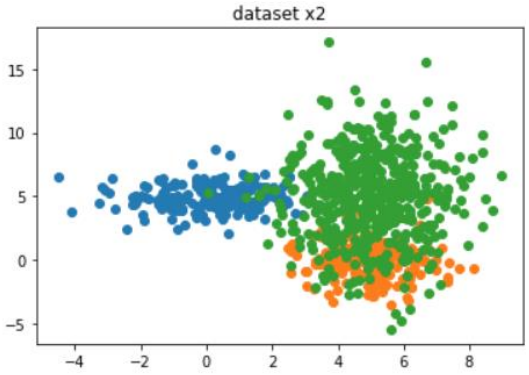
这里利用 numpy.random.multivariate_normal 函数，新写一个可以排除多维高斯分布中心一定距离之外的生成二维高斯分布的函数 make_cluster。生成题目要求中的数据，如下所示。每个种类的数据集产生少量数据和大量数据的两个数据集。

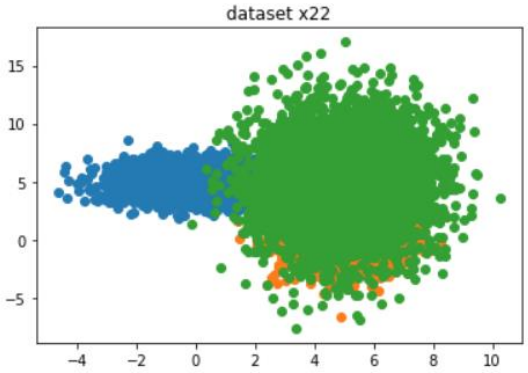
清晰分界的 3 类：



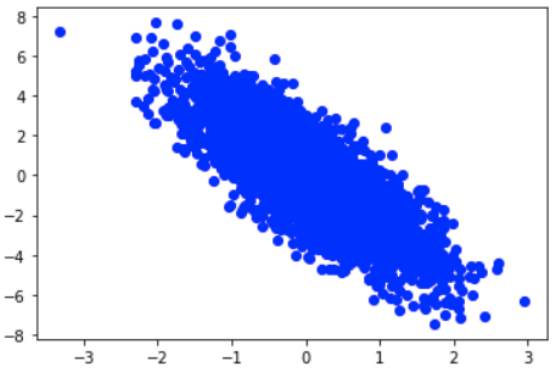


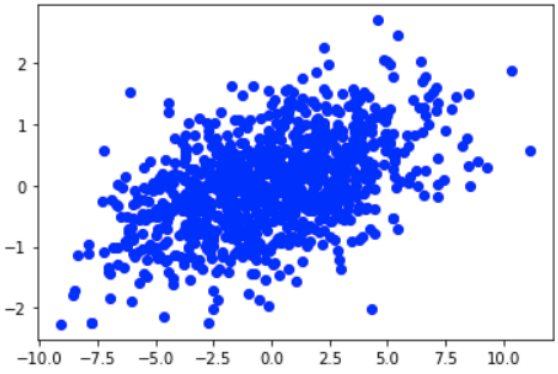
重度重合的 3 类：





一个成分的多维高斯分布：





Part 12-2: K-means clustering

很难看出重度重合的分布数据是分为 3 类的，肉眼观察更接近于 2 类。

用 sklearn.cluster.KMeans 函数，设定聚类数量为 3，对前四个数据进行聚类（后两组数据是单类的数据，无法进行聚类分析）。两个分界清晰的数据集上 k-means 的表现如下所示：

```
dataset:1 train acc:0.99846,test acc:1.0

dataset:2 train acc:0.99067,test acc:0.99022
```

对于重度重叠的数据集，k-means 准确率如下所示：

```
dataset:3 train acc:0.78667,test acc:0.79556

dataset:4 train acc:0.78489,test acc:0.79022
```