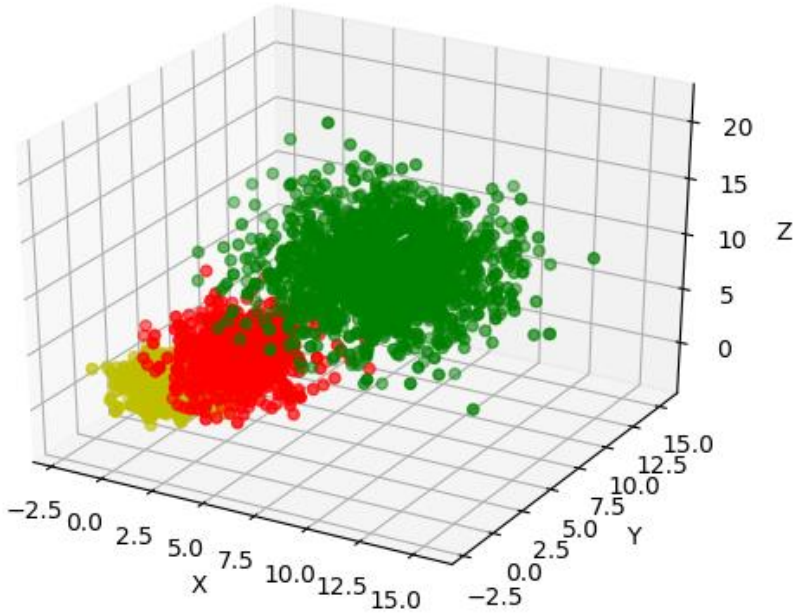
0. 采用 numpy 的正态分布生成 D1、D2 两个样本集，存在 D1.txt D2.txt 中。



对数据分析发现，各变量之间并无相关性，因此所有方差只取对角线，不取协方差。

1. 参考课本第 424 页算法，聚类得到先验概率、类中心与方差：

概率： [0.494375 0.505625]

中心：

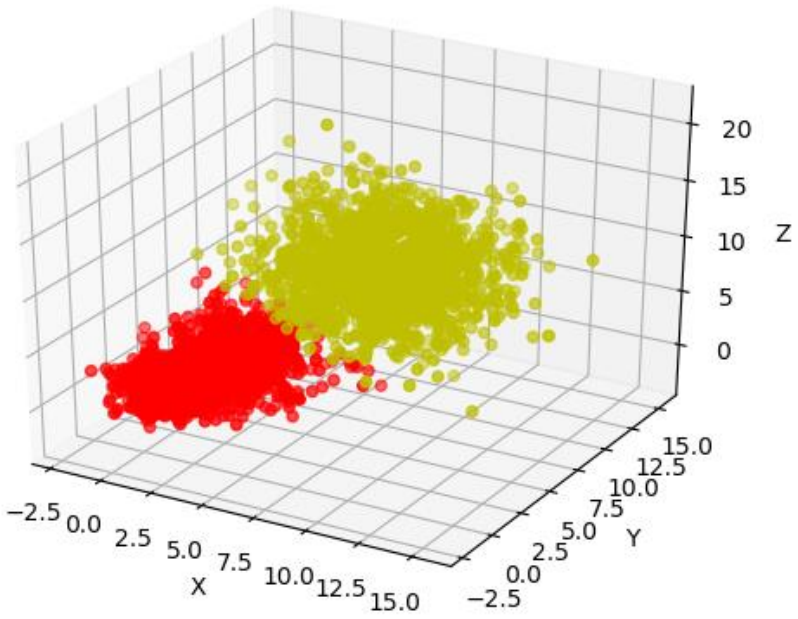
[[7.04378146 8.11445027 9.06403165]

[1.72365823 1.78818073 1.76387357]]

方差：

[[6.15669933 5.81943407 8.39677858]

[2.44600991 2.65178552 3.08617376]]



2. 参考课本第 423 页算法，随机得到初始参数，计算得到类的先验概率、中心和方差：

概率：[0.54214307 0.45785693]

中心：

[[6.75746702 7.77767943 8.64236841]

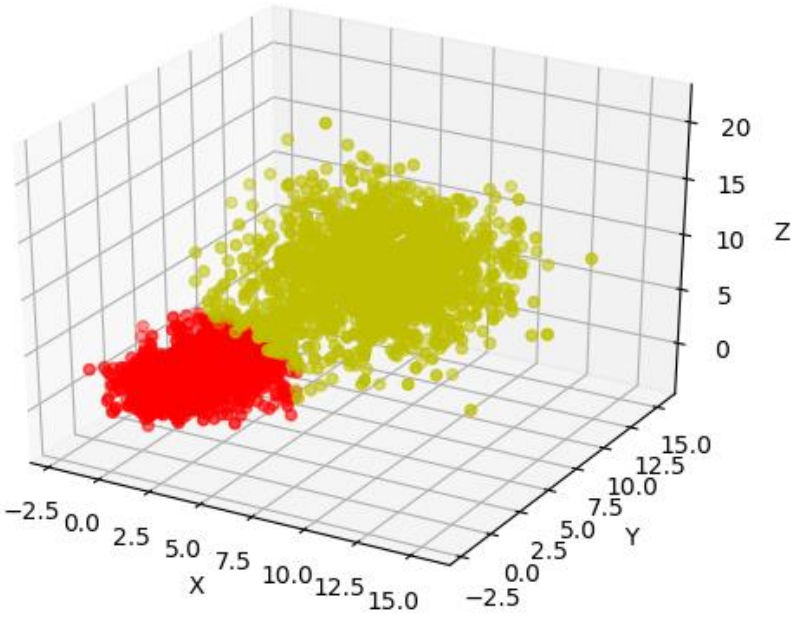
[1.50763299 1.52692938 1.50153683]]

方差：

[[6.72187665 6.76003699 9.85695109]

[1.90027703 1.8749611 2.2535663]]

取不同点对 2 类的后验概率最大者为其分类，可得：



3. 取 1 中得到的参数进行 MLE 计算，可得：

概率：[0.54214292 0.45785708]

中心：

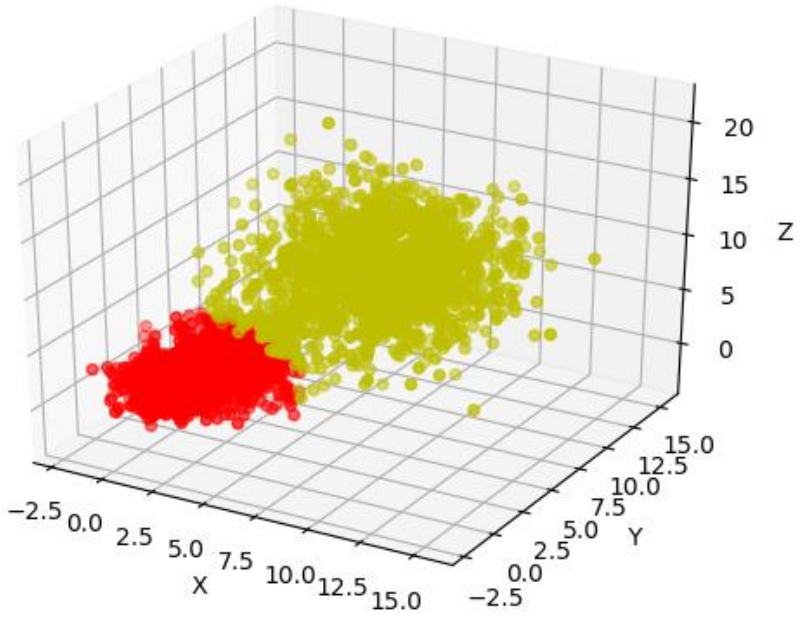
[[6.75746794 7.77768063 8.64236976]

[1.50763366 1.52693004 1.50153761]]

方差：

[[6.72187508 6.76003314 9.85694637]

[1.90027828 1.87496258 2.2535686]]



4. 对 1,2,3 的结果分析，发现二者结果相差不大，kmeans 方法分类更平均，而用 MLE 方法得到的两类则更易区别。使用随机参数和先验估计参数得到的结果几乎没有区别，迭代步数也相差不大。
5. 对聚类得到的结果，采用最近邻分类：离哪个类中心近就是哪类。聚类结果如下：

   概率：[0.255625 0.254375 0.49]

   中心：

   [[7.2361944 8.23001552 11.28716403]

   [6.67858518 7.81895386 6.54889792]

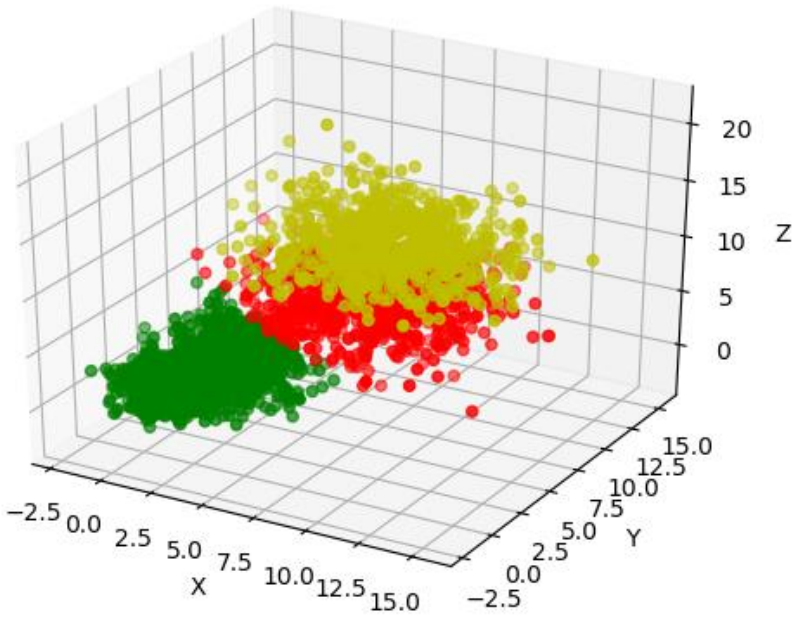
   [1.64321808 1.67956353 1.6770041]]

   方差：

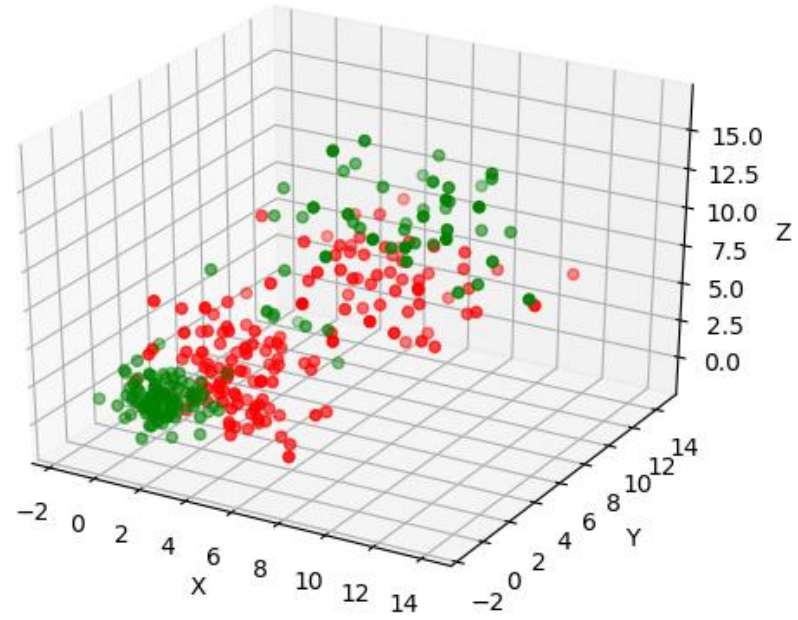
   [[6.35488953 5.88206121 3.38318853]

   [6.11178522 5.98437691 3.1150589]

   [2.19922626 2.27030128 2.83544921]]

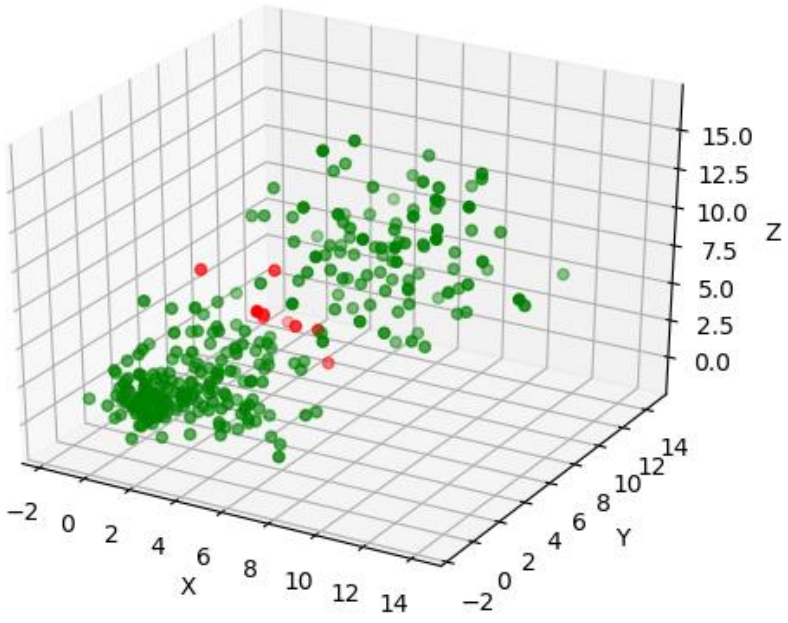


采用 abc 三类，正确率分别为 100%，9%，42%，可以看出分类成果较差。结果如下，其中绿色为被正确分类的点，红色为被错误分类的点：



如果根据聚类结果，将 ab 对应到结果中的第一类，c 对应到第二、第三类，可

以有 95.5%和 100%的正确率，结果如下：



6. 采用 MLE 方法得到结果，采用贝叶斯分类方法：用先验概率乘以该点在正态分布下的概率，取最大者分为其类。聚类结果如下：

概率：[0.19063218 0.30716485 0.50220296]

中心：

[[2.90653722 2.91439997 2.99597159]

[0.92767688 1.00918889 0.97959526]

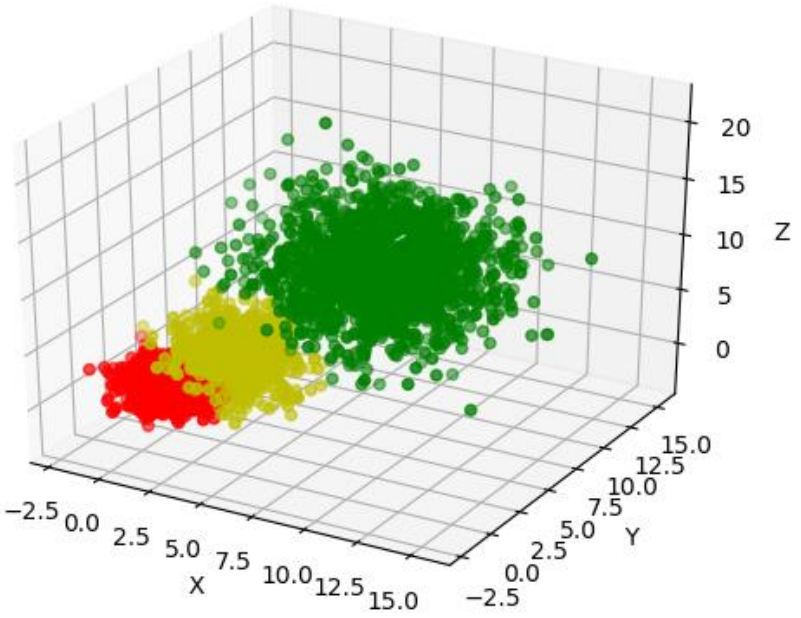
[6.99869404 8.06479599 8.96223979]]

方差：

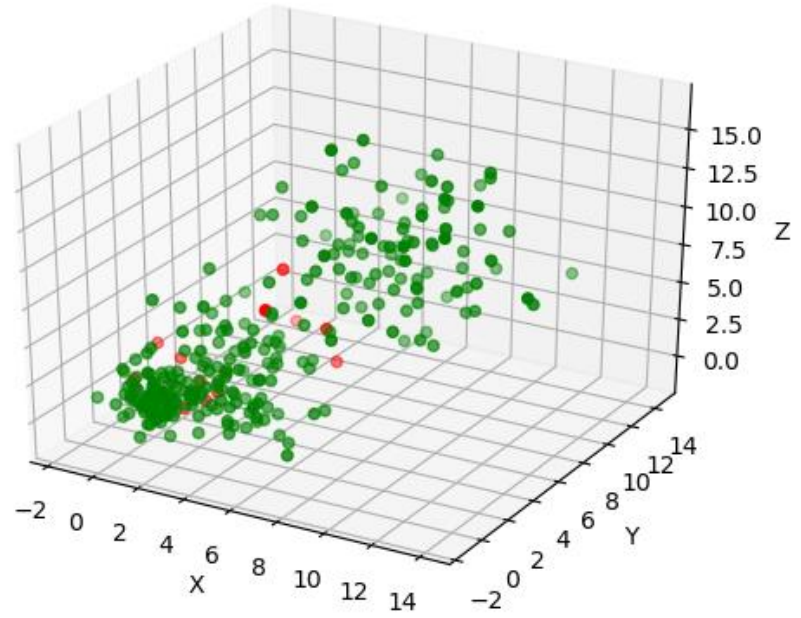
[[1.88811565 2.67756977 4.08583773]

[1.00636117 0.91593082 0.92981827]

[6.30905332 5.97131983 8.94613251]]

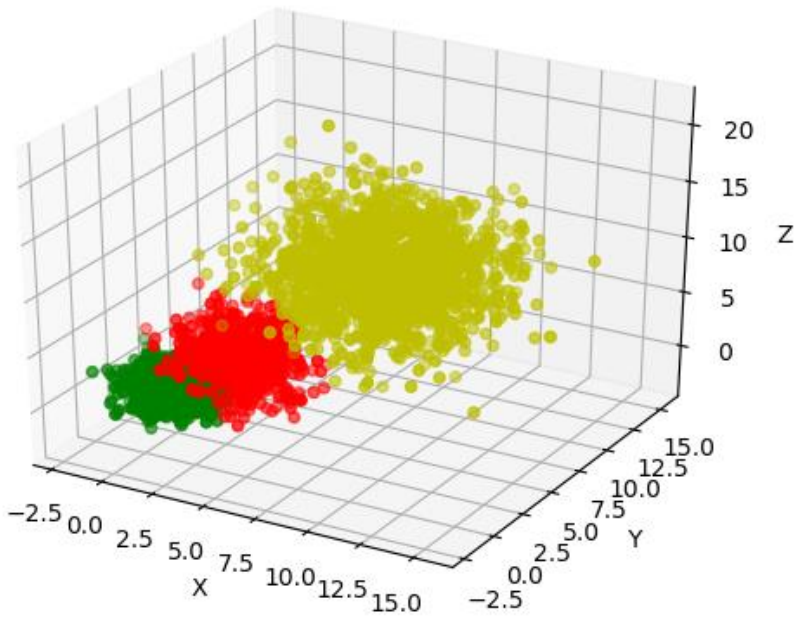


分类正确率为 98%，87%，100%，结果如下：



7. 将 5 中结果作为先验估计值使用 MLE 方法得到结果：

概率：[0.502203　　0.19063233 0.30716468]
中心：
[[6.99869386 8.06479579 8.96223957]
[2.90653591 2.91439855 2.99596976]
[0.92767642 1.00918847 0.97959497]]
方差：
[[6.30905355 5.97132026 8.94613294]
[1.88811608 2.67756916 4.08583642]
[1.00636062 0.91593016 0.92981775]]



分类结果与 6 中完全相同。

8. 分析 5,6,7 结果可知，两个方法聚类的结果相差较大，其中 kmeans 方法无法很好分类 AB，但是能较好地将 AB 与 C 分别。MLE 方法得到的正确率较高，结果显著。