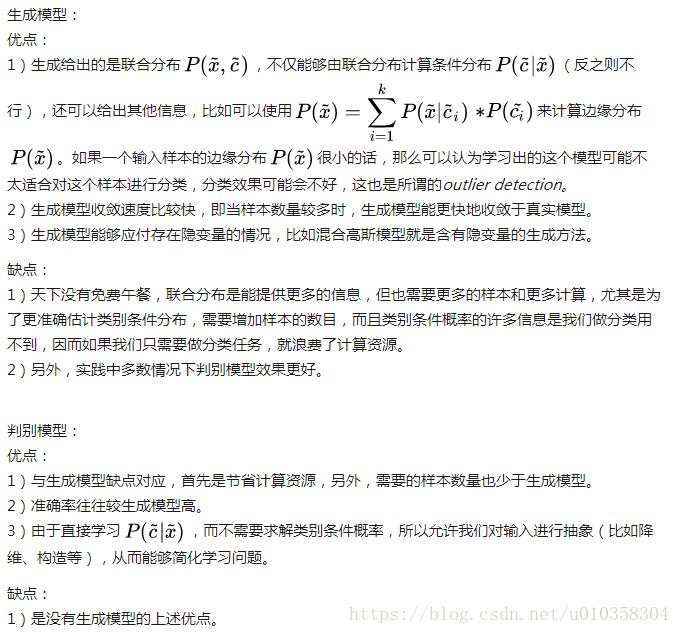
##### 1.什么是生成模型？什么是判别模型？

从概率分布的角度考虑，对于一堆样本数据，每个均有特征Xi对应分类标记yi。

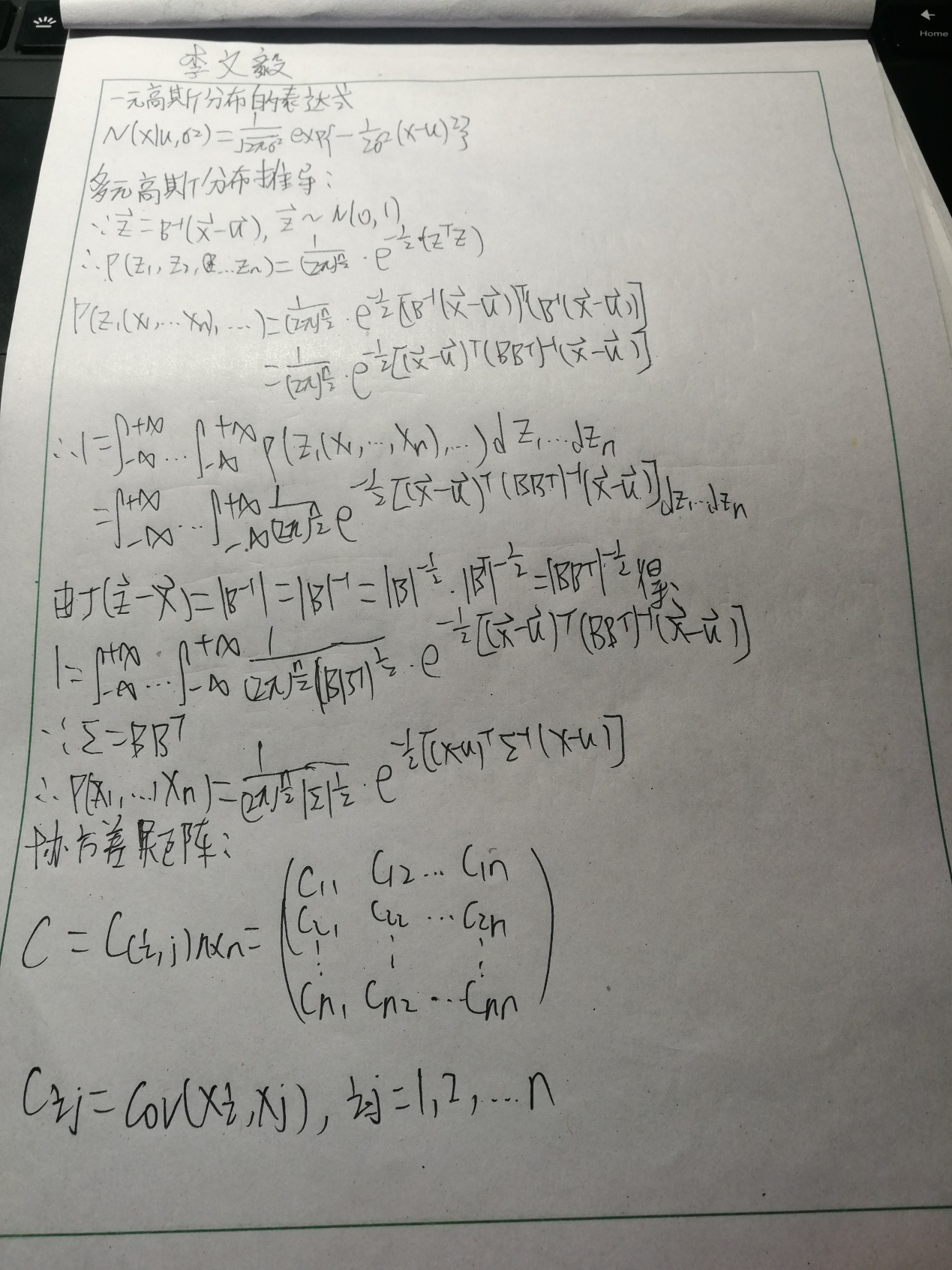
生成模型：学习得到联合概率分布P(x,y)，即特征x和标记y共同出现的概率，然后求条件概率分布。能够学习到数据生成的机制。

判别模型：学习得到条件概率分布P(y|x)，即在特征x出现的情况下标记y出现的概率。

数据要求：生成模型需要的数据量比较大，能够较好地估计概率密度；而判别模型对数据样本量的要求没有那么多。



##### 2.写出一元高斯分布的表达式，并推导多元高斯分布的表达式。同时写出协方差矩阵的具体形式



##### 3.如何判断一个模型是否发生了过拟合

模型在验证集合上和训练集合上表现都很好，而在测试集合上变现很差。

##### 4.什么是拉普拉斯平滑

拉普拉斯平滑（Laplace Smoothing）又被称为加 1 平滑，是比较常用的平滑方法。平滑方法的存在时为了解决零概率问题。

**背景:为什么要做平滑处理?**

  零概率问题，就是在计算实例的概率时，如果某个量x，在观察样本库（训练集）中没有出现过，会导致整个实例的概率结果是0。在文本分类的问题中，当一个词语没有在训练样本中出现，该词语调概率为0，使用连乘计算文本出现概率时也为0。这是不合理的，不能因为一个事件没有观察到就武断的认为该事件的概率是0。

**拉普拉斯的理论支撑：**

  为了解决零概率的问题，法国数学家拉普拉斯最早提出用加1的方法估计没有出现过的现象的概率，所以加法平滑也叫做拉普拉斯平滑。   
  假定训练样本很大时，每个分量x的计数加1造成的估计概率变化可以忽略不计，但可以方便有效的避免零概率问题。

##### 5.写出正态贝叶斯分类器的预测过程

