**生成式分类器**

05

生成式分类器对每个类别的数据生成过程建模，这样我们不仅能对数据进行分类，还能根据学习到的模型产生新的数据。生成式模型通常对数据产生过程有很强的假设，如果假设成立，生成式模型能根据较少的样本就能训练好模型。但如果假设不满足，模型的性能欠佳。

本章我们学习两种常用的生成式分类器：朴素贝叶斯分类器和高斯判别分类器。我们将讨论贝叶斯分类准则、两类模型的假设及其训练，并通过案例，学习Scikit-Learn中这两类分类器的API。

## **标题2** 5.1 生成式分类器

前面两章我们学习了Logistic回归模型和持向量机，这两种算法都是直奔分类目标，计算后验概率，或者判别函数的值，从而实现分类。我们称这类分类算法为**判别式分类器**，因为他们直接寻找不同类别之间的最优分类边界，主要关注不同类别的数据之间的差异，根据这种差异对样本进行分类。由于判别式分类器直接面对预测，往往准确率更高。典型的判别式分类器包括：K近邻，决策树，支持向量机，Logistic回归、条件随机场、提升算法等。

但有时我们更关心数据的生成过程，甚至可以根据学习到的模型生成更多的样本。**生成式分类器**从数据中学习类条件概率和类先验概率，然后根据贝叶斯规则计算出后验概率分布，实现分类。基于数据的联合概率密度分布，可以从分布中产生数据（如随机采样），所以这一类模型被称为**生成式分类器**。典型的生成式分类器有：朴素贝叶斯、隐马尔科夫模型（Hidden Markov Models, HMM）、混合高斯模型和贝叶斯网络等。

类条件概率密度分布描述每类数据的分布情况，反映同类数据的相似度，并不关心各类的分类边界在哪。生成式模型尝试去探索数据是怎么生成的，哪个类别最有可能生成样本，则该样本就属于那个类别。

当特征的维度很高时，类条件概率的学习是一件很困难的事情。通常我们对的形式作出约束，从而简化的学习。具体的，我们将学习两种常用的生成式分类器：朴素贝叶斯分类器和高斯判别分类器。朴素贝叶斯分类器假设在给定类别的情况下，特征与特征之间条件独立，这样联合分布的学习简化为多个一元分布学习。高斯判别分类器假设在给定类别的情况下，特征的联合分布为多元高斯分布。

## **标题2** 5.2 贝叶斯规则

### **5.2.1** 贝叶斯公式

设为试验的样本空间，为的一组事件，若

* ，
* ，

则为样本空间的一个划分。

如果为样本空间的一个划分，则对于每次试验，事件中有且仅有一个事件发生。

设试验的样本空间为，为样本空间的一个划分，为任意随机事件，则根据**全概率公式**，有：

贝叶斯定理：设试验的的样本空间为 ， 为样本空间的一个划分， 为任意随机事件，则有：

在分类任务中，事件为我们观测到输入样本的特征值为，样本空间为响应所有可能的类别，事件对应响应值，为响应可能取值的数目，则在分类任务为在给定输入的情况下，预测响应值的概率为

|  |  |
| --- | --- |
|  | （5-1） |

其中被称为先验概率，表示根据以往经验和分析得到的概率；被称为后验概率，表示根据已经发生的事件（输入样本的特征值为）分析得到的概率；为类条件概率，表示每个类别的样本特征分布。

## **标题2** 5.2 朴素贝叶斯分类器

### **5.2.2** 朴素贝叶斯分类器

朴素贝叶斯（Naïve Bayes）分类器基于贝叶斯定理和特征条件独立假设。朴素贝叶斯分类器中的“朴素”是因为这种方法的思想真的很朴素，因为它假设在给定类别的条件下，特征之间是独立的，即

|  |  |
| --- | --- |
|  | （5-2） |

由于上述条件独立假设，使得特征的类条件概率的学习非常简单，因为个1维的概率密度估计，比维联合概率密度估计简单得多。尽管这些朴素思想的假设过于简单，朴素贝叶斯分类器在很多任务（如文本分类任务）中仍然表现良好，且只需少量的训练数据来估计必要的参数，速度也非常快。

有了类条件概率和类先验概率，然后根据贝叶斯公式（5-1），得到后验概率，从而完成类别预测。

#### 1．类先验

如果类别数，分类任务是一个两类分类任务，类先验分布可用贝努利分布表示。当，分类任务是一个多类分类任务，类先验分布可用分布来表示。由于两类分类是多类分类的特例，下面我们只介绍多类分类的情况。

分布用于描述多类分类的概率分布，其参数为向量，其中，分量表示第个状态的概率。我们用符号表示为：

|  |  |
| --- | --- |
| ， | （5-3） |

其中为指示函数（Indicator function），当括号中的条件满足时，函数值为1，否则为0。

#### 2．类条件概率

在朴素贝叶斯模型中，类条件概率，因此我们只需知道单维特征的分布即可。

特征分布常见的情形有：

**（1） 贝努利分布：**当特征取值只有两种可能时（如某个词语在文档中是否出现），可用贝努利分布表示，其中参数表示在类别的情况下，特征的概率。

**（2）Multinoulli分布：**当特征可取多个可能的离散值时（如商品类别），可用多项分布Multinoulli表示。若特征共有种取值，则参数向量共有维，其中第个元素表示在类别的情况下，特征的概率。

**（3）多项分布：**当特征表示某个事件出现次数（如某个词语在文档中出现的次数），可用多项分布表示。若特征共有种取值，则参数向量共有维，其中第个元素表示在类别的情况下，特征的概率。

（4）**高斯分布**：当特征取值为连续值时（如某种花的花萼长度），可用高斯分布表示，其中参数分别表示在类别的情况下，特征的分布的均值和方差。

### **5.2.3** 朴素贝叶斯分类器的训练

在朴素贝叶斯分类器模型中，为了进行分类预测，我们需要知道类先验的和类条件概率。

#### 1．类先验

给定训练样本，估计只需用到训练样本标签的信息。类先验用Multinoulli分布Multinoulli表示，参数可用极大似然估计得到。

log似然函数为：

其中为的样本的数目。

加入约束，可用拉格朗日乘子法得到拉格朗日函数：

拉格朗日函数分别对求偏导数，并令其等于0，得到

所以得到的极大似然估计：

|  |  |
| --- | --- |
|  | （5-4） |

即类先验为训练样本中第类样本的比例。直观上，我们可以理解为用相对频率来估计概率。

除了极大似然估计，亦可用共轭先验Dirichlet分布表示参数的先验分布，从而得到参数的贝叶斯估计为

|  |  |
| --- | --- |
|  | （5-5） |

可以看出，当所有时，。实际应用中我们更多的采用简化版贝叶斯估计——拉普拉斯平滑（Lapalce smoothing）。拉普拉斯平滑相当于所有*，*即：

|  |  |
| --- | --- |
|  | （5-6） |

其中先验平滑因子 ，相当于每个类别下的样本数目增加1，可防止没有符合条件样本出现频率为0的情况。当训练样本数目足够大时，增加平滑因子并不会对结果产生大的影响。

#### 2．类条件概率

特征分布中的参数亦可用极大似然估计或贝叶斯估计求解。

1. 贝努利分布

当特征分布为贝努利分布时，参数求解推导过程类似上述先验分布参数的求解，只是研究对象为类别的情况下，特征的概率，即参数值为在所有类别标签为的样本中，特征值的样本的比例：

|  |  |
| --- | --- |
|  | （5-7） |

其中先验平滑因子，为标签的样本的数目，为类别标签，且特征值的样本数目。

1. Multinoulli分布

当特征可以取多个离散的值（如商品类别）时，可用Multinoulli分布建模，参数求解推导过程类似上述先验分布参数的求解，只是研究对象为类别的情况下，特征的概率，即参数值为在所有类别标签为的样本中，特征值的样本的比例：

|  |  |
| --- | --- |
|  | （5-8） |

其中为特征所有可能取值的数目，先验平滑因子，为标签的样本的数目，为类别标签，且特征值的样本数目。

1. 多项分布

当特征值表示某种事件发生的次数时，可用多项分布建模。多项分布为Multinoulli分布中试验次数为多次的情况。参数求解公式同Multinoulli分布。

1. 高斯分布

当特征分布为高斯分布时，

|  |  |
| --- | --- |
|  | （5-9） |

模型参数可采用极大似然法来估计。对上述高斯分布，数据的log似然函数为：

其中表示标签的样本的数目。

函数分别对求偏导数，并令其等于0，得到

得到参数的极大似然估计：

|  |  |
| --- | --- |
|  | （5-10） |

即为所有标签的样本的第维特征的均值，为这些样本第维特征的样本方差。

当然，根据特征的物理含义，我们也可以定义其为某种分布，分布的参数可采用极大似然或者贝叶斯方方法估计。如果特征分布不能用某种参数模型建模，还可以采用混合高斯模型、核密度估计、或者直方图表示。

例5-1： 给定如表5-1所示的训练数据，采用朴素贝叶斯分类器，根据用户的属性（日志密度、好友密度、是否使用真实头像），判断用户社交网络服务（Social Networking Services，SNS）账号是否真实。

表**5-1** SNS账号真实性判断案例

|  |  |  |  |
| --- | --- | --- | --- |
| 日志密度 | 好友密度 | 是否使用真实头像 | 账号是否真实 |
| s | s | no | no |
| s | l | yes | yes |
| l | m | yes | yes |
| m | m | yes | yes |
| l | m | yes | yes |
| 续表 | | | |
| 日志密度 | 好友密度 | 是否使用真实头像 | 账号是否真实 |
| m | l | no | yes |
| m | s | no | no |
| l | m | no | yes |
| m | s | no | yes |
| s | s | yes | no |

1. 根据训练数据，计算类先验概率

共有10个样本，，其中账户真实的样本有7个，，不真实账号的样本有3个，，令，根据式（5-6），有：

1. 根据训练数据，计算特征的类条件概率
2. 当类别标签时：，

特征日志密度有s，m，l三种取值，用Multinoulli分布建模，。

在的7个样本中，上述3种特征取值的样本数分别为：1、3、3，即

根据式（5-8），得到：

类似地，得到其他2维特征的类条件分布的参数为：

1. 当类别标签时：，

特征日志密度有s，m，l三种取值，用Multinoulli分布建模，。

在的3个样本中，上述3种特征取值的样本数分别为：2、1、0，即

根据式（5-8），得到：

类似地，得到其他2维特征的类条件分布的参数为：

1. 根据前两步得到的模型参数，对新的样本进行预测

现有一个用户，其日志密度为m，好友密度为m，使用真实头像，则：

，所以该用户是真实账号的可能性更高。

### **5.2.4** 案例分析**——Otto**商品分类

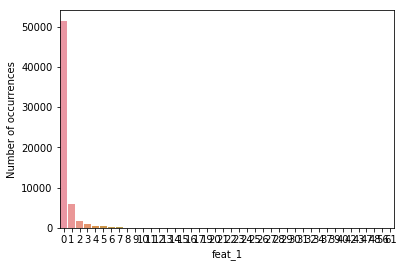
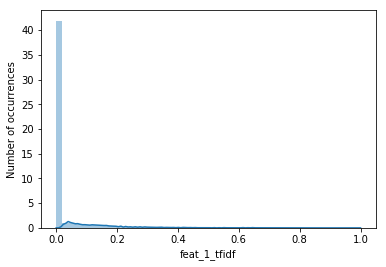
本节我们以Kaggle 2015年举办的奥拓商品分类竞赛数据为例，采用朴素贝叶斯分类器实现商品分类。数据集介绍请见3.7节。

Scikit-Learn中提供4种朴素贝叶斯的分类算法：、和和。Scikit-Learn中这几种朴素贝叶斯模型都是假设所有特征都是同一种分布，读者在使用时需要注意。其中假设特征的类条件概率是为高斯分布，通常用于连续特征值。假设是多项分布，常用于计数特征，如文档分类中的词频特征。虽然要求输入为计数特征，但Scikit-Learn文档也说明TF-IDF特征也可以取得不错的效果。假设是贝努利分布，用于二元离散值特征，如文档分类任务中的特征表示某个词语是否出现。相比，在短文档分析中，也许更有优势。

在本案例中，特征取值均为稀疏的整数特征，特征变化也可将其转换成TF-IDF特征。以第1维特征为例，原始特征和TF-IDF特征的直方图如图5-1所示。对原始特征，我们可以将其视为离散型特征，考虑和分类器。对TF-IDF特征，可考虑采用和。

通常采用默认参数即可，没有需要调优的超参数。可以设置拉普拉斯平滑，一般采用默认值1即可。如果发现拟合得不好，可以采用交叉验证进行超参数调优，可以选择稍大于1或者稍小于1的数。比多出一个参数，默认值为0.0，此时认为每个数据特征都已经是二元（取值为0或1）的。否则的话，小于的会归为0，大于的会归为1。

在Otto商品分类任务上，我们采用默认参数，上述分类器在测试集上的性能如表5-2所示。可以看出，对特征进行二值化，丢失太多信息，分类效果不好。而假设特征的类条件分布为高斯分布， TF-IDF特征的分布离高斯分布很远，所以效果惨不忍睹。效果最好的是采用TF-IDF特征的。但即使这个最好的性能（0.79641），比Logistic回归模型的性能要差（0.63319）。所以朴素贝叶斯分类器的假设太强（特征条件独立、特征的条件分布类型），当实际数据不符合这些假设时，性能很差。相反，判别式分类器（Logistic回归、SVM）直接从数据中找出不同类别数据之间的差异，效果更好。

（a） 原始特征直方图 （b）TF-IDF特征的直方图

图5-1 Otto数据集第1维特征的直方图。

表5-2 Otto商品分类数据集上不同朴素贝叶斯分类器的测试误差

|  |  |  |
| --- | --- | --- |
| 分类器 | 特征 | 模型性能（logloss） |
|  | 原始特征 | 1.77729 |
|  | 原始特征 | 1.29931 |
|  | TF-IDF特征 | 0.79641 |
|  | TF-IDF特征 | 5.22518 |

### **5.2.5** 案例分析——新闻分类

朴素贝叶斯分类器的用武之地之一是文本分类。本节我们对Scikit-Learn自带的新闻分类数据，采用朴素贝叶斯分类器进行分类。

20newsgroups数据集是用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约11,314个新闻组文档，分档分为20个不同主题。我们将其中80%作为训练数据，20%为测试数据。对每个文档，我们用Scikit-Learn中的TfidfVectorizer提取其TF-IDF特征，我们采用2元语法模型，最终得到的特征向量为155,785维。用分类器得到的结果如表5-3所示，总体正确率为0.894388，取得了不错的性能。事实上朴素贝叶斯模型在很多文本分类任务上都能取得很好的性能。我们也比较了Logistic回归， 正确率为0.935484，比朴素贝叶斯分类器性能稍好。

表5-3 新闻分类数据集上朴素贝叶斯分类器的测试误差

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 新闻类别 | Precision | Recall | F1-score | Support |
| alt.atheism | 0.97 | 0.94 | 0.95 | 89 |
| comp.graphics | 0.79 | 0.84 | 0.81 | 99 |
| comp.os.ms-windows.misc | 0.89 | 0.91 | 0.90 | 130 |
| comp.sys.ibm.pc.hardware | 0.86 | 0.80 | 0.83 | 109 |
| comp.sys.mac.hardware | 0.92 | 0.92 | 0.92 | 117 |
| comp.windows.x | 0.92 | 0.89 | 0.91 | 118 |
| misc.forsale | 0.85 | 0.95 | 0.90 | 117 |
| rec.autos | 0.95 | 0.94 | 0.95 | 121 |
| rec.motorcycles | 0.96 | 0.97 | 0.97 | 119 |
| rec.sport.baseball | 0.98 | 1.00 | 0.99 | 113 |
| rec.sport.hockey | 1.00 | 0.96 | 0.98 | 129 |
| sci.crypt | 0.96 | 0.99 | 0.98 | 109 |
| sci.electronics | 0.91 | 0.93 | 0.92 | 120 |
| sci.med | 0.99 | 0.91 | 0.95 | 123 |
| sci.space | 0.97 | 0.97 | 0.97 | 119 |
| soc.religion.christian | 0.93 | 0.98 | 0.95 | 128 |
| talk.politics.guns | 0.97 | 0.96 | 0.97 | 120 |
| talk.politics.mideast | 1.00 | 1.00 | 1.00 | 100 |
| talk.politics.misc | 0.94 | 0.94 | 0.94 | 106 |
| talk.religion.misc | 0.96 | 0.84 | 0.90 | 77 |
| avg / total | 0.94 | 0.94 | 0.94 | 2263 |

## **标题2** 5.3 高斯判别分析

### **5.3.1** 高斯判别分析的基本原理

高斯判别分析假设在给定类别的情况下，每类的特征分布为多元高斯分布，即

|  |  |
| --- | --- |
|  | （5-11） |

其中为特征的维度，为第类的样本的均值向量，为第的样本的协方差矩阵。

将上述类条件概率代入贝叶斯公式

我们计算两类的对数几率为

|  |  |
| --- | --- |
|  | （5-12） |

其中是无关的常数项。当两类的对数几率等于0时，属于两类的概率相等，位于决策边界上。上述对数几率是的二次函数，所以两类的决策边界为二次曲线，该算法被称为二次判别分析（Quadratic Discriminant Analysis, QDA）分类器。

当两个类的协方差时，两类的对数几率为

|  |  |
| --- | --- |
|  | （5-13） |

此时决策边界是的一次函数，两类的决策边界为直线，称为线性判别分析（Linear Discriminant Analysis, LDA）分类器。

综上所述，特征的类条件分布为高斯分布， 且

* 每个类别的协方差矩阵相等：线性判别分析，决策边界为超平面，如图5-2左上角；
* 每个类别的协方差矩阵为对角阵：朴素贝叶斯，如图5-3左上角；
* 其他：二次判别分析，决策边界为二次曲面，如图5-2和图5-3的右侧图。

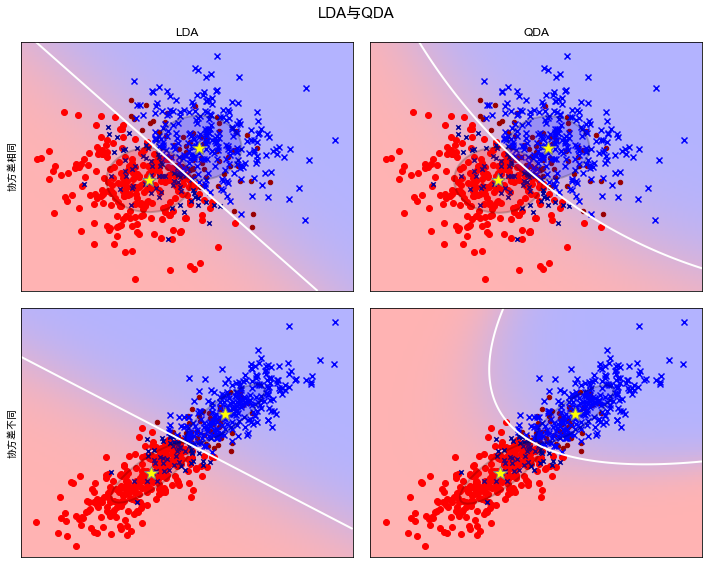


图5-2 LDA与QDA的比较。第一行中两个类的数据的协方差相同，此时LDA与QDA的结果非常接近。第二行中两个类的数据的协方差不同，此时LDA分类器的决策边界与QDA分类器有很大不同。黄色五角星为每类的均值，黑色的细椭圆曲线为每类的协方差矩阵表示，白色粗曲线为分类器的决策边界。

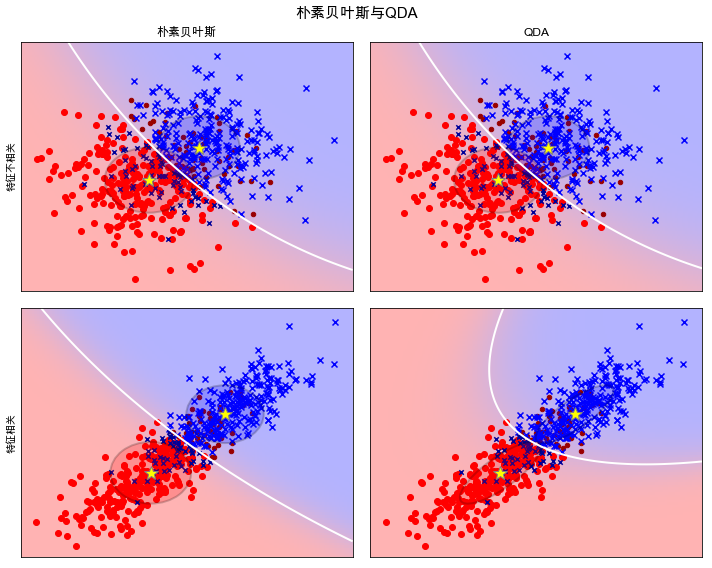


图5-3 朴素贝叶斯与QDA比较。第一行中两类数据的2维特征之间不相关（协方差矩阵为对角矩阵），此时朴素贝叶斯与QDA分类器的结果非常接近。第二行中两类数据的2维特征之间相关，此时朴素贝叶斯分类器的决策边界与QDA分类器有很大不同。黄色五角星为每类的均值，黑色的细椭圆曲线为每类的协方差矩阵表示，白色粗曲线为分类器的决策边界。

### **5.3.3** 高斯判别分析的模型训练

模型训练时，我们需要根据训练数据估计模型参数：每个类别的先验分布、均值向量和协方差矩阵。类先验的参数估计同朴素贝叶斯模型中类先验参数估计。均值向量和协方差矩阵的估计一般采用极大似然估计，即用属于类别的所有样本的均值得到估计、用属于类别的所有样本的样本协方差矩阵估计：

|  |  |
| --- | --- |
|  | （5-10） |

当训练样本数量相比特征维度较小时，上述协方差矩阵是奇异的，可采用收缩（Shrinkage）提升的协方差矩阵预测准确性：，其中是收缩因子，为一个高度结构化的矩阵，如对角阵[7]。

### **5.3.3** 案例分析——**MNIST**手写数字识别

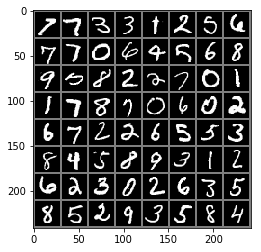


图5-4 MNIST数据集中的手写数字图像。

在深度学习流行之前，LDA分类器也是图像分类（如人脸识别）任务的主流分类器之一。我们对手写体数字识别数据集MNIST采用LDA分类器进行数字识别。

Kaggle上的MNIST数据集是一个手写数字图像数据集，训练集42,000张，测试集 28,000张，共有10个类别（数字0-9），图像是尺寸为（28, 28）的黑白图，每个像素值为0到和255之间数值。图5-4给出了几个示例图像。

我们首先用保持95%能量的主成分分析（Principle Components Analysis, PCA）对原始数据（维）降维，得到154维特征，然后将PCA特征送入LDA分类器。PCA的详细描述请见第10章，这里我们可以将其视为一种图像特征提取方法。5折交叉验证得到正确率为0.868452。这个结果比用SVM分类器要稍差一些。

## **标题2** 5.4 小结

分类器可以采用判别式模型实现，也可也以采用生成式模型实现。通常生成式分类器的分类性能比判别式稍差，因为实际任务中的数据不太符合模型假设的分布。但在假设正确的情况下，生成式模型训练所需的样本更少，并且可以描述数据生成的过程，生成新的样本。

## **标题2** 5.5 练习

1. 采用朴素贝叶斯分类器对水果进行分类。

假设我们有1000个水果的数据。水果是香蕉、桔子或其他水果，我们知道每种水果的三个特征：形状是否为长条、味道是否甜；颜色是否为黄色。这1000个水果的信息如下表所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 水果类型 | 是否为长条 | 是否甜 | 是否为黄色 | 总共 |
| 香蕉 | 400 | 350 | 450 | 500 |
| 桔子 | 0 | 150 | 300 | 300 |
| 其他 | 100 | 150 | 50 | 200 |
| 总共 | 500 | 650 | 800 | 1000 |

1. 根据上述数据，估计每类水果的先验概率；
2. 根据上述数据，估计每个特征的类条件概率；
3. 现有一个水果，形状为长条、味道甜，且颜色是黄色，请问它最可能是哪种水果？
4. 二次判别分析和线性判别分析。

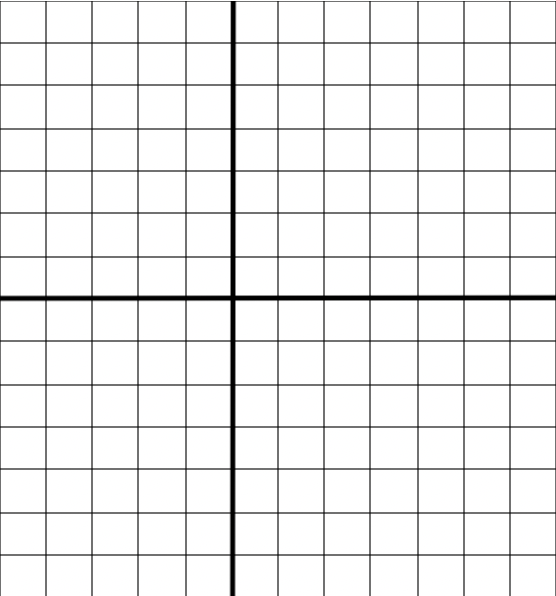
假设2维空间中有12个样本点，分为3类：

第1类样本：，

第2类样本：，

第3类样本：，

1. 计算每类的样本均值和协方差矩阵，；
2. 计算每类的先验概率；
3. 在图中大致画出每个类的QDA判别边界和正态分布（中心位置、长短轴的方向和长度）。



1. 如果采用LDA，会得到好的分类边界吗？
2. 朴素贝叶斯分类器和SVM在垃圾邮件过滤任务均有较好的性能。请对国际文本检索会议提供中文垃圾邮件过滤数据集（trec06c）用朴素贝叶斯分类器分类。作业提供的数据集已经经过数据预处理，每个样本的有2个字段：每个邮件正文的类别（正常邮件为0、垃圾邮件为1）、分词后以空格隔开的邮件正文。
3. 对每个邮件，提起词频特征（TF）和TF-IDF特征；
4. 从上述数据集中随机选择80%数据作为训练集，剩下20%作为测试集，分别采用TF特征和TF-IDF特征，训练朴素贝叶斯分类器，并计较二者在测试集上的性能。注意SVM的超参数调优。
5. 和第（2）题相同的训练集和测试集划分，分别用线性SVM和RBF核的SVM分类器实现垃圾邮件过滤，并与朴素贝叶斯分类器的性能做比较。
6. 请采用QDA对5.3.3节中的MINIST数据集进行手写数字识别，并与LDA的性能相比较。