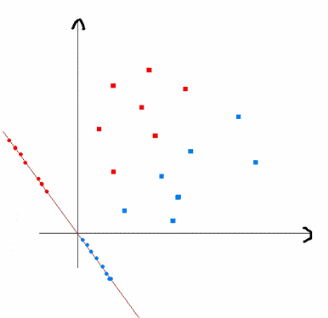
LDA(Linear Discriminant Analysis) 线性判别分析

LDA是一种 supervised learning。Discriminant，一个模型，不需要去通过概率的方法来训练、预测数据，比如说各种贝叶斯方法，就需要获取数据的先验、后验概率等等。LDA是在目前Machine Learning、Data Mining 经典且热门的一个算法。

LDA的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。要说明白LDA，首先得弄明白线性分类器：因为LDA是一种线性分类器。对于K-分类的一个分类问题，会有K个线性函数：

当满足条件：对于所有的j，都有Yk>Yj 的时候，我们就说x属于此类别k。对于每一个分类，都有一个公式去算一个分值，在所有的公式得到的分值中，找一个最大的，就是所属的分类了。

上式实际上就是一种投影，是将一个高维的点投影到一条高维的直线上，LDA最求的目标是，给出一个标注了类别的数据集，投影到了一条直线后，能够使得点尽量的按类别区分开，当k=2即二分类问题的时候，如下图所示



红色的方形的点为0类的原始点，蓝色的方形点为1类的原始点，经过原点的那条线就是投影的直线，从图上可以清楚的看到，红色的电荷蓝色的点被原点明显的分开了，下面推导一下二分类LDA问题的公式：

假设用来区分二分类的直线（投影函数）为：

LDA分类的一个目标是使得不同类别之间的距离越远越好，同一类别之中的距离越近越好，所以需要定义几个关键的值。

类别i 的院士中心点为: （Di表示属于类别 i 的点）

类别 i 投影后的中心点为：

衡量类别 i 投影后，类别点之间的分散程度（方差）为：

最终我们可以得到一个下面的公式，表示LDA投影到w后的损失函数：

我们分类的目标是，使得类别内的点距离越近越好（集中），类别间的点越远越好。分母表示每一个类别内方差只和，方差越大表示一个类别内的点越分散，分子为两个类别各自的中心点的距离的平方，我们最大化J(w)就可以求出最优的w了。想要求出最优的w，可以使用拉格朗日乘子法，但是现在我们得到的J(w)里面，w是不能被单独提出来的，我们得想办法将w单独提出来。

我们定义一个投影钱的各类别分散程度的矩阵，这个矩阵看起来有一点麻烦，其实意思是，如果某一个分类的输入点集Di 里面的点距离和这个分类的中心点mi 越近，则Si里面元素的值就越小，如果分类的点都紧紧地围绕着mi，则S里面的元素值越更接近0.

带入Si，将J(w)分母化为：

同样的将J(w)分子化为：

这样损失函数可以化成下面的形式：

我们将分母限制为长度为1，并作为拉格朗日乘子法的限制条件，带入得到：

这样的式子是一个求特征值的问题，对于N（N>2）分类的问题，结论如下：

这同样是一个求特征值的问题，我们求出第i 大的特征向量，就是对应的Wi了。

PCA：主成分分析与LDA有非常相似的意思，LDA的输入数据是带标签的，而PCA的输入数据是不带标签的，所以PCA是一种unsupervised learning。LDA通常来说是作为一个独立的算法存在，给定了训练数据后，将会得到一系列的判别函数（discriminate function），之后对于新的输入，就可以进行预测了。二PCA更像是一个预处理的方法，它可以将原本的数据降低纬度，而使得降低了维度的数据之间的方差最大（也可以额说投影误差最小，具体在之后的推导里面会谈到）。

有些时候我们会考虑减少方差（比如说训练模型的时候，我们会考虑到方差-偏差的均衡），有的时候我们会尽量的增大方差，从实践来说，通过尽量增大投影方差的PCA算法，确实可以提高我们的算法质量。

两种思路来推导出一个同样的表达式。首先是最大化投影后的方差，其次是最小化投影后的损失（投影产生的损失最小）。

最大化方差法：

假设我们还是讲一个空间中的点投影到一个向量中去。首先，给出原空间的中心点：

假设u1为投影向量，投影之后的方差为：

拉格朗日乘子法：

将上式求导，使之为0，得到：

这是一个标准的特征值表达式了，对应的特征值，u对应的特征向量。上式的左边取得最大值的条件就是最大，也就是取得最大的特征值的时候。假设我们是要将一个D维的数据空间投影到M维的数据空间中（M<D），那么我们取前M个特征向量构成的投影矩阵就是能够使得方差最大的矩阵了。

最小化损失法：

假设输入数据x是在D维空间中的点，那么，我们可以用D个正交的D维向量去完全的表示这个空间（这个空间中所有的向量都可以用这D个向量的线性组合得到）。在D维空间中，有无穷多种可能找到这D个正交的D维向量，哪个组合是最合适的呢？

假设我们已经找到了这D个向量，可以得到：

我们可以用近似法来表示投影后的点：

上式表示，得到的新的x是由前M个基的线性组合加上后D-M个基的线性组合，注意这里的z是对于每个x都不同的，而b对于每个x是相同的，这样我们就可以用M个数来表示空间中的一个点，也就是使得数据降维了。但是这样降维后的数据，必然会产生一些扭曲，我们用J来描述这种扭曲，我们的目标是，使得J最小：

上式的意思很直观，就是对于每一个点，将降维后的点与原始的点之间的距离的平方和加起来，求平均值，我们就要使得整个平均值最小。我们令：

将上面得到的z与b带入降维的表达式：

将上式带入J的表达式得到：

再用上拉普拉斯乘子法，可以得到，取得我们想要的投影基的表达式为：

这里又是一个特征值的表达式，我们想要的前M个向量其实就是这里最大的M个特征值所对应的特征向量。J可以化为：

也就是当误差J是由最小的D-M个特征值组成的时候，J取得最小值。跟上面的意思相同。