一、线性判别分析的基本思想

“给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离；在对新样本进行分析时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别。”

书中的这段话总感觉缺了点什么，缺的就是具体的判断方法，究竟哪个位置才算“正例”和“反例”的分界线呢？

首先，求出训练集“正例”、“反例”的均值中心点。之后，过这两个均值点的中点做LDA直线的垂线。那么，这条垂线应该就是“正例”与“反例”分界的“超平面”。

二、程序运行结果

程序运行结果图：

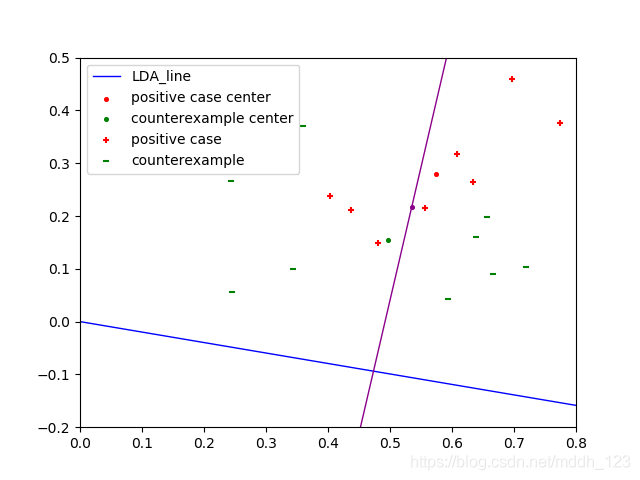
图中，红色的“+”表示“正例”样本，绿色的“-”表示“反例”样本；红色的圆圈代表“正例”均值中心，绿色的源泉代表“反例”均值中心；蓝色的直线为LDA直线；紫色的直线为LDA算法求得的正例、反例的分界线，也就是“超平面”。

**三、结果分析**

LDA算法求得的这个“超平面”的划分效果并不好。图中，正例均值中心在紫色直线右边，因此右边应该是程序经过学习后划分的“正例”范围，左边为“反例”范围。

显然，LDA算法在这个数据集的划分正确率仅为9/17（大约为50%）。50%的正确率意味着什么？假设数据集中17条数据正、反例的频率能够接近二者发生的概率（即假设好瓜、坏瓜出现的概率均为50%左右），假设好瓜、坏瓜的出现相互独立。这就是说，如果你在不知道任何信息的前提下，瞎蒙的正确率也能达到50%。所以，LDA算法在该数据集上根本没有获得任何有价值的信息。

这说明LDA算法只适用于线性可分的数据集，这点和“硬间隔”支持向量机十分类似。（事实上，二者也存在联系。之间联系将在以后介绍）



[**周志华机器学习 CH3.5 编程实现线性判别分析**](https://blog.csdn.net/Elvirangel/article/details/84237089)