Problem4：

用给定分类的二维数据来训练 SVM 模型。要求将(x_1, x_2)转化为高维向量$(1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$，由第三题计算可知，对应的核函数为 2 维的多项式核$\phi(x_i, x_j) = (x_i^T x_j + 1)^2$，经过坐标放缩变换可得。

4.1 用 w1 和 w2 的第一组数据训练 SVM，利用 scikit-learn 中 svm 包训练。在二维平面内分类结果如图 1 所示。黑色实线即为分离超平面，两个数据点均为支撑向量，中心实心点表示数据点，黑色圆圈表示支撑向量。计算得到间隔为[-1.00000002 1.00000001]。简单思考可知，在只有两个数据点的情况下，分类超平面应为两点对称中线。

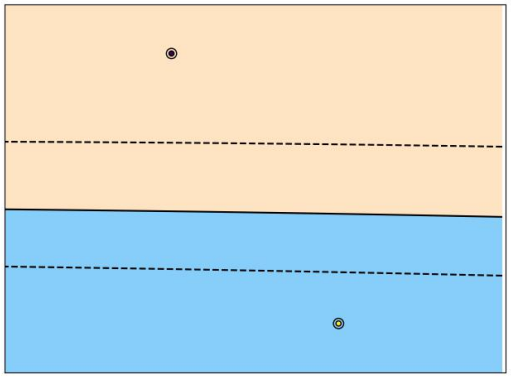
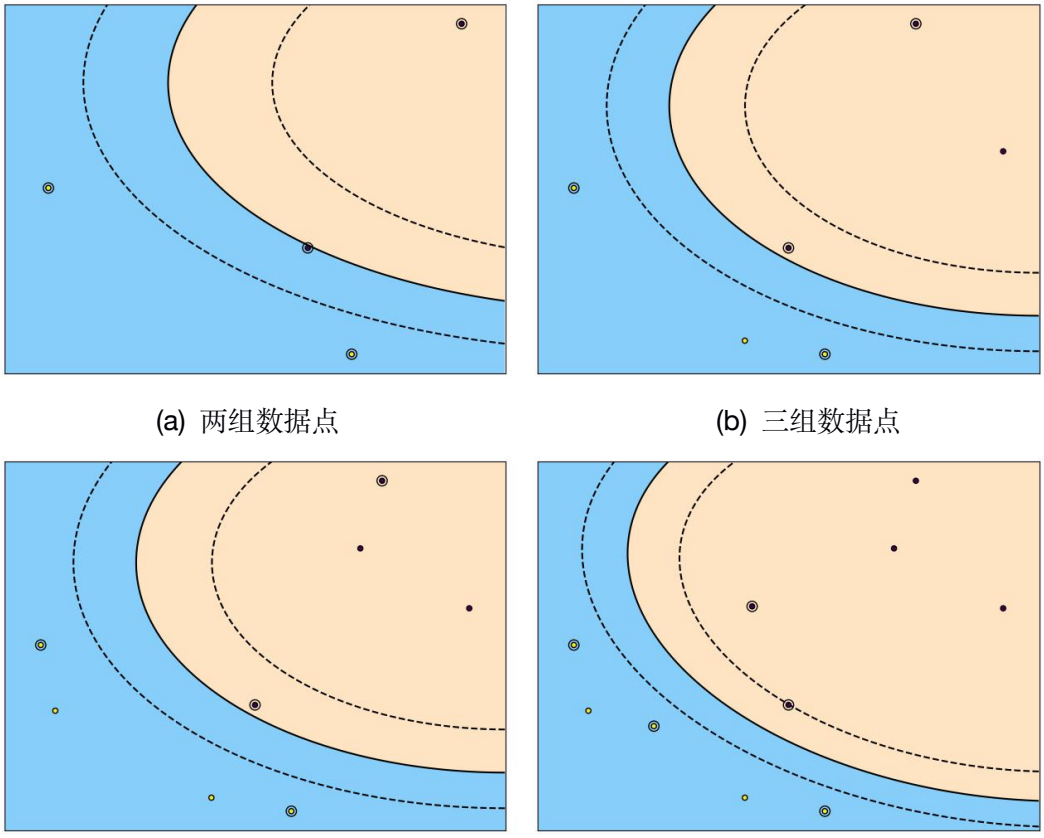


图 1：一组数据点训练得到的 SVM 分类结果

4.2 增加数据点重复实验结果如下图 2 所示。黑色实线即为分离超平面，中心实心点表示数据点，黑色圆圈表示支撑向量。



(a) 两组数据点　　(b) 三组数据点

(c) 四组数据点　　(b) 五组数据点

图 2：多组数据点训练得到的 SVM 分类结果

计算得到间隔为

(a)[0.02315585 -1.00057049 1.00028518 1.00028527]

(b)[-0.1133536 -1.00062296 1.00031147 1.00031135]

(c)[-0.11013534 -1.00037343 1.00018671 1.00018675]

(d)[-0.49332285 -1.00010916 0.99989081 1.00049537 0.99972256]

由此可知，随着样本点的增加，分离超平面不断变化，图 2(a)中在超平面上的样本点逐渐远离分离超平面，说明分类效果不断变优。

增加到六组数据时，二维平面出现了错分数据，即在高维空间不能线性分离两类样本点，如图 3 所示，红框表示即为错分数据。

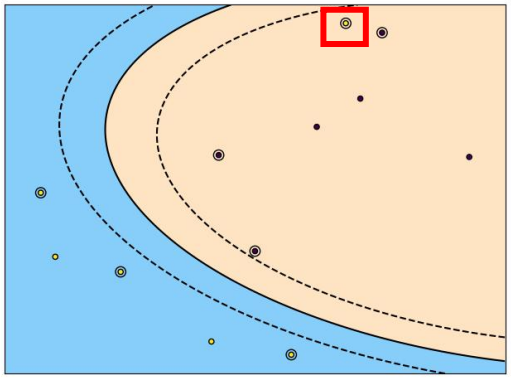


图 3：六组数据分类结果

Reference：

scikit-learn for python:

https://scikit-learn.org/stable/modules/svm.html#svm-classification