# 《机器学习与深度学习》课程

# 实 验 报 告



**姓 名： 金家耀**

**专 业：**  人工智能

**学 号： 1193210320**

**江南大学人工智能与计算机学院**

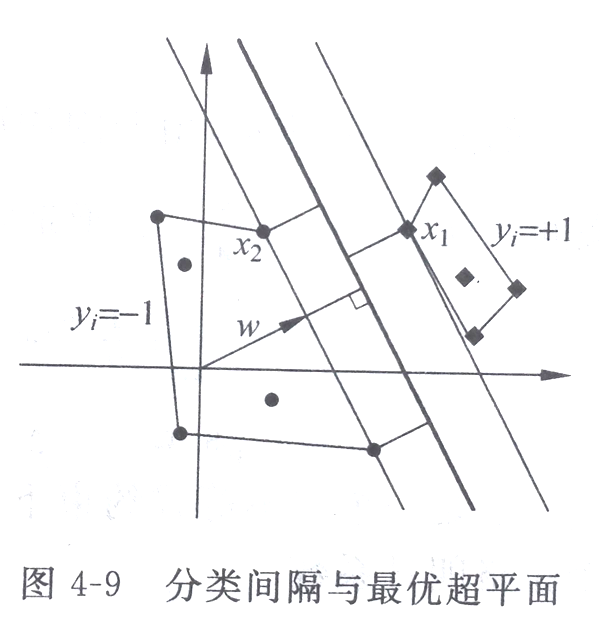
# SVM分类器

**1实验目的**

支持向量机（SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面，在[人像识别](https://baike.baidu.com/item/%E4%BA%BA%E5%83%8F%E8%AF%86%E5%88%AB/1705333?fromModule=lemma_inlink" \t "_blank)、[文本分类](https://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB/7267115?fromModule=lemma_inlink" \t "_blank)等实际应用中广泛采用。本实验的目的在于加深学生对支持向量机的理解，掌握算法的实现过程，体会其在模式分类中的作用。

**2实验原理**

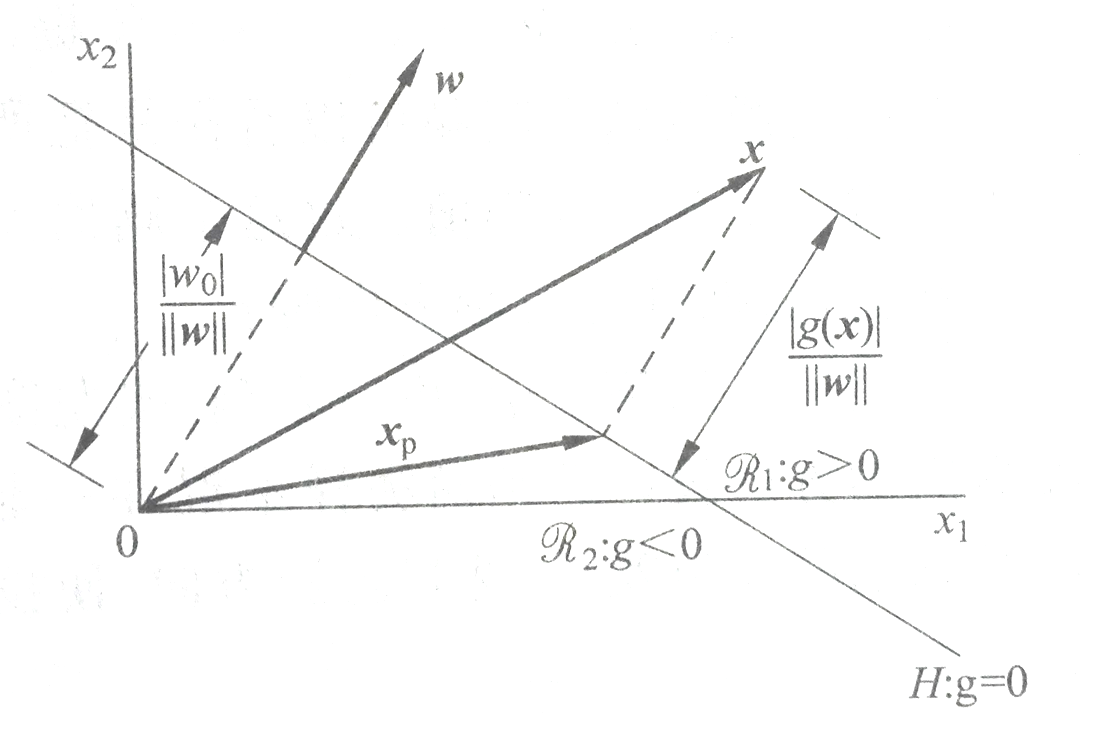
一个超平面，如果能够将训练样本无错分开，并且两类样本中，距离平面最近的样本与超平面的距离最大化，则把这个超平面称为最优分类超平面（也称最大间隔超平面），该最大化的距离称为分类间隔。



样本到分类超平面H的距离r与g(x)关系：

 可得：

而 ，如下图所示



由于线性判别函数的值与  和b的尺度调整有关



之前我们只关心符号（用于分类），对上式计算的值大小不是很关心。在支持向量机的讨论中，为了获得唯一的最优解，我们需要把尺度确定下来。为此我们调整尺度，使得： ，有

最优解应使：

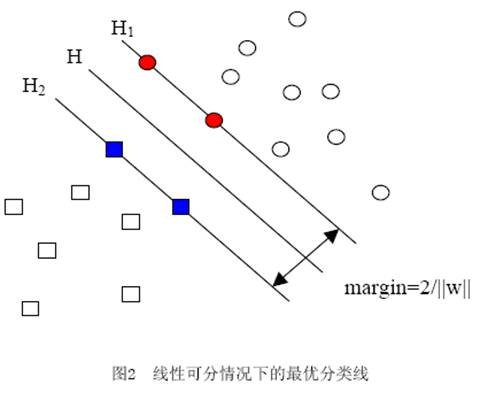




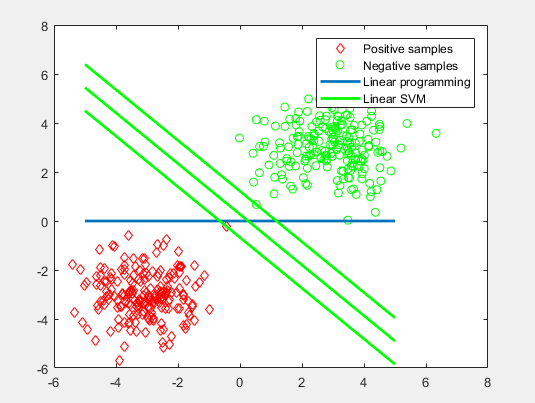


SVM从线性可分情况下的最优分类面发展而来。最优分类面就是要求分类线不但能将两类正确分开(训练错误率为0),且使分类间隔最大。SVM考虑寻找一个满足分类要求的超平面,并且使训练集中的点距离分类面尽可能的远,也就是寻找一个分类面使它两侧的空白区域(margin)最大。

过两类样本中离分类面最近的点且平行于最优分类面的超平面上H1,H2的训练样本就叫做支持向量。



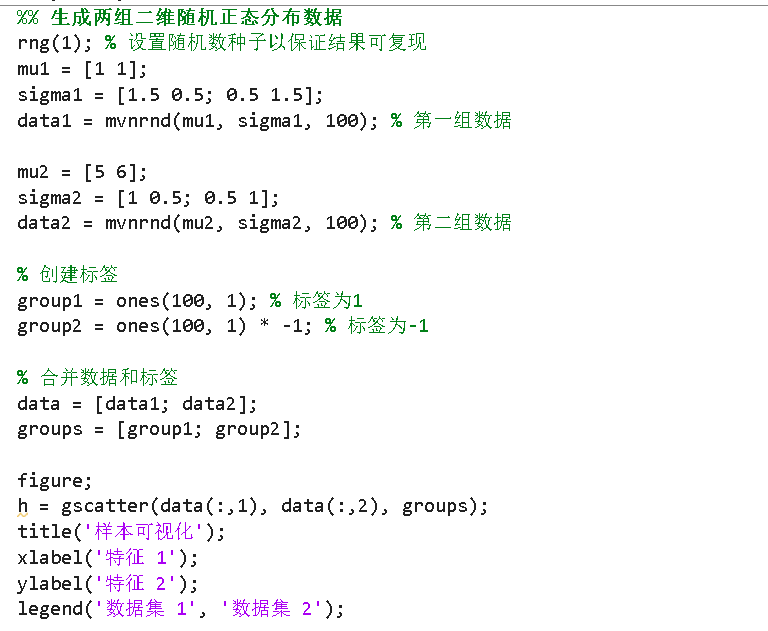
**3实验内容和要求**

1） 自行给出两组二维随机正态分布数组，要求利用二次规划函数进行求解,实现这两组样本的SVM分类，同时展示SVM分类面和最大间隔，可如下图所示：

2）直接利用MATLAB上的SVM工具包实现电离层数据集Ionosphere的分类。该数据集分两类，共351个样本。可将其中340个样本作为训练组，其余11个样本作为测试组，观察不同分组情形下的分类精度差异。同时观察不同核函数类型和参数选择对分类结果的影响。

**4实验代码和结果  
4.1 二次规划与SVM对比**

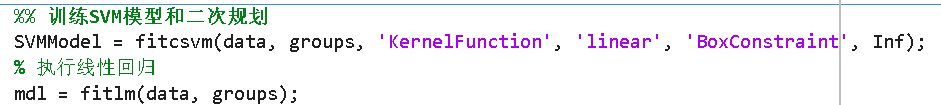
1. 生成两组随机二位样本





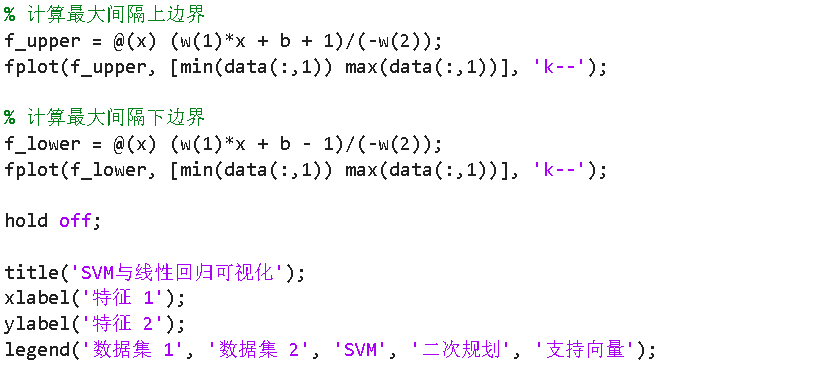
这段代码首先使用两种不同的二维正态分布生成了数据，每种分布包含了100个样本。这些样本以不同的均值和协方差矩阵生成，使它们在空间中呈现不同的分布特征。接着，通过绘制散点图，这些数据以两种不同的颜色展示在同一图中，用来表示两个不同的类别。这个图形化展示有助于直观地观察和理解这些数据在二维空间中的分布和区分情况。

1. 训练SVM和二次规划



1. 可视化SVM和二次规划





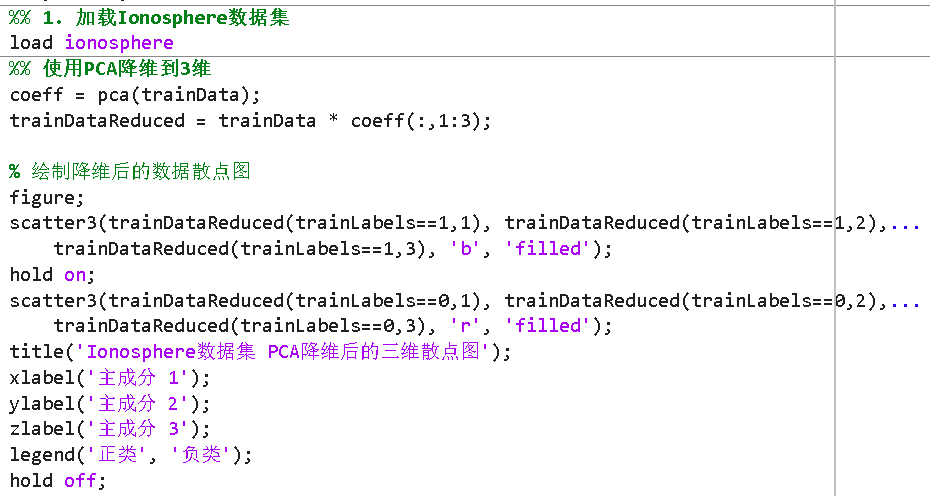
这段代码旨在展示支持向量机（SVM）模型和线性回归在二维数据集上的表现。首先，使用fitcsvm函数训练了SVM模型，采用线性核函数，并设置无穷大的Box约束，这样使得模型的边界没有限制。同时，利用fitlm执行了线性回归分析。接下来的操作是数据可视化和模型绘图。散点图展示了数据集，使用不同颜色标识了两个不同的类别。

在这个基础上，绘制了SVM模型的决策边界和最大间隔。决策边界通过SVM模型的参数计算得出，并以黑色虚线呈现。从SVM模型中提取线性回归系数，用紫色实线绘制了线性回归线。此外，通过黑色圆圈标记出支持向量，它们是在模型构建过程中起关键作用的数据点。最终，图形展示了数据集、SVM模型的决策边界、最大间隔、线性回归线和支持向量的关系，有助于比较和理解这些模型在数据集上的表现。



**4.2 SVM在数据集Ionosphere的表现**

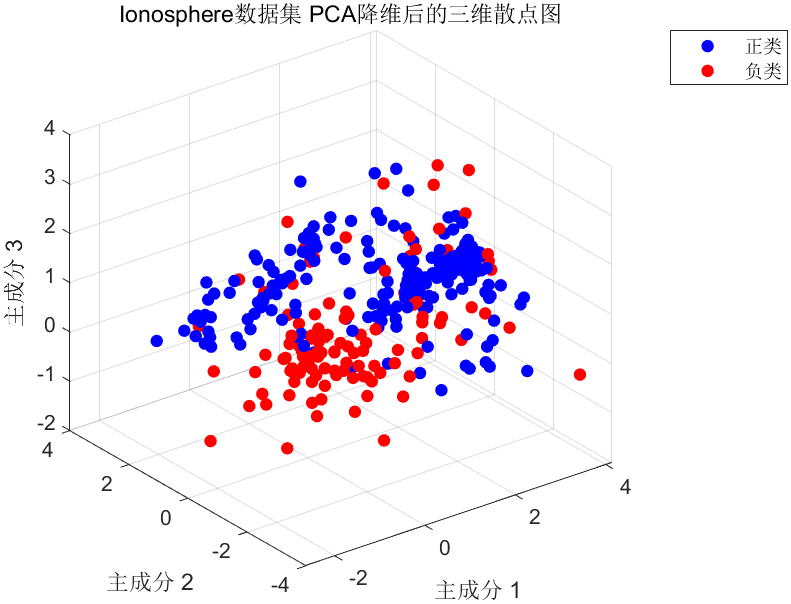
（1）读取lonosphere数据集并用PCA嵌入三维可视化



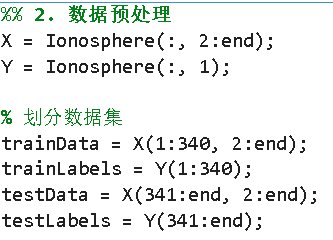
这段代码主要涉及Ionosphere数据集的加载和PCA降维到三维。首先，通过load ionosphere加载数据集，然后利用PCA（主成分分析）方法对训练数据进行降维，将数据降至三维空间。PCA计算的特征向量被用于将原始数据投影到新的三维空间。

接着，利用scatter3绘制了降维后的数据集散点图。散点图中的每个点代表一个样本，其中正类标记为蓝色，负类标记为红色。这个图形化展示有助于观察数据在降维后的三维空间中的分布情况，并区分了正类和负类。xlabel、ylabel和zlabel分别代表了三个主成分在三维空间中的坐标轴，而legend展示了正类和负类的标签说明。

这段代码类似之前提供的数据散点图展示，只不过这次是在降维后的三维空间中进行的展示。

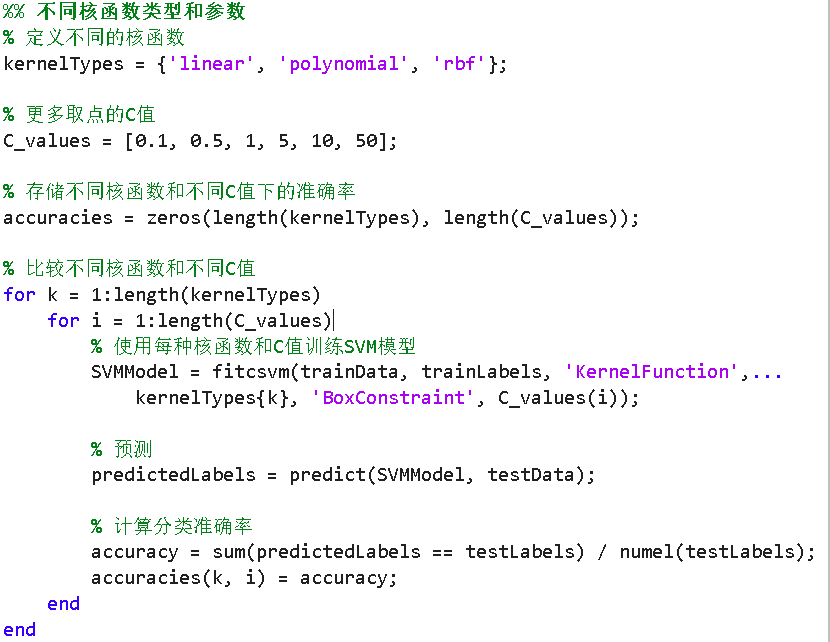


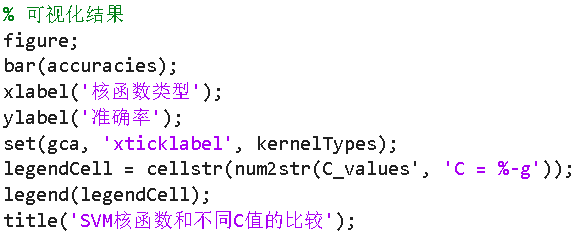
（2）数据预处理与数据集分割



（3）探究不同的核函数与不同的C值与模型性能的关系







这段代码旨在比较不同核函数类型和不同C值对SVM模型在Ionosphere数据集上的分类性能的影响。

首先，定义了不同的核函数类型，包括线性核函数（'linear'）、多项式核函数（'polynomial'）和径向基函数核（'rbf'）。然后，选择了更多不同取值的C值，以包括0.1、0.5、1、5、10和50。这些C值用于控制模型的正则化程度。

接着，初始化了一个矩阵accuracies，用于存储不同核函数和C值组合下的分类准确率。接下来，嵌套的循环用于比较不同核函数和C值的组合。在内部循环中，使用fitcsvm函数训练SVM模型，采用不同的核函数类型和C值。然后，对测试数据进行预测，计算分类准确率，并将其存储在accuracies矩阵中。

最后，通过绘制柱状图可视化结果，x轴表示核函数类型，y轴表示准确率。每个柱状图表示不同C值下的分类准确率，同时使用图例说明了每个柱状图对应的C值。这有助于比较不同核函数类型和C值对SVM模型性能的影响，以选择最佳的模型配置。

1. **实验心得**

对于第一部分，使用两种不同的二维正态分布生成的数据，这有助于直观理解数据集的特征和类别分布。通过可视化数据集，我们可以清晰地观察不同类别之间的分布情况，这对理解模型性能有很大帮助。SVM和二次规划的表现也在二维数据集上得到了很好的展示，直观展现了决策边界、最大间隔和支持向量。这让我们更好地理解了这些模型在数据集上的表现和区分能力。

对于Ionosphere数据集，使用PCA将数据降维到三维并进行可视化，有助于在更高维度中观察数据的分布情况。这种降维和可视化方法为理解数据集提供了一个全新的角度，尤其是对于高维数据的理解。此外，探索不同核函数类型和C值对SVM模型性能的影响，帮助我们选择最佳的模型配置。这种分析有助于理解模型对不同参数的敏感程度，以及如何调整这些参数以获得更好的性能。

总的来说，这些实验深化了对支持向量机（SVM）和二次规划在分类问题上的应用，并展现了数据可视化和参数调整对模型性能的重要性。理解不同模型和参数设置对结果的影响，是提高对机器学习模型理解和调优的关键一步。