svm

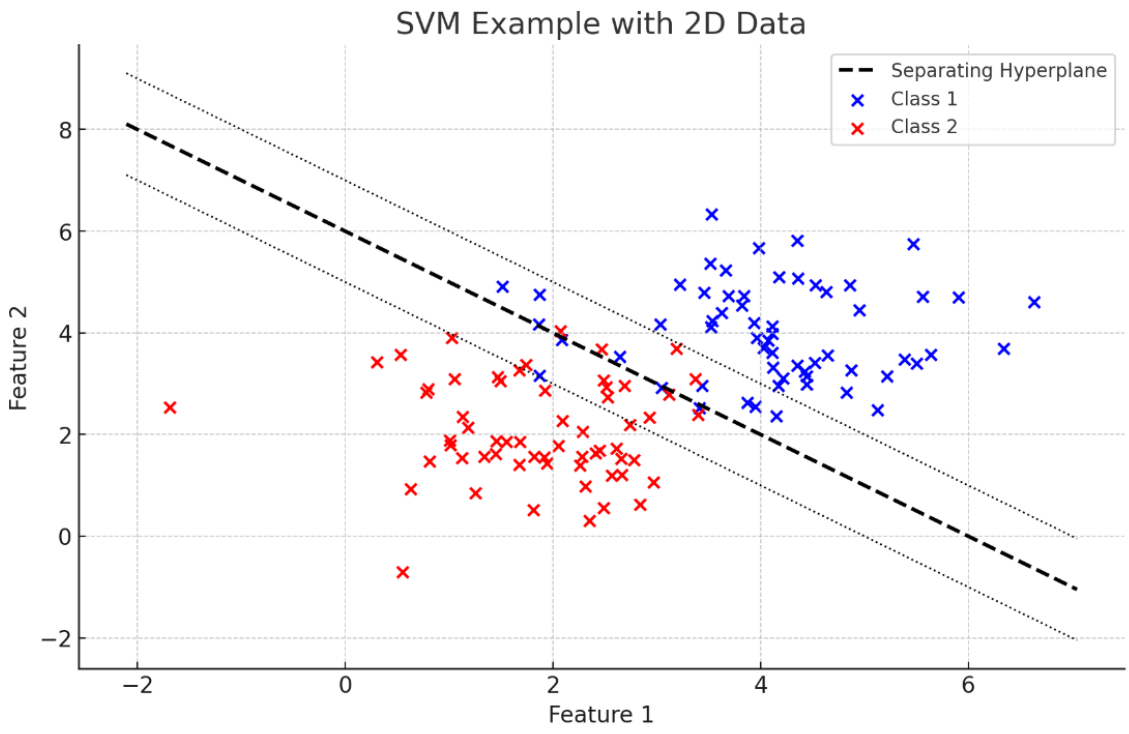
支持向量机（SVM，Support Vector Machines）是一种广泛使用的监督学习算法，主要用于分类问题。**SVM的基本思想是寻找一个最优的分割超平面，使得不同类别的样本之间的间隔最大化。**
在数学上，SVM的目标是求解以下优化问题：

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, \quad \forall i = 1, 2, \ldots, N \end{aligned}$$

其中：

- w是超平面的法向量。
- b 是超平面的截距。
- (x_i, y_i)是训练样本，其中 x_i是特征向量，y_i 是类别标签（通常取值为 −1或 1）。
- N 是训练样本的数量。

实例



在这个示例中，我们创建了一个简单的二维数据集，包含两个类别。蓝色点代表类别1，红色点代表类别2。**我们的目标是找到一个可以将这两个类别尽可能分开的直线，这条直线就是支持向量机（SVM）试图找到的最优超平面。**

黑色虚线代表了这个最优超平面，它**试图最大化两个类别之间的间隔。**间隔是两个类别最近点到超平面的距离，这些最近的点被称为支持向量，并且在图中以黑色圆圈标记出来。在SVM中，这些支持向量是关键的数据点，因为它们直接决定了最终的决策边界。

我们还画出了两条黑色点线，它们表示超平面两边的边际。在这个边际内，没有数据点，这样可以帮助模型有更好的泛化能力。SVM的目标就是最大化这个边际，从而提高模型在未知数据上的表现能力。

在这个例子中，超平面是一条直线（因为我们的数据是二维的），但在更高维的数据中，它可以是一个平面或者超平面。通过选择合适的核函数，SVM可以有效处理数据中的非线性关系，即使在原始特征空间中数据是线性不可分的。

核函数是用来在高维空间中计算数据点之间的相似性的一种方法。核函数的作用是能够在不显式计算高维空间中的点的情况下，通过在原始特征空间中计算内积，间接地计算出在高维空间中的内积。这允许SVM能够处理线性不可分的数据集。

支持向量机（SVM）的训练过程包括以下主要步骤：

1. **选择合适的核函数**：
   - 根据数据的分布和问题的性质选择合适的核函数。对于线性可分的数据，可以选择线性核。对于非线性问题，可以选择RBF、多项式或Sigmoid核。
2. **特征缩放**：
   - 对数据进行预处理，通常包括标准化或归一化，以确保所有特征都在相同的尺度上。这是因为SVM对特征的尺度敏感。
3. **构建优化问题**：
   - 根据选择的核函数和数据，构建一个优化问题，目标是最大化分类间隔，即找到最优超平面。
4. **求解优化问题**：
   - 使用序列最小优化（SMO）算法或其他数值优化算法来找出能够最大化间隔的模型参数，即支持向量。
5. **确定决策边界**：
   - 通过找到的支持向量和它们的系数确定决策边界，这是一个在特征空间中将不同类别分开的超平面。
6. **模型评估**：
   - 使用验证集或交叉验证来评估模型的性能。可以通过调整核函数参数、惩罚参数（C）等来优化性能。
7. **模型应用**：
   - 将训练好的模型应用于测试数据，使用学习到的支持向量和它们的系数来预测新数据的类别。