SVM and Nearest Neighbor Classifiers

21307099 李英骏

1. SVM

支持向量机（SVM）是主要用于分类和回归任务的有监督机器学习算法。主要思想是找到最佳将数据集划分为（两个）多个类别的超平面。

Principle:

给定训练集 $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, 其中 $\mathbf{x}_i$ 是特征向量，$y_i \in \{-1, 1\}$ 是其对应标签。SVM的目标是找到最佳的超平面来分隔数据，使得两个类之间的边距最大化。

超平面可表示为:
$\mathbf{w}^T \cdot \mathbf{x} + b = 0$

其中:

- $\mathbf{w}$ 是权重向量（垂直于超平面）
- b 是偏置
- $\mathbf{x}$ 是输入数据点

分类的决策函数为:
$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \cdot \mathbf{x} + b)$

为了最大化距离，SVM旨在最小化 $\|\mathbf{w}\|$
在以下约束下:
$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall i$

Soft Margin:

真实数据可能是非线性的，并且可能存在噪声。为了处理这种情况，SVM引入了所谓“软间隔”，允许一些误差在确定最大间隔的超平面时。为实现这一点，我们引入了松弛变量 $\xi_i \geq 0$

优化目标变为：
$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$

在以下约束下:
$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$

其中, C 是一个常数，用来控制松弛程度。较大的 C 值意味着更小的容错率。

Kernel Trick:

在数据不可线性分隔的情况下，SVM使用核函数将数据映射到一个在其中它变得线性可分隔的高维空间。常用核函数:

- 线性: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- 多项式: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + c)^d$
- 径向基函数 (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

2. Nearest Neighbor Classifiers

最近邻分类器是基于实例的学习模型。思想是根据训练数据中的最近邻点的标签来对新数据点进行分类。

Principle:

训练数据集 $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ 和新数据点 $\mathbf{x}$, 最近邻分类器将训练数据集中 $\mathbf{x}$ 的最近点的标签分配给 $\mathbf{x}$ 。

常用的距离度量是欧氏距离等，定义为:
$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$
其中 m 是特征的数量。

k-Nearest Neighbors (k-NN):

数据点 $\mathbf{x}$ 的分类由其k个最近邻点的主要标签决定。k 可能会影响分类器的性能。

Weighted k-NN:

邻居对分类决策的影响由其到 $\mathbf{x}$ 的距离加权。更近的邻居有更大的$power$。

$\text{weight}(\mathbf{x}_i) = \frac{1}{d(\mathbf{x}_i, \mathbf{x})^2}$