1 核方法: 近邻法与支持向量机

1.1 密度估计的非参数法

对于随机变量 $\boldsymbol{x}$, 如果知道其概率密度函数 $p(\boldsymbol{x})$, 就能够计算出其在某一区域 R 内取值的概率. 现在, 我们需要根据 $\boldsymbol{x}$ 的一组采样点 $\mathcal{D}:\{\boldsymbol{x}_n\}_{n=1}^N$ 估计概率密度函数 $p(\boldsymbol{x})$. 这就是**密度估计 (density estimation)** 问题.

> **定义 1.1 密度估计** 给定随机变量 $\boldsymbol{x}$ 的一组采样点 $\mathcal{D}:\{\boldsymbol{x}_n\}_{n=1}^N$, 密度估计是根据 $\mathcal{D}$ 估计随机变量 $\boldsymbol{x}$ 的概率密度函数 $p(\boldsymbol{x})$ 的过程.

密度估计有两类办法:

(1) 参数法: 假定 $p(\boldsymbol{x})$ 具有已知的带参数形式, 那么问题转化为应用最大似然法确定参数的值.

(2) 非参数法: 不对 $p(\boldsymbol{x})$ 作任何假设, 而是直接根据样本 $\mathcal{D}$ 估计 $p(\boldsymbol{x})$.

本节介绍的直方图方法, 核密度估计法和近邻法都属于非参数法.

1.1.1 直方图方法

直方图是我们熟知的表示随机变量分布的方法. 将数据空间划分为若干个小区域, 称为**区间 (bins)** 或**箱 (buckets)**, 然后统计每个区间内的样本点数目, 最后通过归一化得到概率密度函数的估计. 具体地, 设数据空间被划分为 M 个区间 $\{\mathcal{R}_j\}_{j=1}^M$, 每个区间的体积为 V, 则在区间 $\mathcal{R}_j$ 内的概率密度函数估计为

$$p_{\mathcal{R}_j}(\boldsymbol{x}) = \frac{1}{NV}\sum_{n=1}^N \mathbb{I}(x_n \in \mathcal{R}_j) = \frac{n_j}{NV}$$

其中 n_j 即 $\mathcal{R}_j$ 中的样本点的数目.

直方图的数学原理可以推导如下.

证明. 根据概率密度函数的定义, 随机变量 $\boldsymbol{x}$ 落在某一区域 $\mathcal{R}$ 内的概率为

$$P = \int_{\mathcal{R}} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

对 $\boldsymbol{x}$ 随机采样 N 次, 有 K 个点落入 $\mathcal{R}$ 的概率服从二项分布:

$$p\left(K|N,P\right) = C_N^K P^K (1-P)^{N-K}$$

如果 N 和 K 都很大, 那么上述二项分布是一个窄峰, 我们就可以近似地认为 $P \approx \dfrac{K}{N}$. 另一方面, 如果区域 $\mathcal{R}$ 足够小, 那么可以认为在该区域内 $p(\boldsymbol{x})$ 近似为常数, 即 $P = p(\boldsymbol{x})V$. 结合上述两式即可得

$$p(\boldsymbol{x}) = \frac{K}{NV}$$

这就说明了直方图方法的合理性. □

直方图方法的优点是简单直观, 但缺点也很明显: 结果曲线不光滑, 并且高维空间下将因维度灾难而效果变差.

1.1.2 核密度估计法

1.1.3 近邻法

1.2 核方法的主要思想

1.3 支持向量机

支持向量机通过核方法进行非线性分类. 它的主要想法是: 在所有可能的分类超平面中, 选择一个使得分类间隔最大的超平面作为最终的分类超平面.

1.3.1 支持向量机的数学原理

我们用更严谨的语言描述上述问题. 考虑线性分类模型

$$y(\boldsymbol{x}) = \boldsymbol{w}$$