PART2 KNN 模型

2020-06-1

河北师范大学 软件学院

基本内容

1. KNN 是什么的简称？此处的 K 值是什么意思？
2. 掌握基于 KNN 近邻法进行分类或回归的实现步骤。
3. KNN 近邻决策时，那些因素会影响决策结果？
4. 如何面向分类或回归任务，采用 m-折交叉验证的方式进行 K 值优选？你是如何评价每个备选 K 值的？
5. 若采用 KNN 法进行两类别的分类，K 值的设定会有哪些考虑？

练习题

1. 给定来自三种类别花型的训练样本集 $\boldsymbol{D}=\left\{\left(\boldsymbol{x}_i, y_i\right), \mathrm{i}=1, \cdots, N\right\}$，其中每个样本的输入部分分别由四种特征（如：花瓣长、花瓣宽、花萼长、花萼宽）描述，并且 $\boldsymbol{y}_i \in\{1,2,3\}$，请采用 $\boldsymbol{K}$ 近邻法，对任意观测样本 $\boldsymbol{x} \in R^4$ 的类别 y 进行预测（要求：为确保取得尽可能好的分类性能，在描述你的实现步骤中尽量体现你的处理细节）

解：

STEP1. 首先规范化预处理训练样本集$\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$的输入部分。

估计 $\begin{cases}\mu^{(j)}=\dfrac{1}{N}\sum_{i=1}^{N}x_i^{(j)},\ j=1,2,3,4\\ \sigma^{(j)}=\sqrt{\dfrac{1}{N}\sum_{i=1}^{N}\left(x_i^{(j)}-\mu^{(j)}\right)^2},\ j=1,2,3,4\end{cases}$ 并保存

对于$(\boldsymbol{x}_i,y_i)\in\boldsymbol{D}$，$x_i^{(j)}\Leftarrow\dfrac{x_i^{(j)}-\mu^{(j)}}{\sigma^{(j)}}$ $j=1,2,3,4$

STEP2. 基于欧式距离度量，并采用m-fold CV(m折交叉验证)方式选择K

尝试着描述一下这个优选的过程，你会采用什么指标来评价每个备选的K？

STEP3. 对样本$\boldsymbol{x}=\left[x^{(1)}\cdots x^{(4)}\right]$预处理：$x^{(j)}\Leftarrow\dfrac{x^{(j)}-\mu^{(j)}}{\sigma^{(j)}}$ $j=1,2,3,4$

并基于预处理的训练集$\boldsymbol{D}$内找到**K个近邻，记为**$N_K(\boldsymbol{x})$

$N_K(\boldsymbol{x})=N_{K,\omega_1}(\boldsymbol{x})\cup\cdots\cup N_{K,\omega_C}(\boldsymbol{x})$

STEP4. 结合指定的**分类规则**，预测$\boldsymbol{x}$的类别y.

$$\hat{y}=\arg\max_{j\in\{1,2,\ldots,C\}}\left|N_{K,\omega_j}(\boldsymbol{x})\right|$$

注意：可将$\dfrac{\left|N_{K,\omega_j}(\boldsymbol{x})\right|}{K}$视为$\boldsymbol{x}$关于$\omega_j$类的后验概率。

2. 给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$，$y_i\in\{1,2\}$. 请采用 *K* 近邻法，对任意观测样本 $x\in R^d$ 的类别 y 进行预测？此时关于 *K* 值的取值你是如何考虑的？（要求：为确保取得尽可能好的分类性能，在描述你的实现步骤中尽量体现你的处理细节）

提示：对于两类别的分类，K 值应为奇数。

3. 给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$，针对如下两种情况，采用 *K* 近邻法，分别对任意观测 $x\in R^d$ 产生的输出 y 进行预测：

(1) $y_i\in\{1,2,\ldots,C\}$；　**(2)** $y_i\in R$

（要求：为确保取得尽可能好的预测性能，在描述你的实现步骤中尽量体现你的处理细节）

哪些因素会影响基于 *K* 近邻的决策结果？