PART2 KNN 模型

2023.06

基本内容

1. KNN 的英文全称？K 的意义？
2. KNN 近邻决策时，哪些因素会影响模型的决策性能？
3. 若采用 KNN 法进行两类别的分类，K 值的设定会有哪些考虑？
4. 掌握基于 KNN 近邻法进行分类的完整实现流程。
5. 掌握基于 KNN 近邻法进行回归的完整实现流程。
6. 如何采用 m-折交叉验证的方式面向分类任务进行 K 值优选。你是如何评价每个备选 K 值的？
7. 如何采用 m-折交叉验证的方式面向回归任务进行 K 值优选。你是如何评价每个备选 K 值的？

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 $\boldsymbol{y}_i \in\{1,2,3\}$，若采用“欧式距离”度量样本之间差异，并按照等权投票法决策，请给出基于 $\boldsymbol{K}$ 近邻法对任意观测样本 $\boldsymbol{x} \in R^d$ 的类别 y 进行预测的完整流程。（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处理细节）

解：

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

估计各特征的均值与标准差 $\begin{cases}\mu^{(j)}=\dfrac{1}{N}\sum\limits_{i=1}^{N}x_i^{(j)}, j=1,2,3,4\\ \sigma^{(j)}=\sqrt{\dfrac{1}{N}\sum\limits_{i=1}^{N}\left(x_i^{(j)}-\mu^{(j)}\right)^2}, j=1,2,3,4\end{cases}$ 并保存

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

STEP2. 采用欧式距离度量，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. 对输入样本$\boldsymbol{x}=\left[x^{(1)}\cdots x^{(4)}\right]$进行预处理： $x^{(j)} \Leftarrow \dfrac{x^{(j)}-\mu^{(j)}}{\sigma^{(j)}}$ $\quad j=1,2,3,4$

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

$$N_K(\boldsymbol{x})=N_{K,1}(\boldsymbol{x})\cup N_{K,2}(\boldsymbol{x})\cup N_{K,3}(\boldsymbol{x})$$

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

$$\hat{y}=\underset{j\in\{1,2,3\}}{\arg\max}\left|N_{K,\mathrm{j}}(\boldsymbol{x})\right|$$

注意：可将$\dfrac{\left|N_{K,j}(\boldsymbol{x})\right|}{K}$视为$\boldsymbol{x}$关于第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** $\boldsymbol{x}\in R^d$ **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：要有完整的实现流程；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** $\boldsymbol{x}\in R^d$ **产生的输出** y **进行预测，要求给出完整的实现流程。**

(1) $y_i\in\{1,2,...,C\}$；

(2) $y_i\in R$

哪些因素会影响基于 ***K*** **近邻的决策结果？**

4. K-近邻回归. 给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$， $y_i\in R$，请完成如下工作：

(1)若采用等权投票法决策，对 K-近邻回归模型进行学习，并对任意观测 $\boldsymbol{x}\in R^d$ 产生的输出 y 进行预测，给出完整的实现流程；

(2)哪些因素会影响基于 K-近邻回归模型的决策性能？

解：

(1)

STEP1. 训练集$\boldsymbol{D}$内各样本特征取值的预处理，并记录每种特征取值预处理的使用参数。
预处理参数计算？如何预处理？

STEP2. 明确$\begin{cases}\textbf{距离度量}\text{为欧式距离}\\ \textbf{输出预测规则}\text{为等权平均}\end{cases}$，在此基础上，采用交叉验证方式，选择近邻数**K**值的大小。

STEP3. 对输入样本$\boldsymbol{x}$，基于上述记录的参数，进行同样方式的预处理；
在此基础上，在预处理的训练集$\boldsymbol{D}$内找到它的前**K个近邻，记为**$N_K(\boldsymbol{x})$

STEP4. 结合指定的**输出预测规则(**等权平均**)**，对$\boldsymbol{x}$的输出y进行预测
即：该近邻中各样本目标答案的均值即为该样本的预测输出。

(2)各样本是否进行预处理、选择何种预测规则、采用何种距离度量、是否进行 K 值选择将直接影响模型预测性能。

5. K-近邻分类. 给定训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i),\mathrm{i}=1,\cdots,N\}$，其中 $\boldsymbol{x}_i\in R^d$，$y_i\in\{1,2,...,C\}$，请完成如下工作：

(1)若采用等权投票法决策，对 K-近邻分类模型进行学习，并对任意观测 $\boldsymbol{x}\in R^d$ 产生的输出 y 进行预测，给出完整的实现流程；

(2)哪些因素会影响基于 K-近邻分类的决策性能？

答：(1)

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

预处理训练集： $\tilde{x}_i^{(j)}\Leftarrow\dfrac{x_i^{(j)}-\mu^{(j)}}{\sigma^{(j)}}\qquad j=1,\ldots,d$

将预处理的数据集记为$\widehat{D}=\{(\tilde{x}_i,y_i),i=1,\ldots,N\}$

STEP2. 明确等权投票的决策规则，**基于欧式距离度量，**
并采用交叉验证方式选择K

STEP3. 对样本***x***预处理： $\tilde{x}^{(j)} \Leftarrow \dfrac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}} \quad j = 1,\dots,d$

并在预处理的训练集内找到该样本的前**K个近邻**

STEP4. 找到K个近邻的出现次数最多的类别，作为该样本的预测输出。

(2)样本的输入部分是否进行预处理、选择何种预测规则、采用何种距离度量、是否进行 K 值选择。