Machine learning: Homework #2

5 18 2020

Problem2

动手实现 Naïve Bayes 算法。本次实验使用一个较小的英文数据集 (50 个实例) 和中文数据集“The TREC 2006 Chinese Public Corpus – 60MB (trec06c.tgz)”其中有 60000+ 多个邮件，垃圾邮件和正常邮件的比例为 2：1。由于因为编码原因有许多邮件可能采用了混合编码（非 gb2312），所以一些邮件打开之后出现非常生僻的中文字符。在本次实验中可以直接忽略了这些文件，所以剩下的有用文件一共有 51069 个。这也是总的数据集大小。请参考 index 文件确定垃圾邮件路径。

同学们也可以自己寻找一些大小合适（10000 以上）的数据集。

数据处理要求

要求对英文邮件进行简单停用词处理。

要求对中文邮件进行分词处理，简单的停用词处理。有能力的同学可以进一步清洗数据。

参数估计

要求从以下三种参数估计方法中至少选择一种：

1. 离散特征:

$$\left[\hat{\theta}_{jc}\right]_{\alpha}=\frac{\sum_{i=1}^{n} I\left(y_{i}=c\right) I\left(x_{i\alpha}=j\right)+l}{\sum_{i=1}^{n} I\left(y_{i}=c\right)+l K_{\alpha}}$$

2. 多项式特征:

$$\hat{\theta}_{\alpha c}=\frac{\sum_{i=1}^{n} I\left(y_{i}=c\right) x_{i\alpha}+l}{\sum_{i=1}^{n} I\left(y_{i}=c\right) m_{i}+l \cdot d}$$

3. 连续特征:

$$\mu_{\alpha c} \leftarrow \frac{1}{n_{c}} \sum_{i=1}^{n} I(y_{i}=c) x_{i\alpha} \quad \left(n_{c}=\sum_{i=1}^{n} I(y_{i}=c)\right)$$

$$\sigma_{\alpha c}^{2} \leftarrow \frac{1}{n_{c}} \sum_{i=1}^{n} I\left(y_{i}=c\right)\left(x_{i\alpha}-\mu_{\alpha c}\right)^{2}$$

最终结果汇总在一个文档（pdf/doc/docx/ppt etc.）中即可，待检查。实验报告模板已上传。