模式识别第六次作业

孙浩森 201928013229100

Email: sunhm15@gmail.com

1 Adaboost 算法的设计思想

核心思想 从弱学习算法出发, 反复学习, 得到一系列弱分类器; 然后组合这些弱分类器, 构成一个强分类器。

基本做法 改变训练数据的概率 (权重) 分布, 提高那些被前一轮弱分类器分错的样本的权重, 降低已经被正确分类的样本的权重, 从而使得错分的样本在下一轮弱分类器中得到更多关注。针对不同的训练数据的分布, 调用弱学习算法来学习一系列分类器。

组合方法 采用加权 (多数) 表决的方法, 加大分类错误率较小的弱分类器的权重, 使其在表决中起更大的作用。

2 模型选择的基本原则

- 不存在一个与具体应用无关的、普遍适用的“最优分类器”
- 不存在与问题无关的最优的特征/属性集合。世界上不存在分类的客观标准, 一切分类的标准都是主观的。
- 简单有效原理, 如果对训练数据分类的效果相同, “简单的”分类器往往优于“复杂的”分类器。
- 我们应该选择尽可能简单的分类器或模型

3 分类器集成的基本方法

- 通过处理训练数据 (bagging, boosting), 比如, 对训练样本进行随机分组, 对错分样本进行加权。
- 通过处理特征, 比如, 每次只选择一部分特征来训练分类器
- 通过处理类别标号, 比如, 对多类问题, 一对一策略、一对多策略。
- 通过改进学习方法, 比如, 变更学习参数 (如多核学习)、模型结构 (如神经网络结构) 等

4 Hard-Margin SVM 的优化目标

优化目标为最大化最小间距，即，

$$\begin{aligned}&\arg\max_{w,b}(w,b,D)\\=&\arg\max_{w,b}\arg\min_{x_i\in D}d(x_i|w,b)\\=&\arg\max_{w,b}\arg\min_{x_i\in D}\frac{|b+x_iw_i|}{\sqrt{\sum_{i=1}^{d}w_i^2}}\end{aligned}$$

此时的任务为

$$\begin{aligned}&\arg\max_{w,b}\arg\min_{x_i\in D}\frac{|b+x_iw_i|}{\sqrt{\sum_{i=1}^{d}w_i^2}}\\s.t.\quad&\forall x_i\in D:y_i(x_iw+b)\geq 0\end{aligned}$$

此时可以对 w 和 b 进行归一化，使得 $|b+x_iw_i|$ 最小值变为 1，该问题可以变为

$$\begin{aligned}&\arg\min_{w}\sum_{i=1}^{d}w_i^2\\s.t.\quad&\forall x_i\in D:y_i(x_iw+b)\geq 1\end{aligned}$$

5 Hinge Loss 在 SVM 中的意义

hinge loss 可以用来表示 svm 的损失函数，其形式如下，

$$\sum_{i=1}^{n}[1-y_i(wx_i+b)]_+ +\lambda\|w\|^2$$

第一项是损失，第二项是正则化项。这个公式就是说 $y_i(wx_i+b)$ 大于 1 时 loss 为 0，否则 loss 为 $1-y_i(wx_i+b)$。对比感知机的损失函数 $[-y_i(wx_i+b)]_+$ 来说，hinge loss 不仅要分类正确，而且置信度足够高 (即 margin 足够大) 的时候，损失才为 0，对学习有更高的要求。

6 编程题

代码见 svm.py, 选择类别 2 和 5 两个类别，C=1，若用线性核函数，准确率为 0.969。若用多项式核函数，准确率为 0.997。若选用 RBF 核作为核函数，准确率可以达到 0.999。