机器学习笔记–随机森林和 *Bagging*

空修菜

1 装袋（bagging）

1.1 Bootstrap 和 Aggreation

1. “**bagging**”是“bootstrap aggreation”的简写.
2. Bootstrap 是对训练集进行抽样的方法. X 是训练集, 每次从样本中抽取 $|X|$ 个样本, 允许重复. 即有可能某个样本点被多次重复抽到, 而另外的样本点从未被抽到过.
3. 每当得到一个新样本, 就进行一次训练, 得到一个新模型.
4. Bootstrap 可以在一定程度上处理数据集不平衡的问题.
5. 通过 Bootstrap 得到的新数据集 $Z_i, i = 1, 2, \ldots, M$ 都有一个与之对应的模型（estimator）$G_i(\cdot)$. 此时，定义 **aggregate predictor** G 为:

$$G(X) = \sum_{i=1}^{M} \frac{G_i(x)}{M}.$$

   - 利用式子 (**??**), 有

$$\mathrm{Var}(\bar{G}) = \mathrm{Var}(G) = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2.$$

   - 从上式可知, 当增加模型数量的时候, 第二项分母增大, 所以模型的方差 $\mathrm{Var}(G)$ 就降低了, 因此过大的方差就被降低了.

6. bagging 加上决策树就得到了随机森林. 随机森林的“随机”体现在:
   - feature 的选择是随机的. 若有 p 个特征, 随机选取的特征数为 $\lfloor\sqrt{p}\rfloor$, 这就使得所训练的模型间的相关系数 ρ 变小;
   - 训练集的选择是随机的 (bootstrap).

7. 在 bagging 中,还有一个被称为"**out-of bag estimation**"的优点. bootstrap 得到的数据是随机抽取的，可以证明被抽到的样本点占到整个训练集 S 的 $\frac{2}{3}$, 所以，剩余的 $\frac{1}{3}$ 样本点就可以用来估计模型的误差，这个误差称为"袋外误差"(out of bag error).

1.2 随机森林（Random Forest）

1. 对于决策树来说, 存在偏差低, 方差高, 预测效果差的情况;

2. 对训练集 S 进行 bootstrap 后分别得到 m 个新的训练集 $Z_i, 1 \leq i \leq m$, 分别对每一个训练集 Z_i 训练一颗决策树 T_i, 总共就有 m 棵决策树. 这样就得到了一个森林.

3. 得到决策树集合 $\{T_i : i = 1, 2, \ldots, m\}$ 后, 最后的预测函数为 T:

$$T(x) = \frac{1}{m} \sum_{i=1}^{m} T_i(x).$$

4. 随机森林与决策树不同的体现在节点的分类标准的选择上. 对于每一棵树 T_i，在选择 feature 以及切分点点时候，并不全部考虑 p 个 features, 而是"随机地不重复"选取 p 个 feature 中的 t 个进行考虑 $(t < p)$, 一般地，取 $t = \lfloor \sqrt{p} \rfloor$(取小于等于 $\sqrt{p}$ 的最大整数).

5. 模型优点:

   - 降低了模型误差
   - 提高了准确性
   - 有额外的验证集 (set out of bag)

6. 模型缺点:

   - 偏差 bias 增加
   - 模型复杂更难解释
   - 耗时更长