精简版机器学习知识

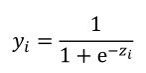
**感知机**：二分类的线性分类模型。学习算法是基于随机梯度下降法的对损失函数的最优化算法，即一次随机选取一个误分类点使其梯度下降（使分离超平面向该误分类点的一侧移动，以减少该误分类点与超平面间的距离，直至被正确分类）。算法由于采用不同的w、b初值或不同的迭代顺序，解可以不同。仅当训练数据集线性可分时，学习算法是收敛的。

KNN：基本分类与回归方法，可用于多分类。K值的选择（交叉验证选择最优的K），距离度量（欧氏距离），分类决策规则（多数表决）是K近邻法的三个基本要素。为了提高K近邻搜索的效率，构造kd树。Kd树是二叉树，搜素被限制在空间的局部区域内，效率大为提高。

缺点：预测精度一般，对于大部分问题比不上基于树的方法。当样本存在范围重叠时，K近邻的分类精度很低

**朴素贝叶斯：**基于贝叶斯定理与特征条件独立假设的分类方法，属于生成模型。将后验概率最大的类作为x的类输出（等价于期望风险最小化），先验概率和条件概率计算时分子分母分别加上λ=1和kλ，避免所要估计的概率为0，称为拉普拉斯平滑。优点：过程简单且速度较快，因为它的预测过程实际就是进行概率的乘积。对于多分类问题也很有效，计算复杂度不会有大幅度的上升缺点：属性个数比较多或者属性之间相关性较大时，分类效果不好。

**SVM：**二类分类模型，在特征空间上的间隔最大使它有别于感知机，还包括核技巧，使它成为实质上的非线性分类器。学习策略就是间隔最大化，等价于正则化的合页损失函数最小化问题。最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾难”。 SVM模型具有较好的泛化能力，因为其本身的优化目标是结构风险最小，而不是经验风险最小。缺点：１. 训练时间长　2. 模型预测时，预测时间与支持向量的个数成正比。当支持向量的数量较大时，预测计算复杂度较高。3. 不方便解决多分类问题

**LOGISTIC:**实际上，“分类”是应用逻辑回归的目的和结果，但中间过程依旧是“回归”。因为通过逻辑回归模型，我们得到的计算结果是0-1之间的连续数字，可以把它称为“可能性”（概率）。然后，给这个可能性加一个阈值，就成了分类。例如，算出贷款违约的可能性>0.5，将借款人预判为坏客户.逻辑斯蒂回归得到的是概率值，解决的其实是二分类问题(可推广用于多分类).回归模型就是将线性回归的结果输入一个Sigmoid函数，将回归值映射到0~1，表示输出为类别的概率。Sigmoid函数：

优点：１.直接对分类的概率建模，无需实现假设数据分布，从而避免了假设分布不准确带来的问题。2.不仅可预测出类别，还能得到该预测的概率，这对一些利用概率辅助决策的任务很有用；3.对数几率函数是任意阶可导的凸函数，有许多数值优化算法都可以求出最优解。（梯度下降法，牛顿法）

缺点：１.Logistic回归模型本质上还是一种线性模型，只能做线性分类，不适合处理非线性的情况，一般需要结合较多的人工特征处理使用。2.Logistic回归对正负样本的分布比较敏感，所以要注意样本的平衡性

**决策树：**决策树是一种树状结构模型，可以进行基本的分类与回归，是集成方法经常采用的基模，主要涉及三要素，分别是**特征选择、决策树的生成和决策树的剪枝**。从最初的没有信息，到最后的百分百确信，这是一个减少不确定性的过程。使用决策树决策的过程就是我们不断减少信息熵的过程，直到它降为0。我们希望信息熵能下降得快一点。这就涉及到决策树的构建了，我们该怎么构建才能使得这棵决策树的信息熵在判断分支中下降得最快呢？CART 决策树用于分类任务：采用**基尼系数**作为特征选择标准，**Gini指数越小，说明数据集越集中**，纯度越高。所以选择的是切分之后Gini指数尽量小的切分方法

CART决策树用于回归任务：采用平方误差最小化准则进行特征选择，把最终各个叶子中所有样本对应结果值的均值或者中位数当作预测的结果输出  
CART分类树与回归树采用二叉树来对每一个特征进行划分

生成算法：在所有可能的特征A以及他们所有可能的切分点a中，选择基尼指数最小的特征及其对应的切分点作为最优特征与最优切分点。

优点：时间复杂度较小，适合高维数据，决策树算法可处理连续和种类字段，:挖掘出来的分类规则准确性高, 便于理解, 决策树可以清晰的显示哪些字段比较重要, 即可以生成可以理解的规则

缺点：对噪声比较敏感，容易过拟合，在训练集噪声较大时得到的模型容易过拟合，处理特征关联性较强的数据时表现不太好。

**随机森林：**采用CART作为基学习器，而且它在Bagging模型的基础上再进一步，每次训练基学习器时，除对样本随机采样外，对样本的特征也进行随机采样，树模型本身具有较强的非线性拟合能力。

优点：１.RF各个基学习器之间没有强关联，因此学习过程可以分开，实现并行化　２. RF每次训练时，各个基学习器只是抽取样本的部分特征进行训练，因此对于样本特征维度很高的情况，RF仍能高效地训练模型　３. RF的每个基学习器只是随机抽取部分样本和部分特征进行学习，因此模型的泛化能力较强　４. 采用的基学习器是CART决策树，而CART决策树对缺失值不敏感，因此RF对部分特征缺失也不敏感。５. RF训练模型后可以顺便输出各个特征对预测结果的重要性，因此可以辅助我们进行特征选择。  
缺点：１. 决策树模型的一个缺点是对**噪声**比较敏感，所以RF在某些噪声较大的样本集上容易陷入过拟合

**XGBOOST**

XGBoost（eXtreme Gradient Boosting）是GBDT的一种改进形式，具有很好的性能，XGBoost采用的是解析解思维，即对损失函数进行二阶泰勒展开（传统的GBDT/GBRT模型只用到了损失函数的一阶导信息（一阶泰勒展开）），求得解析解，然后用这个解析解作为“增益”来辅助建立 CART 回归树，最终使得整体损失达到最优。

优点：１.XGBoost 支持自定义损失函数，只要满足定义的损失函数二阶可导即可，这大大增加了处理问题的灵活性。２.在GBDT的基础上加入了一个正则化项，用于控制模型的复杂度，正则化项里面包含了树的叶子节点数和各个叶子节点输出值的平方之和。从方差—偏差角度来看，正则化项可以降低模型的方差，使学习出来的模型更加简单，防止模型过拟合。３.XGBoost 模型借鉴了随机森林的做法，支持对特征进行抽样，这也可以起到降低过拟合风险和减少计算量的作用。４. 当遇到一个负“增益”时，GBDT/GBRT 会马上停止分裂，但 XGBoost会一直分裂到指定的最大深度，然后回过头来剪枝。如果某个节点之后不再有负值，则会除掉这个分裂；但是如果负值后面又出现正值，并且最后综合起来还是正值，则该分裂会被保留。５. XGBoost 对于不同节点遇到的特征缺失将采用不同的处理方式，并且会逐渐学习出处理缺失值的方式，当后面再遇到有缺失特征时就可以按学习出的处理方式进行处理，这样更加科学。６. XGBoost支持并行：XGBoost的并行不是在tree粒度上，而是在特征粒度上。xgboost在训练之前，预先对数据进行了排序，然后保存为block结构，后面的迭代中重复地使用这个结构，大大减小计算量。这个block结构也使得并行成为了可能，在进行节点的分裂时，需要计算每个特征的增益，最终选增益最大的那个特征去做分裂，那么各个特征的增益计算就可以开多线程进行。

LightGBM 在很多方面会比 XGBoost 表现的更为优秀，比如更快的训练效率、低内存使用、更高的准确率、支持并行化学习、可处理大规模数据、支持直接使用类别特征等。其在原理上与GBDT算法和XGBoost算法类似，都采用损失函数的负梯度作为当前决策树的残差近似值，去拟合新的决策树。