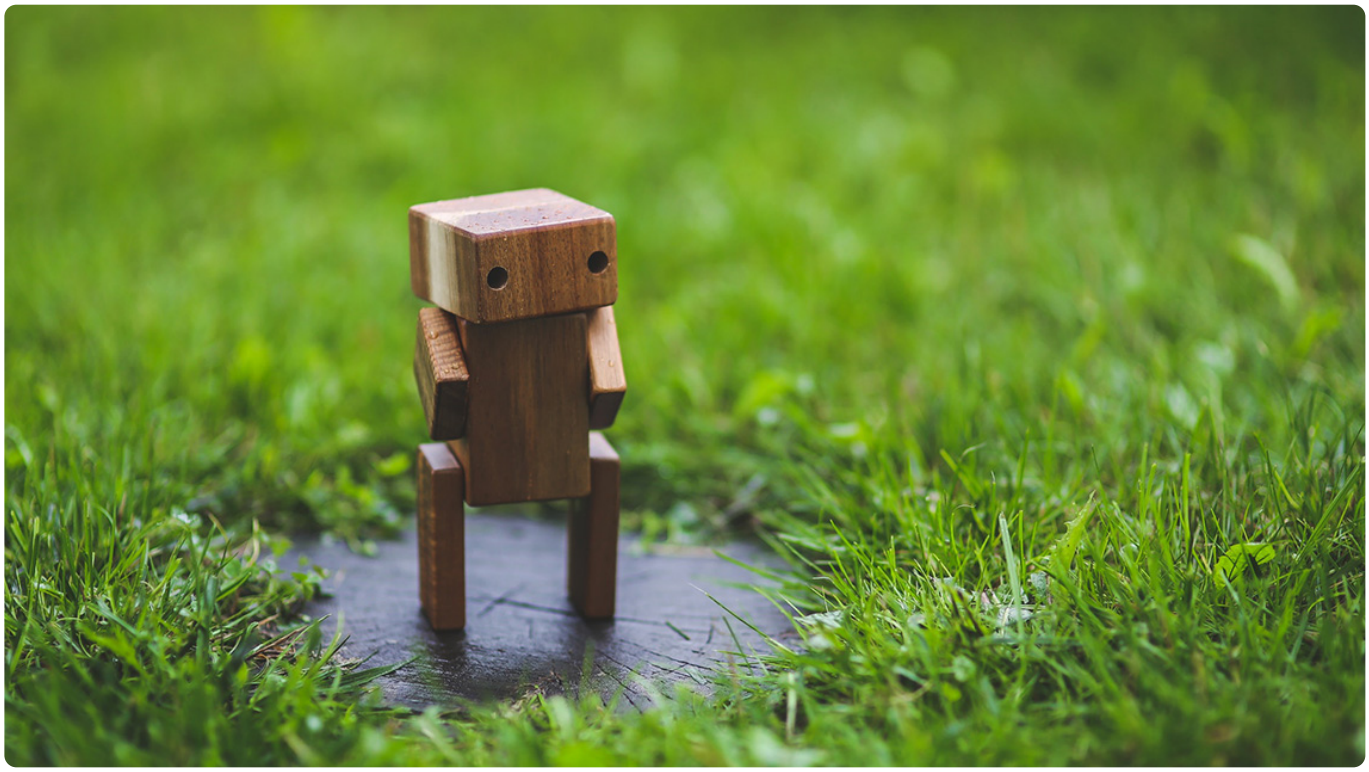
040 | 机器学习排序算法经典模型：GBDT

2018-01-03 洪亮劼

AI技术内参

进入课程 >



讲述：初明明
时长 06:55　大小 3.17M

这周我们讨论机器学习排序算法中几个经典的模型，周一分享了配对法排序中的一个经典算法，即排序支持向量机（RankSVM），这个算法的核心思想是把支持向量机应用到有序数据中，试图对数据间的顺序进行直接建模。

今天，我们来聊一聊利用机器学习进行排序的一个重要算法：**“梯度增强决策树”（Gradient Boosted Decision Tree）**。长期以来，包括雅虎在内的很多商业搜索引擎都利用这种算法作为排序算法。

梯度增强决策树的历史

梯度回归决策树的思想来源于两个地方。首先是“**增强算法**”（Boosting），一种试图用弱学习器提升为强学习器的算法。这种算法中比较成熟的、有代表性的算法是由罗伯特·施

派尔（Robert Schapire）和约阿夫·福伦德（Yoav Freund）所提出的**AdaBoost 算法**[1]。因为这个算法两人于 2003 年获得理论计算机界的重要奖项“哥德尔奖”（Gödel Prize）。罗伯特之前在普林斯顿大学任计算机系教授，目前在微软研究院的纽约实验室工作。约阿夫一直在加州大学圣地亚哥分校任计算机系教授。

增强算法的工作机制都比较类似，那就是先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，使得先前基学习器做错的训练样本在后续受到更多关注，然后基于调整后的样本分布来训练下一个基学习器。如此重复进行，直到基学习器数目达到事先制定的值，最终将所有的基学习器进行加权结合。如果你对“偏差 - 方差分解”（Bias-Variance Decomposition）有耳闻的话，那么，Boosting 主要关注降低偏差。**在实际效果中，增强算法往往能基于泛化性能相当弱的学习器构建出很强的集成结果。**

AdaBoost 提出后不久，机器学习学者和统计学家杰罗姆·弗赖德曼（Jerome H. Friedman）等人发表了一篇论文 [2]，**从“统计视角”解释 AdaBoost 实质上是基于加性模型（Additive Model）以类似牛顿迭代法来优化指数损失函数（Loss Function）**。于是受此启发，杰米姆提出了“**梯度增强**”（Gradient Boosting）的想法。这也就是梯度回归决策树思想来源的第二个地方，也是直接根源。如果你希望对“梯度增强”有进一步的了解，可以见参考文献 [3]。

最早把“梯度增强”的想法应用到搜索中，是雅虎研究院的学者于 2007 年左右提出的 [4]&[5]。之后，Facebook 把梯度增强决策树应用于新闻推荐中 [6]。

梯度增强的思想核心

我们刚才简单讲了增强算法的思路，那么要想理解梯度增强决策树，就必须理解梯度增强的想法。

梯度增强首先还是增强算法的一个扩展，也是希望能用一系列的弱学习器来达到一个强学习器的效果，从而逼近目标变量的值，也就是我们常说的标签值。而根据加性模型的假设，这种逼近效果是这些弱学习器的一个加权平均。也就是说，最终的预测效果，是所有单个弱学习器的一个平均效果，只不过这个平均不是简单的平均，而是一个加权的效果。

那么如何来构造这些弱学习器和加权平均的权重呢?

梯度增强采用了一个统计学或者说是优化理论的视角，使得构造这些部分变得更加直观。

梯度增强的作者们意识到，如果使用“梯度下降”（Gradient Descent）来优化一个目标函数，最后的预测式可以写成一个加和的形式。也就是，每一轮梯度的值和一个叫“学习速率”（Learning Rate）的参数共同叠加起来形成了最后的预测结果。这个观察非常重要，如果把这个观察和我们的目标，也就是构造弱学习器的加权平均联系起来看，我们就会发现，**其实每个梯度的值就可以认为是一个弱学习器，而学习速率就可以看作是某种意义上的权重。**

有了这个思路，梯度增强的算法就很容易构造了。

首先，这是一个迭代算法。每一轮迭代，我们把当前所有学习器的加权平均结果当作这一轮的函数值，然后求得针对某一个损失函数对于当前所有学习器的参数的一个梯度。**然后，我们利用某一个弱学习器算法，可以是线性回归模型（Linear Regression）、对数几率模型（Logistic Regression）等来拟合这个梯度。最后，我们利用“线查找”（Line Search）的方式找到权重。**说得更直白一些，那就是我们尝试利用一些简单的模型来拟合不同迭代轮数的梯度。

梯度增强的一个特点就是梯度下降本身带来的，那就是每一轮迭代一定是去拟合比上一轮小的一个梯度，函数对目标的整体拟合也是越来越好的。这其实也就是增强算法和梯度下降的一个完美结合。

梯度增强决策树以及在搜索的应用

理解了梯度增强，那么梯度增强决策树也就容易理解了。简单来说，**梯度增强决策树就是利用决策树，这种最基本的学习器来当作弱学习器，去拟合梯度增强过程中的梯度。**然后融合到整个梯度增强的过程中，最终，梯度增强决策树其实就是每一轮迭代都拟合一个新的决策树用来表达当前的梯度，然后跟前面已经有的决策树进行叠加。在整个过程中，决策树的形状，比如有多少层、总共有多少节点等，都是可以调整的或者学习的**超参数**。而总共有多少棵决策树，也就是有多少轮迭代是重要的**调节参数**，也是防止整个学习过程过拟合的重要手段。

参考文献 [5] 和 [6]，就是雅虎的科学家第一次把刚才提到的这个思路用于搜索问题中，训练排序算法。在应用的时候，有一些细节的调整，比如**损失函数的设定**。这里，作者们采用了配对法排序学习方法，那就是不直接去拟合相关度，而是拟合两个不同文档相关度的差值。具体来说，就是针对某一个查询关键字，我们利用算法来最小化对文档相关度差值的预测，也就是说我们不希望把相关度高的文档放到相关度低的后面。

在这些论文中，还有后续的很多研究中，利用梯度增强决策树算法进行排序训练得到的效果比当时的其他算法都有大幅度的提升。因此，这也就慢慢地奠定了**梯度增强决策树作为一种普适的机器学习排序算法**的地位。值得说明的是，梯度增强决策树的成功，一部分来自于增强算法，另一部分来自于把很多决策树堆积起来的想法。这两个思路都是在机器学习中被反复验证、行之有效的“模式”。

小结

今天我为你讲了梯度增强决策树算法的基本原理，这是一个利用机器学习技术来学习排序的基础算法。作为配对法排序学习的一个经典算法，梯度增强决策树算法有着广泛的应用 。一起来回顾下要点：第一，我们简要介绍了梯度增强决策树提出的历史。第二，我们详细介绍了增强算法的核心以及梯度增强的思路。第三，我们简要介绍了梯度增强决策树的核心以及如何利用这个算法来训练排序问题。

最后，给你留一个思考题，梯度增强的思路能和神经网络模型结合吗?

欢迎你给我留言，和我一起讨论。

参考文献

1. Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 55, 1 (August 1997), 119-139, 1997.
2. Friedman, Jerome; Hastie, Trevor; Tibshirani, Robert. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann. Statist. 28 (2000), no. 2, 337--407, 2000.
3. Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of Statistics (2001): 1189–1232, 2001.
4. Zhaohui Zheng, Hongyuan Zha, Tong Zhang, Olivier Chapelle, Keke Chen, and Gordon Sun. A general boosting method and its application to learning ranking functions for web search. Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07), J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.). Curran Associates Inc., USA, 1697-1704, 2007.
5. Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. A regression framework for learning ranking functions using relative relevance judgments. Proceedings of the 30th annual international ACM SIGIR conference on Research

and development in information retrieval (SIGIR '07). ACM, New York, NY, USA, 287-294, 2007.
6. Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. Practical Lessons from Predicting Clicks on Ads at Facebook. Proceedings of the Eighth International Workshop on Data Mining for Online Advertising (ADKDD'14). ACM, New York, NY, USA, , Article 5 , 9 pages, 2014.

论文链接

A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting

Additive logistic regression: a statistical view of boosting

Greedy function approximation: a gradient boosting machine

A general boosting method and its application to learning ranking functions for web search

A regression framework for learning ranking functions using relative relevance judgments

Practical Lessons from Predicting Clicks on Ads at Facebook

上一篇　039 | 机器学习排序算法经典模型：RankSVM

下一篇　041 | 机器学习排序算法经典模型：LambdaMART

精选留言 (2)

写留言

黄德平
2018-12-13

残差网络估计是受到GBDT的启发

展开

范深
2018-01-17

神经网络与增强梯度最简单的结合，就是把多个神经网络作为弱分类器串联起来？我相信还有更妙的结合点：）