

Modify's Programming

[博客园](#)[首页](#)[新随笔](#)[联系](#)[订阅](#)[管理](#)

随笔 - 12 文章 - 0 评论 - 160 阅读 - 22万

本博客所有内容以学习、研究和分享为主，如需转载，请联系本人，标明作者和出处，并且是非商业用途，谢谢！

Email:rongshunlin@126.com

微信公众号: ModifyAI

 Flag Counter

昵称: ModifyBlog

园龄: 6年8个月

粉丝: 300

关注: 3

[+加关注](#)

逻辑回归的常见面试点总结

1.简介

逻辑回归是面试当中非常喜欢问到的一个机器学习算法，因为表面上看逻辑回归形式上很简单，很好掌握，但是一问起来就容易懵逼。所以在面试的时候给大家的第一个建议不要说自己精通逻辑回归，非常容易被问倒，从而减分。下面总结了一些平常我在作为面试官面试别人和被别人面试的时候，经常遇到的一些问题。

2.正式介绍

如何凸显你是一个对逻辑回归已经非常了解的人呢。那就是用一句话概括它！**逻辑回归假设数据服从伯努利分布,通过极大化似然函数的方法，运用梯度下降来求解参数，来达到将数据二分类的目的。**

这里面其实包含了5个点 1：逻辑回归的假设，2：逻辑回归的损失函数，3：逻辑回归的求解方法，4：逻辑回归的目的，5:逻辑回归如何分类。这些问题是考核你对逻辑回归的基本了解。

• 逻辑回归的基本假设

- 任何模型都是有自己的假设，在这个假设下模型才是适用的。逻辑回归的**第一个基本假设是假设数据服从伯努利分布**。伯努利分布有一个简单的例子是抛硬币，抛中为正面的概率是 p ，抛中为负面的概率是 $1 - p$ 。在逻辑回归这个模型里面是假设 $h_{\theta}(x)$ 为样本为正的的概率， $1 - h_{\theta}(x)$ 为样本为负的概率。那么整个模型可以描述为

$$h_{\theta}(x; \theta) = p$$

- 逻辑回归的第二个假设是假设样本为正的的概率是

< 2021年4月 >

$$p = \frac{1}{1 + e^{-\theta^T x}}$$

- 所以逻辑回归的最终形式

$$h_{\theta}(x; \theta) = \frac{1}{1 + e^{-\theta^T x}}$$

• 逻辑回归的损失函数

- 逻辑回归的损失函数是它的极大似然函数

$$L_{\theta}(x) = \prod_{i=1}^m h_{\theta}(x^i; \theta)^{y^i} * (1 - h_{\theta}(x^i; \theta))^{1-y^i}$$

• 逻辑回归的求解方法

- 由于该极大似然函数无法直接求解，我们一般通过对该函数进行梯度下降来不断逼近最优解。在这个地方其实会有个加分的项，考察你对其他优化方法的了解。因为就梯度下降本身来看的话就有随机梯度下降，批梯度下降，small batch 梯度下降三种方式，面试官可能会问这三种方式的优劣以及如何选择最合适的梯度下降方式。
 - 简单来说 批梯度下降会获得全局最优解，缺点是在更新每个参数的时候需要遍历所有的数据，计算量会很大，并且会有很多的冗余计算，导致的结果是当数据量大的时候，每个参数的更新都会很慢。
 - 随机梯度下降是以高方差频繁更新，优点是使得sgd会跳到新的和潜在更好的局部最优解，缺点是使得收敛到局部最优解的过程更加的复杂。
 - 小批量梯度下降结合了sgd和batch gd的优点，每次更新的时候使用n个样本。减少了参数更新的次数，可以达到更加稳定收敛结果，一般在深度学习当中我们采用这种方法。
- 其实这里还有一个隐藏的更加深的加分项，看你了不了解诸如Adam，动量法等优化方法。因为上述方法其实还有两个致命的问题。
 - 第一个是如何对模型选择合适的学习率。自始至终保持同样的学习率其实不太合适。因为一开始参数刚刚开始学习的时候，此时的参数和最优解隔的比较远，需要保持一个较大的学习率尽快逼近最优解。但是学习到后面的时候，参数和最优解已经隔的比较近了，你还保持最初的学习率，容易越过最优点，在最优点附近来回振荡，通俗一点说，就很容易学过头了，跑偏了。

日	一	二	三	四	五	六
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

搜索

 找找看

 谷歌搜索

常用链接

[我的随笔](#)
[我的评论](#)
[我的参与](#)
[最新评论](#)
[我的标签](#)

随笔分类

[机器学习\(3\)](#)
[自然语言处理\(4\)](#)

随笔档案

2019年9月(2)

2019年8月(1)

2019年7月(2)

2019年1月(1)

2018年5月(1)

2018年1月(1)

2017年11月(2)

2017年10月(2)

最新评论

1. Re:机器学习算法GBDT的面试要点总结-上篇

代码试了一下与预期不符。
计算损失函数L的时候没有在内存循环重置L为0，需要在代码66行上面加一个L=0。

--rulcy

2. Re:TextCNN 代码详解(附测试数据集以及GitHub 地址)

- 第二个是如何对参数选择合适的学习率。在实践中，对每个参数都保持的同样的学习率也是很合理的。有些参数更新频繁，那么学习率可以适当小一点。有些参数更新缓慢，那么学习率就应该大一点。这里我们不展开，有空我会专门出一个专题介绍。

- **逻辑回归的目的**

- 该函数的目的便是将数据二分类，提高准确率。

- **逻辑回归如何分类**

- 逻辑回归作为一个回归(也就是y值是连续的)，如何应用到分类上去呢。y值确实是一个连续的变量。逻辑回归的做法是划定一个阈值，y值大于这个阈值的是一类，y值小于这个阈值的是另外一类。阈值具体如何调整根据实际情况选择。一般会选择0.5做为阈值来划分。

3.对逻辑回归的进一步提问

逻辑回归虽然从形式上非常的简单，但是其内涵是非常的丰富。有很多问题是可以进行思考的

- **逻辑回归的损失函数为什么要使用极大似然函数作为损失函数？**

- 损失函数一般有四种，平方损失函数，对数损失函数，HingeLoss0-1损失函数，绝对值损失函数。将极大似然函数取对数以后等同于对数损失函数。在逻辑回归这个模型下，对数损失函数的训练求解参数的速度是比较快的。至于原因大家可以求出这个式子的梯度更新

$$\theta_j = \theta_j - (y^i - h_{\theta}(x^i; \theta)) * x_j^i$$

这个式子的更新速度只和 x_j^i ， y^i 相关。和sigmoid函数本身的梯度是无关的。这样更新的速度是可以自始至终都比较的稳定。

- 为什么不选平方损失函数的呢？其一是因为如果你使用平方损失函数，你会发现梯度更新的速度和sigmoid函数本身的梯度是很相关的。sigmoid函数在它在定义域内的梯度都不大于0.25。这样训练会非常的慢。

- **逻辑回归在训练的过程当中，如果有很多的特征高度相关或者说有一个特征重复了100遍，会造成怎样的影响？**

- 先说结论，如果在损失函数最终收敛的情况下，其实就算有很多特征高度相关也不会影响分类器的效果。
- 但是对特征本身来说的话，假设只有一个特征，在不考虑采样的情况下，你现在将它重复100遍。训练以后完以后，数据还是这么多，但是这个特征本身重复了100遍，实质上将原来的特征分成了100份，每一个特征都

感谢博主，介绍地很详细，有系统。
我想问下，这个可以修改为自己的数据再跑吗？
我看了下训练数据“vocal”没有文件后缀，打开后乱码~但是测试集里可以打开，类似txt

--sunfish21

3. Re:机器学习算法GBDT的面试要点总结-上篇

求下篇啊

--黎明咬破夜的唇

4. Re:逻辑回归的常见面试点总结

“逻辑回归的第一个基本假设是假设数据服从伯努利分布。”，这应该是二分类的吧，如果是多分类是什么假设呢

--yyl424525

5. Re:机器学习算法GBDT的面试要点总结-上篇

楼主你好，请问你这里使用GBDT处理分类问题时，使用的损失函数还是MSE？为什么不是log-loss呢？

--我得冲向图书馆了

是原来特征权重值的百分之一。

- 如果在随机采样的情况下，其实训练收敛完以后，还是可以认为这100个特征和原来那一个特征扮演的效果一样，只是可能中间很多特征的值正负相消了。
- **为什么我们还是会在训练的过程当中将高度相关的特征去掉？**
 - 去掉高度相关的特征会让模型的可解释性更好
 - 可以大大提高训练的速度。如果模型当中有很多特征高度相关的话，就算损失函数本身收敛了，但实际上参数是没有收敛的，这样会拉低训练的速度。其次是特征多了，本身就会增大训练的时间。

4.逻辑回归的优缺点总结

面试的时候，别人也经常问到，你在使用逻辑回归的时候有哪些感受。觉得它有哪些优缺点。

在这里我们总结了逻辑回归应用到工业界当中一些优点：

- 形式简单，模型的可解释性非常好。从特征的权重可以看到不同的特征对最后结果的影响，某个特征的权重值比较高，那么这个特征最后对结果的影响会比较大。
- 模型效果不错。在工程上是可以接受的（作为baseline），如果特征工程做的好，效果不会太差，并且特征工程可以大家并行开发，大大加快开发的速度。
- 训练速度较快。分类的时候，计算量仅仅只和特征的数目相关。并且逻辑回归的分布式优化sgd发展比较成熟，训练的速度可以通过堆机器进一步提高，这样我们可以在短时间内迭代好几个版本的模型。
- 资源占用小,尤其是内存。因为只需要存储各个维度的特征值，。
- 方便输出结果调整。逻辑回归可以很方便的得到最后的分类结果，因为输出的是每个样本的概率分数，我们可以很容易的对这些概率分数进行cutoff，也就是划分阈值(大于某个阈值的是一类，小于某个阈值的是一类)。

但是逻辑回归本身也有许多的缺点：

- 准确率并不是很高。因为形式非常的简单(非常类似线性模型)，很难去拟合数据的真实分布。
- 很难处理数据不平衡的问题。举个例子：如果我们对于一个正负样本非常不平衡的问题比如正负样本比 1000 0:1.我们把所有样本都预测为正也能使损失函数的值比较小。但是作为一个分类器，它对正负样本的区分能力不会很好。

阅读排行榜

1. 机器学习算法GBDT的面试要点总结-上篇(112017)
2. 逻辑回归的常见面试点总结(42233)
3. 【原创】文本分类算法TextCNN原理详解（一）(18092)
4. 教你如何最快入门用户画像(14455)
5. 使用LSTM和Softmx来进行意图识别(9339)

评论排行榜

1. 机器学习算法GBDT的面试要点总结-上篇(44)
2. 如何两个月刷400道leetcode(33)
3. 2018年年终总结(25)
4. 教你如何最快入门用户画像(16)
5. 逻辑回归的常见面试点总结(16)

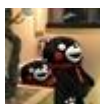
推荐排行榜

- 处理非线性数据较麻烦。逻辑回归在不引入其他方法的情况下，只能处理线性可分的数据，或者说，处理二分类的问题。
- 逻辑回归本身无法筛选特征。有时候，我们会用gbdt来筛选特征，然后再上逻辑回归。

5.总结

在这篇博文里面，我总结了在日常面试的过程当中逻辑回归可能会遇到的问题，欢迎大家在下面留言和我进行交流，将你们面试遇到的关于逻辑回归的问题发出来，我会尽力去解答的。

分类: [机器学习](#)

[好文要顶](#)[关注我](#)[收藏该文](#)

ModifyBlog

关注 - 3

粉丝 - 300

[+加关注](#)

13

[推荐](#)

2

[反对](#)

« 上一篇: [使用paddle来进行文本生成](#)

» 下一篇: [机器学习算法GBDT的面试要点总结-上篇](#)

posted @ 2017-10-27 09:29 ModifyBlog 阅读(42237) 评论(16) 编辑 收藏

[刷新评论](#) [刷新页面](#) [返回顶部](#)

登录后才能查看或发表评论，立即 [登录](#) 或者 [逛逛](#) [博客园首页](#)

园子动态:

- 致园友们的一封检讨书：都是我们的错
- 数据库实例 CPU 100% 引发全站故障
- 发起一个开源项目：博客引擎 fluss

最新新闻:

- 杀死八种癌症！中国要有自己的抗癌“神药”了？
- 烧光84亿造不出一台量产车！拜腾德国子公司已进入临时破产管理状态