机器学习入门：模型评估指标

原创 开发中心-杨晨 锡银电子化 2019-05-17

作者：杨晨，就职于软件开发中心渠道组，对建模及调优有一定经验与理论基础。

引言

我行于2018年底成功地将机器学习与百度助贷项目结合，实现了第一个人工智能场景的落地。由于机器学习入门难，理论性强，向他人解释训练好的模型便是个麻烦事儿。本文将为您解除这样的烦恼。

1 概述

当前形式

近几年各界对人工智能的兴趣激增，自2011年以来，开发与人工智能相关的产品和技术并使之商业化的公司已获得超过总计20亿美元的风险投资，而科技巨头更是投资数十亿美元收购那些人工智能初创公司。人工智能领域的相关报道铺天盖地。前两天，特朗普签署行政命令，旨在推动美国在人工智能(AI)领域的发展，同时削弱中国在这一领域的强劲势头。 可见人工智能的研究竞争已经进入白热化阶段，已经上升到国家竞争的层面。

项目背景

我行对前沿科技始终保持着高敏锐度，在这科技变革的大时代里，争作科技变革的引领者。在去年年底的百度助贷项目中，成功将机器学习应用于贷前准入，在降低不良率的同时，提升了贷款准入的通过率。体现出了机器学习的巨大应用价值。然而，当模型建好之后，如何体现模型的价值，如何向业务人员解释模型的优越性对于模型开发者来讲是个头疼的问题。

2
原理

模型训练好，必须要通过各种指标去衡量模型的好坏，也就是模型的泛化能力。模型的评估指标有很多，笔者在刚开始学习的时候，也是焦头烂额，有时候自己理解了，但又很难跟别人解释清楚，本节将以理论知识结合图片的形式，详细介绍分类模型的各种评估指标以及ROC和AUC值。

01

混淆矩阵

对于二分类的模型，预测结果与实际结果分别可以取0和1。我们用N(negative)和P(positive)代替0和1，T(true)和F(false)表示预测正确和错误。将他们两两组合，就形成了下图所示的混淆矩阵（注意：组合结果都是针对预测结果而言的）。

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图1.混淆矩阵

则图中的四种结果解释为：

(a) TP：预测为1，实际为1，即预测正确；

(b) FP：预测为1，实际为0，即预测错误；

(c) FN：预测为0，实际为1，即预测错确；

(d) TN：预测为0，实际为0，即预测正确；

02

准确率

准确率的定义是预测正确的结果占总样本的百分比。

公式：准确率=(TP+TN)/(TP+TN+FP+FN)

即，绿色部分和/ （ 绿色部分和+ 红色部分和 ）

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图2.准确率

实际应用场景中，由于样本不平衡的问题，导致了得到的高准确率结果含有很大的水分。即如果样本不平衡，准确率就会失效。这样就衍生出了另外两个指标：**精准率和召回率。**

03

精准率

精准率（Precision）又叫查准率，是指在所有被预测为正的样本中实际为正的样本的概率。

公式：精准率=TP/(TP+FP)

即，绿色部分/ （ 绿色部分+ 红色部分）

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图3.精准率

04

召回率

召回率（Recall）又叫查全率，是指在实际为正的样本中被预测为正样本的概率。

公式：召回率=TP/(TP+FN)

即，绿色部分/ (绿色部分+ 红色部分)

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图4.召回率

以信用卡逾期为背景，召回率越高，代表实际逾期用户被预测出来的概率越高，它的含义类似：宁可错杀一千，绝不放过一个。所以召回率的提高，往往意味着精准率的下降。

05

灵敏度、特异度、真正率、假正率

灵敏度(Sensitivity) = TP/(TP+FN)，即实际为正样本预测成正样本的概率

特异度(Specificity) = TN/(FP+TN)，即实际为负样本预测成负样本的概率

真正率(TPR) = 灵敏度= TP/(TP+FN)，即实际为正样本预测成正样本的概率

假正率(FPR) = 1- 特异度= FP/(FP+TN)，即实际为负样本预测成正样本的概率

不难看出：

召回率 = 灵敏度= 查全率= 真正率= TPR = TP/(TP+FN)

它们都是指实际正样本中预测为正样本的概率。

灵敏度、真正率：绿色部分/ (绿色部分+ 红色部分)

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图5.灵敏度、真正率

(1-特异度)、假正率：绿色部分/ (绿色部分+ 红色部分)

		实际	
		1	0
预测	1	TP	FP
	0	FN	TN

图6.特异度、假正率

真正率和假正率这两个指标跟正负样本的比例是无关的。所以当样本比例失衡的情况下，准确率不如这两个指标好用。

06

ROC曲线

ROC（Receiver Operating Characteristic）曲线，又称接受者操作特征曲线。ROC曲线的横坐标是假阳性比值（假正率），纵坐标是真阳性比值（真正率）。

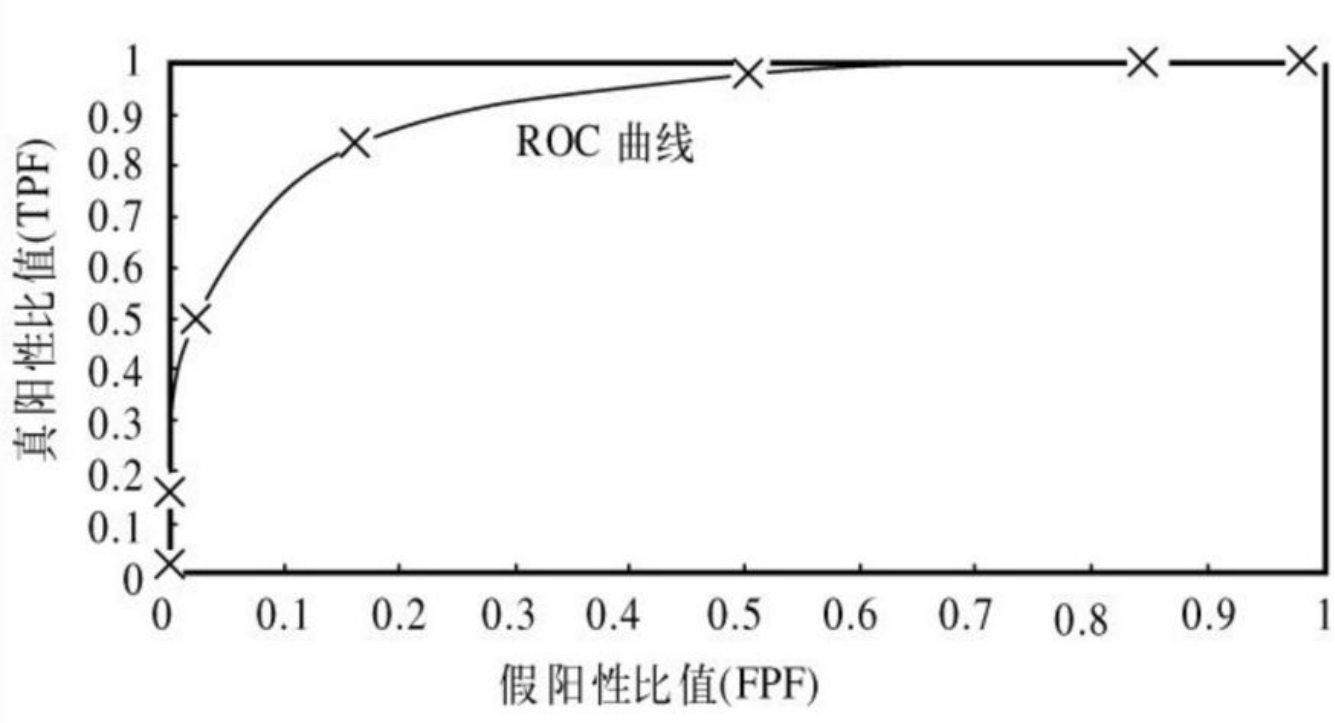

图7.ROC曲线

假正率反应了模型虚报的响应程度，真正率反应了模型预测响应的覆盖程度。所以我们希望，假正率越小，真正率越高越好，即虚报的少，覆盖的多。也就是说，TPR越高，FPR越低，模型就越好。反应到ROC图形上，也就是取现越陡峭，越朝着左上方突出，模型效果越好。

07

AUC值

AUC是基于ROC曲线的，被称为曲线下面积（Area Under Curve）。

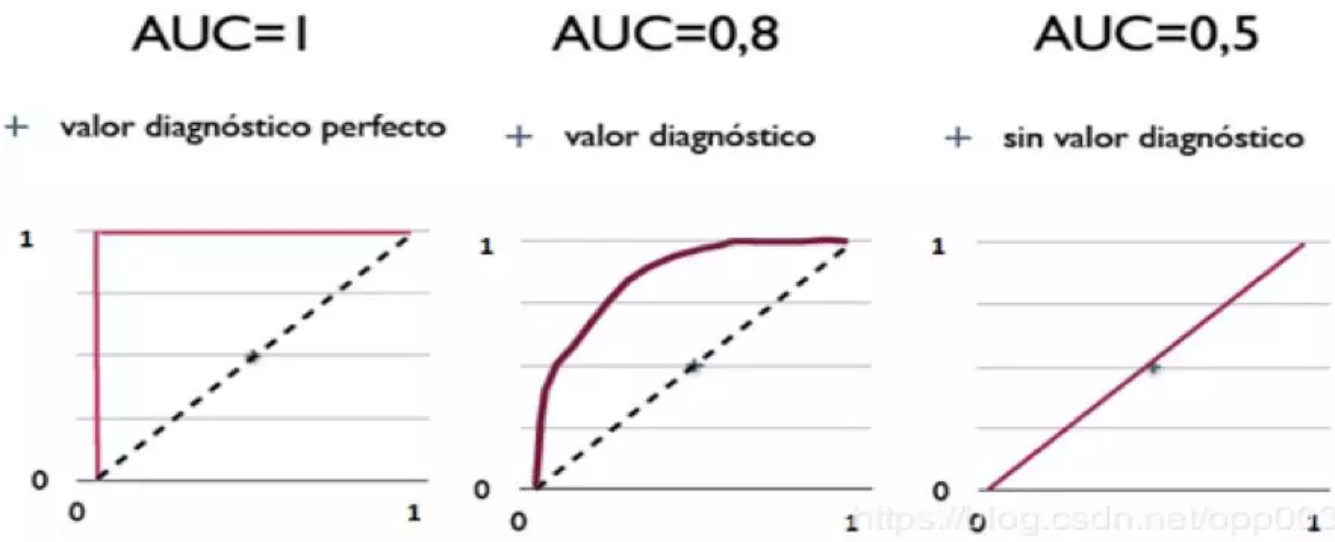

图8.不同AUC值效果展示

如上图所示，在ROC曲线图上，如果我们连接对角线，它的面积正好是0.5。对角线的实际含义是：随机判断响应与不响应，正负样本覆盖率应该都是50%，表示随机效果。ROC曲线越陡越好，所以理想值就是1，一个正方形，而最差的随机判断都有0.5，所以一般AUC的值是介于0.5到1之间的。

3
项目

百度助贷项目中，我们将机器学习算法应用于贷前准入规则中。旨在降低贷款不良率，同时能够让更多的客户准入。在经过业务访谈、特征筛选、特征工程、数据处理等环节后，我们采用GBDT算法建模，得到最终的模型。

模型虽然得到了，但是作为开发人员，你必须向业务人员证明，为什么你的模型会比专家规则好。经过上一节的理论介绍后，我们知道可以从**召回率**、**精准率**和**通过率**三个模型评估指标来介绍我们的模型。

召回率在该项目中的意义是：模型预测的真实坏人数（即会产生逾期的用户）与总的真实坏人数的占比。也就是说，在一群真实的坏人里面，模型能抓出多少真实坏人。召回率为1时，代表模型预测的坏人覆盖了所有真实的坏人，没有放过一个坏人。

精准率在该项目中的意义是：模型预测的真实坏人数与模型预测的总坏人个数的占比。也就是说，模型预测的总坏人数里有多少是真的坏人。精准率为1时，代表模型预测的坏人都是坏人，没有“错杀”一个好人。

通过率在该项目中的意义是：模型预测成好人的个数与总人数的占比。即贷款申请成功的人数占总人数的比。一样的，该值越高越好。

在现实中，召回率和精准率总是相违背的。这也很好理解。因为实际真实坏人数是固定的，如果我们希望提高召回率，那么只要提高预测成坏人的个数即可，然而提高预测成坏人的个数，意味着模型会将更多的好人预测成坏人，这样精准率就降低了。所以召回率和精准率在实际应用场景中是鱼和熊掌，很难兼得。项目中模型最终的表现结果如图9所示：

				预测坏人数		预测的真实坏人数		召回率		精准率		通过率	
	总样本量	真实坏人数	真实坏人占比	AI	规则	AI	规则	AI	规则	AI	规则	AI	规则
第一批	1444	300	20.77%	112	126	51	32	17%	10.67%	45%	25%	92.24%	91.27%
第二批	2242	1025	45.72%	224	203	149	100	14.54%	9.75%	66.50%	49.20%	90%	90.94%
第三批	4399	2026	46.05%	467	405	320	214	15.79%	10.56%	68%	52.80%	89.30%	90.80%

图9.模型表现

很明显，采用机器学习建模来判断客户是否给与准入的表现要远好于传统的专家规则。

4
总结

通过本文的介绍，不管是人工智能的研究者还是普通的业务人员都可以很好的评估一个模型的表现效果。知道如何评价模型，那么如何建立好的模型呢？这就需要我们理解业务知识，掌握各种算法，不断学习，不断专研。