评估机器公平性的规范原则学习

Derek Leben
leben@pitt.edu
匹兹堡大学约翰斯敦分校

抽象的

有许多不兼容的方法可以衡量机器学习算法的公平结果。本文的目标是将受保护群体（种族、性别、性取向)的成功率和错误率描述为一个分配问题,并根据道德和政治哲学的不同规范原则描述该问题的可能解决方案。这些规范性原则基于分配问题中的各种相互竞争的属性:意图、补偿、应得、同意和后果。每个原则都将应用于样本风险评估分类器,以展示不同公平性指标集背后的哲学论点。

CCS 概念

·社会和专业话题→计算/技术政策；·计算方法论→ 人工智能的哲学/理论基础；机器学习。

关键词

公平,机器学习,政治哲学,歧视,算法决策

ACM 参考格式： Derek Leben。 2020. 评估机器学习公平性的规范原则。在 2020 年 AAAI/ACM 人工智能、伦理和社会会议 (AIES 20) 会议记录中,2020 年 2 月 7 日至 8 日,美国纽约州纽约市。ACM,美国纽约州纽约市,7 页。 https://doi.org/10.1145/3375627.3375808

1 简介

机器学习算法越来越多地用于公共和私营部门,以做出有关招聘、财务申请、大学录取、医疗诊断和监禁的决策。我们最关心的算法是产生二进制输出的分类器,例如贷款申请的“批准/拒绝”,囚犯的“危险/安全”,求职者的“雇用/通过”,医疗诊断的“恶性/良性”,等等。

这些算法的吸引力很明显;它们可以极大地提高决策的效率、准确性和一致性。因为他们是在重要资源和机会的分配中发挥作用,我们有义务确保算法也没有对历史上代表性不足的群体的歧视性偏见。

然而,有很多方法可以衡量二元分类器是否确实没有对这些受保护群体的偏见,正如 Kleinberg 等人。 [11]和 Chouldechova [3]已经证明,根据每个公平性指标在数学上不可能实现对等。在这种情况下,Binns [2]指出,可能需要来自道德和政治哲学的规范原则来证明做出哪些设计选择是合理的。本文试图朝这个方向迈出更详细的一步。如果我们将成功率和错误率描述为一个分配问题,那么道德和政治哲学中有非常具体的规定可以用来证明使用一组公平指标而不是另一组指标是合理的。

2 机器学习中的奇偶校验指标

如果我们有一个二元分类器( ),它是在某些数据集(,) 上训练的,其中 x 是输入值的向量,y 是分类（0 或 1） ,那么( )将为我们提供预测值类别,对于一些新的数据集，.例如,在标记为“1=癌症”或“0=无癌症”的皮肤标记图像上训练的算法将判断新皮肤标记是否癌变。这里，-values 是像素的模式，-values是图像是否被标记为癌性的，( )是一个分数,它将产生^ = 1 表示癌症和^ = 0表示没有癌症的分类。然后可以通过预测类别与新数据的实际类别之间的匹配来评估模型的输出。因为预测值和实际值有两个可能值,所以比较分为真阳性 (TP)、真阴性 (TN)、假阳性 (FP) 和假阴性 (FN) 类别。

理想模型的所有输出都在 TP 和 TN 类别中,但这是不现实的。所有分类器都会有一定数量的误差,问题是如何相对于这个误差来评估模型（TP 和 TN 结果）的成功。考虑一个结果在包含它的另一组结果中的比率将产生条件概率。例如,考虑所有阳性预测中真阳性率将给出阳性预测率,这也是概率:(= 1| ^ = 1)。这个指标告诉我们有多少积极的预测实际上具有目标特征。另一方面,我们也可以询问有多少负面预测实际上缺乏特征,即负面预测率:(= 0| ^ = 0)。

确定分类器在处理它时是否不公平

A 组与 B 组比较,我们可以比较两组的成功率或错误率。如果是X 组的成功率或错误率,那么最重要的奇偶校验指标如下：

AIES 20,2020 年 2 月 7 日至 8 日,美国纽约州纽约市

ACM 国际标准书号 978-1-4503-7110-0/20/02。 . . 15.00
美元https://doi.org/10.1145/3375627.3375808

人口均等：

$$\frac{(\hat{} \quad = 1)}{+} = \frac{(\hat{} \quad = 1)}{+}$$

阳性预测平价：

$$\frac{(= 1 | \hat{} \quad = 1)}{+} = \frac{(= 1 | \hat{} \quad = 1)}{+}$$

负预测奇偶校验：

$$\frac{(= 0 | \hat{} \quad = 0)}{+} = \frac{(= 0 | \hat{} \quad = 0)}{+}$$

误报奇偶校验：

$$\frac{(\hat{} \quad = 1 | = 0)}{+} = \frac{(\hat{} \quad = 1 | = 0)}{+}$$

机会均等：

$$\frac{(\hat{} \quad = 1 | = 1)}{+} = \frac{(\hat{} \quad = 1 | = 1)}{+}$$

让我们将这些奇偶校验率应用于示例数据集。想象一下,我们训练了一个风险评估算法来辅助假释判决,我们想评估它在对待白人和黑人囚犯时是否不公平。假设共有 600 名白人囚犯（A 组)和 200 名黑人囚犯（B 组）。每个组的结果如图 1 所示。

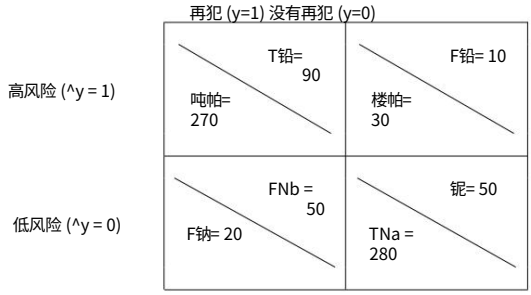

图 1:样本风险评估分类器

这个算法对白人或黑人囚犯不公平吗?答案取决于我们用来评估公平性的平等指标。
就阳性预测值而言,该模型对两个人口统计组均一视同仁:白人囚犯为 90% (270/300),黑人囚犯为 90% (90/100)。这意味着它正确地预测了两组的再犯罪率相同,这就是Northpointe 为其 COMPAS 算法辩护以免受偏见指控的方式。更一般地说,这个样本分类器也满足人口统计平等,因为 50% 的黑人和白人囚犯都是被标记为“高风险”。Feldman 等人[6]认为该指标的修改版本满足不同影响法中非歧视的法律标准。通过这些指标,人们可以争辩说该算法在道德和法律上都是公平的.

通过其他指标,人们可以声称该算法是不公平的。
其中一项指标是被错误识别的和平囚犯的不平等（误报率）,黑人囚犯为 16% (10/60),但白人囚犯仅为 9.6% (30/310) 。这是 Pro Publica 在指控 COMPAS 评估存在偏见时使用的指标。算法可以被称为不公平的另一种方式是诉诸于从两组中正确识别的危险囚犯比率的不平等,白人为 93% (270/290),但黑人仅为 64% (90/140)).因为这使得那些真正具有该特征的人能够被正确识别,我们可以称之为机会均等（该标签显然在积极条件有益的环境中效果更好,例如那些实际上有资格胜任所雇工作的人） . Hardt 等人 [7]将相等的误报率和机会均等的组合称为“机会均等”。

鉴于我们不可能在所有指标中实现均等,我们应该更关心哪些指标?为了论证为什么我们应该更喜欢一种利率而不是另一种利率,我们必须求助于分配商品的规范原则。

3 规范分配原则

根据 Binns (2018) 的建议,本文将组间的成功率和错误率描述为分布问题。如果这种描述是恰当的,那么可以通过它们与公平分配的规范原则的接近程度来评估上述均等指标。在道德和政治哲学中,规范原则可以分为两类,结果论和道义论（图 3）。本节将总结这些原则;可以从[16]中的哲学角度和 [13]中的福利经济学角度找到更多细节。

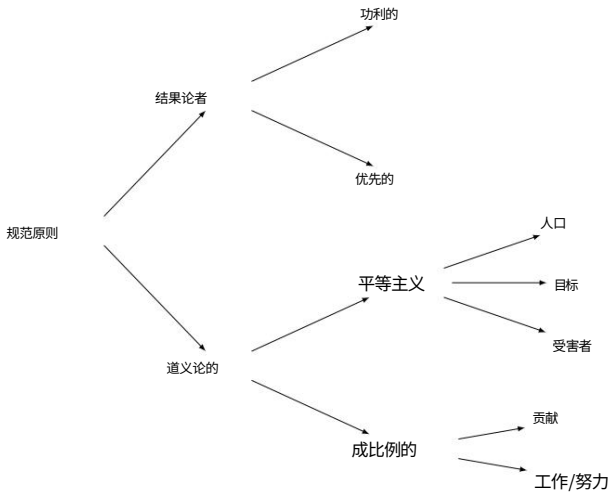

图 2:规范分配原则系列

假设我们是一个中央计划者,拥有一定数量的总商品 ,我们希望将其分配给两个人,[illegible],其中 (1 . . .) 是表示可能分配的向量,使得每个和 (1 . .) 的矢量总和是幸福感的度量。效用函数被解释为衡量爱丽丝和鲍勃幸福结果的指标。该效用函数可以通过一个系数打折,它根据需要或相对价值适当地修改这些患者的效用。配备了适当折扣[illegible]义者 [12] 的总和,要么最大化优先主义者 [1]的最小值:

功利主义原则（Maxisum）：

UT( ) = 最大参数 =1 (,)

优先原则（Maximin）：

PR( ) = arg 最大最小值 (,)

在将 10 美元的整数部分分配给 Alice 和 Bob 时,Prioritarian 显然会选择 (5,5) 的分配。如果结果没有权重,那么功利主义者将对所有分配无动于衷,因为所有潜在分配的总和为 10 美元。然而,如果我们对结果进行加权以说明边际效用递减,或者当另一个参与者获得更大份额时包括对两个参与者的嫉妒权重,那么功利主义者也会选择 (5,5)。

相比之下,道义论方法将评估基于权利的分配。平等主义方法赋予人口中的每个成员平等的权利,从而平等的份额。平等权利也可以只分配给符合条件的一部分人口,也许作为对预期受益人的奖励[10],或者作为对那些为该过程做出牺牲的人的补偿[14]。

其他道义论方法并不赋予细分人群平等的权利,而是根据每个人对生产的贡献,按比例调整每个人的权利（和分配） 。

形式上,道义原则可以用损失函数ℓ(,) 来表征,它衡量分配结果与某些公平标准之间的距离。最简单的公平标准仅基于人口规模,其中 是每个人对总商品的期望值,如果我们将10 美元整数分配给 Alice 和 Bob,则平均分配将为每人 5 美元.更一般地说,当我们无法实现完全平等的分配时,均等主义者会更喜欢与此公平标准的距离最小化的分配: .

平等主义原则：

EG( ) = 最小参数 =1 ℓ(,)

比例道义原则将根据进入生产过程的因素进行分配,例如:进入生产的每个人口成员的资源， ( ),部署这些资源的实际工作量资源， ( ),以及部署这些资源的运气， ( )。假设运气和工作耗尽了一个人资源的可能来源,所以= + （资源=运气+工作）。像 Rand [15]这样的自由主义者关注每个人在同意时的总贡献,其中一个更富有的[illegible]斗志旺盛的年轻新贵更大的份额.相比之下,基于沙漠的方法[17]认为权利与个人努力成正比,运气对分配的影响要最小化。1这两个原则总结为：

自由主义原则：

LB( ) = arg 最小值 =1 ℓ(,)

沙漠原则：

DS( ) = arg 最小值 =1 ℓ(,)

例如,假设爱丽丝为制作贡献了 3 美元,而鲍勃只贡献了 2 美元。自由主义者会确定 Alice 对商品有权利,而 Bob 只有对商品的权利,正确的分配结果是 (6,4)。

另一方面,如果事实证明爱丽丝的工作只产生了她的 1 美元投资,而鲍勃努力工作产生了他所有的投资,那么沙漠原则将确定鲍勃欠它,因此中央计划者必须随机化在(3,7) 或 (4,6)的同样好的整数分配之间。

的货物,而爱丽丝只有权

将规范分布原则定义为优化问题使得将它们插入机器学习算法变得相对简单。对于后果论者来说,公平性原则已内置于数据本身,因为每组输入和分类(1, 1)现在也与效用值1 配对,并且任何最大化准确性的学习算法都将根据公平性进行加权考虑因素。对于道义学家,可以添加损失函数作为对最大化准确性目标的约束。这是继Thornton 等人的重要发现之后。 [19]结果论和道义论原则属于机器学习中的自然类别。它还使我们能够看到所有公平原则如何旨在最大限度地提高公平边界内的准确性。

4 道义论方法

鉴于不可能在所有指标上实现全体人口的平等,平等主义者可能会尝试根据预期接受者或受害群体对人口进行细分。在二元分类器评估的上下文中,基于意图的方法可以假设正= 1 或负特征=^ 0的分类是模型的目标,而特征,因此只有测量 ( ) 是有条件的分类

1幸运均等主义者不是直接按努力比例分配商品,而是通过平均运气的影响来间接优先考虑努力。沙漠和运气均等主义原则对于我们的目的来说看起来是一样的,所以这个微妙之处将被忽略。

是那些设计师负有道德责任的,其形式如下:

(| ^)

这些包括人口统计均等和正/负预测均等。其他指标中的不平等是康德所说的“可预见的”伤害,或者在这种情况下,可预见的不平等,设计师应该为此承担道德责任。在这方面， Northpointe 在强调平等时采取了基于意图的立场COMPAS 在人口统计和积极预测均等方面的影响。该论点假设犯罪分类是风险评估的预期目标,而所有其他结果都是附带损害。这些论点在关于预测性警务的辩论中也很常见,警察声称他们使用统计方法“去犯罪的地方”,任何不公平的错误率都只是可预见的副作用。

另一方面,基于补偿的分配正义方法侧重于哪个群体因分类器的先前状态而变得更糟。正如诺齐克强调的那样,这将分配正义问题转变为矫正正义问题。 Saleiro 等人。 [18]在他们的名为 Aequitas 的“偏见和公平审计工具包”中采用了这一立场,该工具包区分了辅助干预和惩罚性干预。辅助性干预是那些赋予接受者本来不会享受的好处的干预,而惩罚性干预会增加成本或以前不存在的惩罚。

Aequitas 然后提出惩罚性分类器应该通过 FP 率的平等来评估（对那些不值得惩罚的人给予惩罚）,而辅助分类器应该通过FN 率的平等来评估（未能奖励那些应得的人）.

要指定哪些胎次率很重要,Aequitas 需要将其中一个分类值指定为不良结果， 。对于风险评估算法,其中^ = 1 表示危险,则= 1。对于贷款资格算法,其中^ = 1 表示合格,则= 0。我们关心的利率具有以下形式:

[(^ =)| (= (1 −))]

该公式正确地生成了风险评估算法的 FP 率和贷款资格算法的 FN 率。描述补偿观点的另一种方式是利用“非恶意”原则,该原则在伤害而非利益上强加平等主义。正如 Saleiro 等人所说:“..... [提供] 帮助误报的个人不会受到伤害他们,但失踪的人可能对他们有害。”

对补偿方法有两个主要的反对意见。
首先,它允许任意大的不平等以获得有益的结果;当一个风险评估分类器包含被错误释放的黑人和白人危险囚犯之间的巨大差异时,它可以被判断为可以接受。其次,它假设了“富裕”和“贫困”之间的可疑区分,在考虑分类器的所有影响时可能会失效。正如后果论者会指出的那样,释放危险的囚犯可能不会让他们的境况更糟,但肯定会让公众的境况更糟。

与其在选定的群体中实施平等主义,不如
自由主义原则完全根据资源进行分配
每个人都为生产做出了贡献。将其转化为二元分类器上下文的一种方法是将某些人群中某种特征的普遍性视为该群体在分类中的“投资”。如果群体中特征的先验分布为 (= 1),则Group 的贡献是:(= 1|)

在我们的样本数据中,白人囚犯再犯罪的预期为 48.3% (290/600),而黑人囚犯再犯罪的预期为 70% (140/200)。这是从规范的自由意志主义原则思考价值的一种方式。如果白人囚犯的比例为 48%,黑人囚犯的比例为 70%,那么每组都有权获得至少与其初始贡献一样公平的成功率。例如,如果黑人和白人囚犯在 FP 或 FN 率上的差距超过目标特征的原始差距的大小,那么自由主义者可能会称该分类器不公平。然而,在这个范围内的不平等并不是不公平的,因为它与人口特征的原始不平等成正比。尽管最初的不平等可能是运气不好或历史压迫的结果,但自由主义者并不认为我们的分类器有责任减轻这些因素。

在经典的自由主义中,原始贡献比例的任何差异都被认为是不公平的。如果爱丽丝为生产贡献了 3 美元,而鲍勃只贡献了 2 美元,那么 50-50 的货物分配与 70-30 的分配一样错误,因为理想的分配是 60-40。然而,在分类器的情况下,理想情况是两组在预测措施上都取得 100% 的成功,而在错误措施上两组都取得 0% 的成功。因此,从公平的角度来看,我们只关心确保群体之间的不平等不超过目标特征中预先存在的不平等。形式上,这可以通过将超过此先验比率的比率与不超过此先验比率的比率分开来实现。设∈为超过原始组间贡献率的比率子集:

$$\epsilon : \quad \frac{*}{*} > \frac{}{}$$

我们力求最小化的自由意志主义“成本”变成了这些不公平比率与原始贡献率之间的差异。一旦我们将这种差异最小化,就可以优化模型的准确性,并且群体之间任何剩余的不平等都是“不幸的,但不公平。”自由主义者的目标不是创造更多的平等,而只是不扩大我们社会先前存在的不平等。

基于沙漠的方法拒绝这种对人群中某种特征先前流行的关注,因为这可能是不公正环境（历史压迫或运气)的结果。在二元分类器中,Desert 的拥护者经常采用人口统计均等的修改版本,称为条件人口统计均等,其中受保护群体在预测类别中的代表必须相等,条件是某些被认为合法的因素[9]。如果( )是数据中作为工作结果的特征集,那么沙漠理论家关心的分类器是:

(^ = (1, 0)| ( ))

对于风险评估算法， ( )中的特征可能是基于合法因素（如先前的定罪)的特征,而不是基于

父母的教育水平,即使后者被证明是目标变量的极其准确的预测因子。

沙漠理论家面临的挑战是准确说明( )中的合法限制应该是什么。最广泛的限制可能只是那些实际上有资格进行分类的人,其中( ) = (= 1),这就是 Hardt 等人。 [7]呼吁机会均等:

(^ = 1| = 1)

哈特在一篇博客文章中激发了这个想法,他在文章中展示了目标营销的人口平等问题:

> 例如,考虑一家豪华连锁酒店,它向一部分富有的白人(他们可能会光顾酒店)和一部分不太富裕的黑人(他们不太可能光顾酒店)提供促销活动。这种情况显然非常棘手,但只要每个群体中有相同比例的人看到促销活动,人口平等就完全没有问题。

Hardt 强调,重要的是向可能同样数量光顾酒店的白人和黑人消费者提供促销活动,而不是仅仅确保向同等数量的白人和黑人消费者提供促销活动。对于我们的示例风险评估算法,在考虑实际危险的囚犯时,该模型预测白人囚犯的高风险率为 93%,黑人囚犯为 64%,因此未能满足机会均等。

沙漠理论家还有其他可能的方法来限制条件平价的数据。例如,我们可以查看那些具有一定数量先验的囚犯,并要求具有 3 个先验的白人囚犯的预测应与具有 3 个先验的黑人囚犯的预测相同。 Dwork 等人的 [5] “通过意识实现公平”模型将衡量个人距离而不是群体结果。Dwork 明确引用罗默等幸运平等主义者的工作作为哲学灵感。在她的模型下,每个人都被分配了一定的距离一个评估沙漠的度量空间,而评估模型公平性的方法是通过该度量空间内每个组的个体之间的平均距离。例如,一名白人囚犯可能由于纯粹的运气而出生在更特权的背景下,因此,即使白人和黑人囚犯的前科数量相同(例如,三个前科),通过平均或抵消他们背景特权的影响,我们也可以将白人囚犯的“努力”评为低于黑人囚犯,因此,预测不同程度的风险。有理由认为,关注个人努力也可以更好地预测未来的行为,但这更像是一个结果论者的论点。沙漠理论家纯粹关心对过去行为的奖励和惩罚。

对沙漠原则的主要反对意见是无法正确区分标记为“运气”和“工作”的因素。例如,那些努力工作的人之所以具有这种性格特征,是因为他们历史上的一系列因素(养育、榜样、教育、遗传等),这些因素本身并不是通过工作获得的。

5 结果论者的方法结果论者将以某种程度的怀疑态度听取关于各种公平指标的辩论,因为社会正义不取决于群体之间的平等或不平等,而是取决于不平等对这些群体中个人幸福的影响。因此,像人口均等这样的度量标准本身对功利主义者来说是无趣的,因为在 ^ = 0 的类别中,仅仅是代表性的不平等并不能告诉我们关于= 1 或分类与总效用之间的关系的任何信息。相反,如果我们可以根据总体社会成本对每个结果的相对权重进行估计,那么我们就可以简单地设计模型来优化社会成本而不是平等。

如果 是每个类别的平均比率:和( ) ( )是每个类别的平均 , , , , 加权效用,那么功利主义者将简单地选择最大化效用总和的

( ) ( ) ( )

Prioritarians 会最大化最小值:

( ) ( ) ( )

每个类别的实用程序将根据分类器的类型而有所不同。对于贷款资格算法, FP 产生的负效用比 FN 产生的负效用差得多,因为前者代表投资的完全损失。

然而,对于刑事司法中的风险评估算法,FN 的负效用可能比FP 的负效用差得多,这仅仅是基于预防的暴力数量。

将无辜者关押所造成的痛苦可能小于让极其危险的囚犯获释所造成的痛苦。道义学家会拒绝这个计算我们愿意为清洁街道接受多少无辜入狱者的整个项目,但如果囚犯和公众的总成本之间没有区别,那么使用“享乐”的共同货币价值观”是必要的要求。

让我们首先为样本分类器中的每个结果类别分配效用 ( ),包括囚犯和公众的效用。 “效用的人际测量”存在着臭名昭著的挑战,但目前这些价值只是规定。假设释放危险的囚犯将以暴力犯罪的形式对公众造成平均 400 单位的伤害,而拘留任何人囚犯将对他们和他们的家人造成 100 单位的伤害,和平囚犯被不必要地监禁会造成额外 100 单位的痛苦。涉及和平囚犯(FP 和 FN)的结果对公用事业没有影响。这些公用设施的地图表示在表 1 中:

表 1:风险评估结果的示例实用程序

结果 囚徒效用 公共效用 总效用			
TP	-100	400	300
[illegible]	-200	0	-200
前线	100	-400	-300
[illegible]	100	0	100

结果论方法中的公平性体现在我们如何根据相对社会成本为每个组的结果分配权重 ( )。例如,[illegible] 延续黑人社区的历史剥夺循环。同样,释放和平的黑人囚犯可能会带来更大的社会效益。在我们的样本数据中,黑人再犯占他们所在群体的比例是白人囚犯占他们所在群体的比例的 1.4 倍(70%:48%)。在那种情况下,我们可能会权衡黑人囚犯持续误报的负面社会成本要差 1.4 倍,而黑人囚犯真阴性的积极社会效益要好 1.4 倍。这种加权特征是在结果主义方法中实现更大平等的主要工具。使用这些值,图 4 表示我们的示例分类器的实用程序。

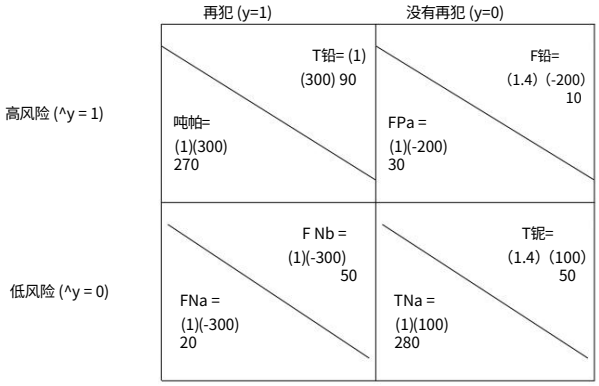

图 3:具有加权效用的列联表。

结果主义的好处是将公平直接纳入数据,并针对整体利益进行优化。由于黑人囚犯中 TN 的额外收益和 FP 的额外成本,效用优化也将增加群体之间的均等性。与 Corbett-Davies 等人的担忧一致。 [4],这种平价将与牺牲过多公共安全的社会成本相平衡。

对结果主义有很多反对意见。我们已经考虑了效用的人际比较问题。另一个反对意见认为,即使我们可以为每个类别分配特定的实用程序,对整个数据中的所有受保护组系统地这样做仍然是不现实的。结果论者有时会通过提倡一种规则功利主义来回应这些反对意见,这种规则功利主义采用的启发式方法已被证明可以有效地在小规模环境中最大化幸福。结果论者可能会提出与公平指标类似的东西。 Hu 和 Chen [8]最近证明了将每个 ML 公平性指标转化为相应效用计算的正式可能性。如果可以模拟不同模型对总效用的影响,并表明公平性指标在排名最高的模型类别中被实例化,那么这将是一个强有力的结果论者论证使用作为默认指标。

6 结论

可以理解,采用 ML 算法的公司希望避免出现指责他们歧视的负面新闻。问题在于,任何适用于在受保护群体中分布不均的特征的ML 算法在某种意义上都可能被指责为歧视。假设该公司最初设计其模型以产生成功率的平等,那么明天的标题可能是:

算法犯错误的次数超过

如果公司调整其模型以产生相同的错误率,那么标题可以是:

算法批准更多值得的成员
比应得的成员

当公司试图创造机会均等时,下一个标题可能是:

算法提供了比

面对这些艰难的选择,公司可能会故意对模型的细节含糊其辞,或者遵从未来的行业标准,或者干脆完全放弃对公平的承诺。然而,我会鼓励开发 ML 算法的团队不要气馁。相反,工程师和计算机科学家应该意识到设计模型涉及做出艰难的道德承诺。试图忽略这些选择,或者只是“取平均”只会产生不负责任的结果。

对“这个模型对群体公平吗?”的回答永远是:“根据哪个规范原则是公平的?”强调模型的预期输出将更关心成功率的平等,而强调那些因模型而处于不利地位的人将更关心错误率的平等。对相称性的担忧会导致人们更关心匹配群体内某种特征的预先存在的流行程度,或者个人可以对这种流行程度负责的程度。

衡量对受模型影响的每个人的总体影响（而不仅仅是少数人的权利)将导致将公平指标纳入社会成本和收益的一般计算中。如果我们选择一种方法,那么其他人就会受到影响。但这是道德选择的本质,减轻负面新闻的唯一负责任的方法是对它们做出一致的反应,而不是忽视它们。

致谢

作者非常感谢 Alexandra Chouldechova 在写作的各个阶段提出的深思熟虑的评论和建议。

参考

[1]马修·阿德勒。 2011. 福祉和公平分配:超越成本效益分析。
牛津大学出版社,纽约,纽约。
[2]鲁本·宾斯。 2018. 机器学习中的公平性:政治哲学的教训。机器学习研究论文集 8 (2018),
1-11。
[3]亚历山德拉·乔德乔娃。 2017. 具有不同影响的公平预测:累犯预测工具中的偏差研究。大数据 5 (2017),153–163。

[4] Sam Corbett-Davies、Emma Pierson、Avi Feller、Sharad Goel 和 Aziz Huq。 2017.
算法决策和公平成本。在第 23 届ACM SIGKDD 知识发现和数据挖掘国际会议论
文集中。
ACM,哈利法克斯,NS,加拿大,797–806。
[5] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。 2012. 通过意识实现
公平。在过程中。第三届 ITCS。 214-26。
[6] Michael Feldman、Sorelle A Friedler、John Moeller、Carlos Scheidegger 和 Suresh
Venkatasubramanian。 2017. 证明和消除不同的影响。在过程中。第 21 次 SIGKDD。美国
计算机协会。
[7]莫里茨·哈特、埃里克·普赖斯和纳蒂·斯雷布罗。 2016.监督学习机会均等。在过程中。第 29 次 NIPS。 3315-23。
[8]胡莉莉、陈以玲． 2018.公平的福利和分配影响
分类。 (2018)。. arXiv 预印本 arXiv:1807.01134。
[9] Faisal Kamiran、Indre liobait・e和 Toon Calders。 2013. 量化可解释的
歧视并消除自动决策中的非法歧视。
知识和信息系统 35 (2013), 613–644。
[10] 伊曼纽尔·康德。 1788. 实践理性批判。牛津大学出版社。
[11] Jon Kleinberg、Sendhil Mullainathan 和 Manish Raghavan。 2017.风险评分公平确定中的固有权衡。在过
程中。第八届ITCS。 ITCS。
[12] 约翰·斯图尔特·穆勒。 1861. 功利主义。哈克特,纽约,纽约。
[13]赫维磨坊。 2003.公平分工与集体福利。麻省理工学院出版社,剑桥,
嘛。
[14] 罗伯特·诺齐克。 1974. 无政府状态、国家和乌托邦。 Basic Books,纽约,纽约。
[15] 安·兰德。 1961. 自私的美德。企鹅出版社,纽约,纽约。
[16]约翰·罗默。 1971. 分配正义理论。哈佛大学出版社,马萨诸塞州剑桥市。
[17]约翰·罗默。 1998. 机会均等。哈佛大学出版社,剑桥,
嘛。
[18] Pedro Saleiro、Benedict Kuester、Abby Stevens、Ari Anisfeld、Loren Hinkson、Jesse London 和 Rayid
Ghani。 2018. Aequitas:偏见和公平审计工具包。 (2018)。 arXiv 预印本 arXiv:1811.05577。
[19] Sarah Thornton、Selina Pan、Stephen Erlien 和 Christian Gerdes。 2016. 将道德考量纳入自动车辆控制。
在 IEEE智能交通系统交易中。 1-11。