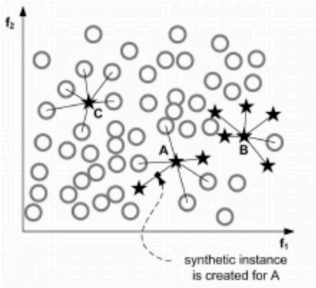
# 为什么不要用 GAN 生成不平衡的样本？

样本不平衡问题是分类任务中一个常见的现象，由于正样本的数量远远大于负样本，如果仅仅训练模型来获得高的 accuracy，那么很可能得到一个难以检测出负样本的模型。因此评价模型的时候我们会结合其他指标比如某类的精度（当样本被分为该类时有多大的可信程度），召回率（该类的样本有多少被正确检测出来），AUC（ROC 曲线下的面积）等等。

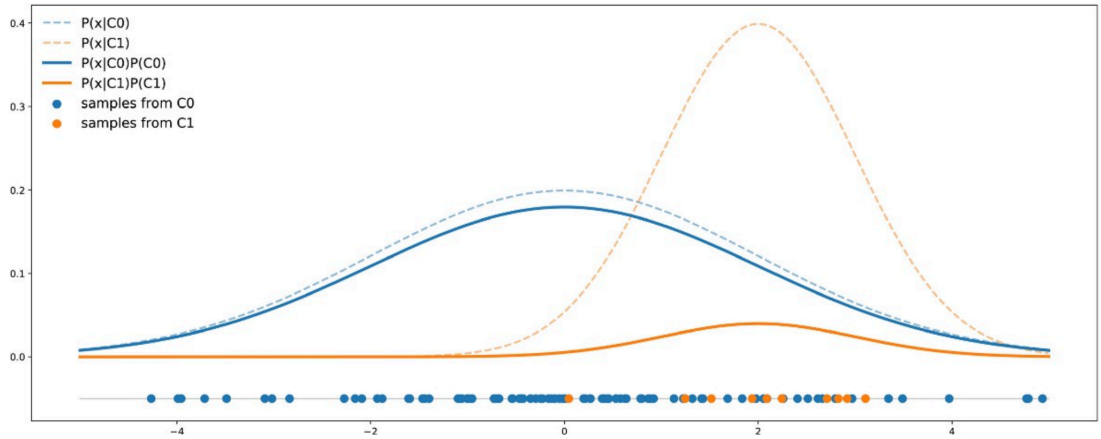
解决数据不平衡问题的常见方法有重新采样+集成学习、调整不同样本的权重、调整概率阈值（比如大于 0.8 才认为是正类），以及 SMOTE 等生成更多样本的方法。SMOTE 主要思想就是随机选负类样本的一个近邻，然后在与近邻点的连线上取新的点，从而得到人工合成的新样本。




既然可以合成样本，我们是不是可以用基于 GAN 或者 RL 的一些当下比较流行的生成方法来进行合成呢。直观上想好像是可行的，但是经过简单的调研我们可以发现，对于非图片类的分类问题，这样做意义不大---并不会比 SMOTE 这种方法优化多少。本文分为两个部分，第一部分描述为什么样本不平衡的数据难以学习，第二部分我们看两篇用 GAN 套在这个场景上的水文有什么问题。

## 1. What's wrong with the imbalance dataset?

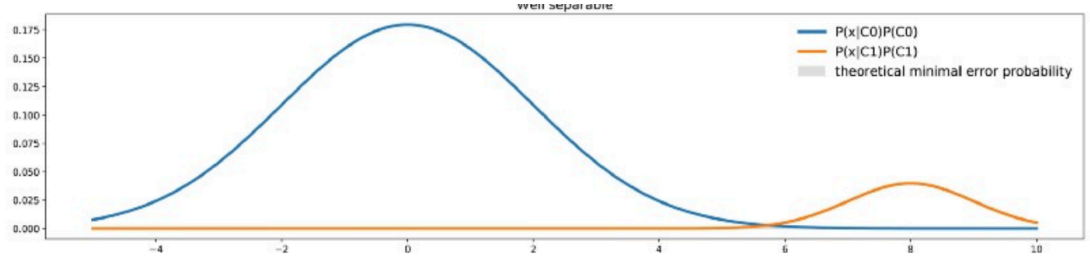
在 blog[1]中 Baptiste Rocca 指出，数据不平衡实际上导致了数据更加不可分离。既然是要讨论生成模型，我们首先假设两个类的的样本分别采样自分布 P(x|C0),P(x|C1)，我们要学习的是 x 产生自某一个类的概率 P(C0|x)，假设产生样本的两个分布都是高斯分布，且是可分的---两个分布不完全重合，而类别不均衡可以看做先验概率 P(C0)和 P(C1)相差很多。



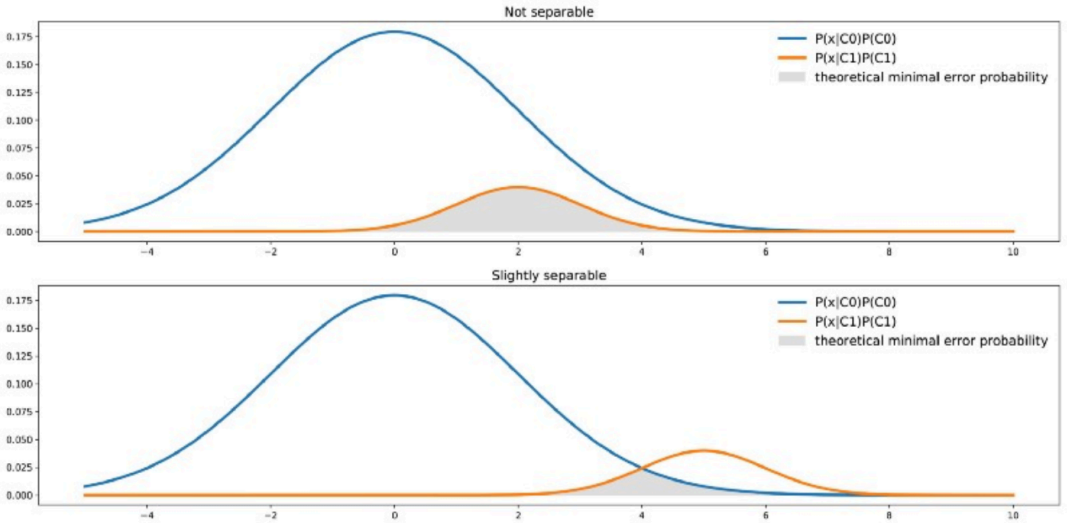

可以看到，原本可分的两个虚线由于先验概率相差太多变得不能分离，在任何点判断是 C0 的概率要大于 C1 的概率：

$$\mathbb{P}(C0|x) = \frac{\mathbb{P}(x|C0)\mathbb{P}(C0)}{\mathbb{P}(x)} > \frac{\mathbb{P}(x|C1)\mathbb{P}(C1)}{\mathbb{P}(x)} = \mathbb{P}(C1|x)$$

当然不平衡不一定会导致不可分，如果本身产生正负样本的分布就是不重叠的，那么先验概率将不会有很大的影响：




现实中这种情况很少，抛开这种情况，样本 x 属于某个类的概率分布为图中位于上方的曲线（蓝线接橙线）连起来再归一化（即与坐标轴围成的面积的1）。而最小误差概率为下图阴影部分的面积：




我们的所有方法都是为了让这个面积变小，基于重新采样或者合成数据的方法（相当于同分布的采更多的负例）可以看做是增加负例的先验概率，而这本身就不一定有效，比如在上图图一中，P(C1)变大在早期一定导致阴影部分的面积（占整个曲线下方面积比例）增大。

2. 一些水文

文章[2]中提出，既然我们通过给样本加权来处理不均衡问题，那么是否可以对负样本加不同的权重来使分类的性能更好呢？具体来说，用生成器 G 来为不同的负样本生成权重，同时 D 来验证在不同权重的情况下分类的效果。文章给了这样的 loss：

$$\min_G \max_D \frac{1}{|S^+|}\Sigma_{x\in S^+}\log D(x) + \\ \Sigma_{x\in S^-}G(x)\times\log(1-D(x)) + \\ \lambda\times\Sigma_{x\in S^-}G(x)\times\log(G(x)),$$

其中第三部分是一个增加给负样本权值的熵的 loss。整个过程如下图：

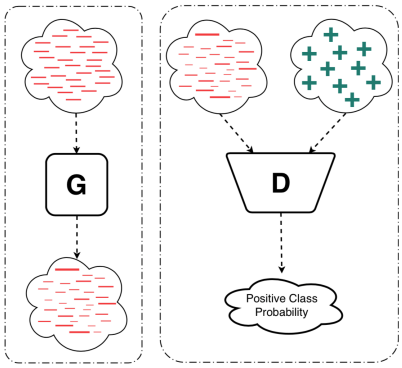



Figure 1: Illustration of the proposed adversarial classification framework. *G* re-weights the negative samples and *D* is trained using the positive and the re-weighted negative samples.

我们讨论一下他的做法，首先这个 loss 的大 bug 是 G 的目标是使得数据更加不可分，也就是说 D 的上限就已经是原本的分类器的效果，不可能更好，可能是因为跑不出效果作者又加了个熵增的项，如果这一项权重非常大，就相当于所有负样本概率相等，所以硬套 GAN 是没什么意义的。

抛开这个模型，给负样本不同权重是否能提高分类效果呢（类似于把他的 G 的 loss 改为 max），答案是在训练集上是可以的，在测试集上不会有提升。更改样本权重相当于改分布 P(x|C1)使得该分布与 P(x|C0)差异更大，然而真实分布时不会变的，所以也是没有意义的。

文章[3]提出了一个更加讲的通生成模型，基于 cGAN 来学习不同类别样本的分布。具体来说，将样本的 label 作为额外的 condition 输入到 G 和 D 中，然后生成器来生成对应类的样本，判别器来区分样本来自生成器还真是数据。也就是说，我们要学习的 G 就是将噪音分布 transfer 到 P(x|C0)的 NN，这是值得一试的，但是效果可能并不会比过采样或者 SMOTE 好很多，一方面对于维度低、特征不丰富的情况，D 很容易区分两者，导致 G 只能生成和真实样本非常近似的点，另一方面有时调整先验 P(C0)和 P(C1)的比例，甚至可能得到更差的误差下限。

在目标检测领域（给定图片或者视频来区分目标和背景）也存在样本不均衡问题，也有一些基于 gan 的方法。比如将图片作为生成器输入来生成旋转或者部分遮挡的图片来丰富样本集[4]，此时生成器相当于一个 AE，其中往往会人为设计变形的网络结构（比如在中间的卷积层设计 mask 块），这种比较启发式的方法类似于对一般数据选近邻来合成新的数据，只不过该方法多出来用 D 来判断生成的图片是不是真实，这是为了保证图片只产生了必要的形变而能够正确的解码。而 D 对于 SMOTE 这种基于 KNN（非参模型）进行合成的方法是多此一举的。也可以尝试用 NN 对输入数据加噪音，但是一般分类场景中我们经常可以知道每个维度的属性（比如用户年龄），可以得到均值方差等统计量，直接用高斯泊松等常用分布建模可解释性会更好一点。

所以最后的结论是，如果样本的维度很高、不稀疏（类似图片）且负样本的数量也足够多，基于 cGAN 的生成模型可以试一试。其他情况可能仍然用概率阈值或者对正负样本分错给不同的惩罚可能会更好一点。

[1]. https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28
[2]. Montahaei E, Ghorbani M, Baghshah M S, et al. Adversarial classifier for imbalanced problems[J]. arXiv preprint arXiv:1811.08812, 2018.
[3]. Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Systems with applications, 2018, 91: 464-471.
[4]. Wang X, Shrivastava A, Gupta A. A-fast-rcnn: Hard positive generation via adversary for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2606-2615.