# UNSUPERVISED AND SEMI-SUPERVISED LEARNING WITH CATEGORICAL GENERATIVE ADVERSARIAL NETWORKS

# Abstract

In this paper we present a method for learning a discriminative classifier from unlabeled or partially labeled data. Our approach is based on an objective function that trades-off mutual information between observed examples and their predicted categorical class distribution, against robustness of the classifier to an adversarial generative model. The resulting algorithm can either be interpreted as a natural generalization of the generative adversarial networks (GAN) framework or as an extension of the regularized information maximization (RIM) framework to robust classification against an optimal adversary. We empirically evaluate our method– which we dub categorical generative adversarial networks (or CatGAN) – on synthetic data as well as on challenging image classification tasks, demonstrating the robustness of the learned classifiers. We further qualitatively assess the fidelity of samples generated by the adversarial generator that is learned alongside the discriminative classifier, and identify links between the CatGAN objective and discriminative clustering algorithms (such as RIM).

本文中我们提出了从未标记或部分标记的数据训练判别分类器的方法。我们的方法基于目标函数，该目标函数权衡观察样本和预测的categorical class distribution之间的互信息量,可用来对抗生成模型分类器的robustness.该算法可以被解释为GAN的通用框架扩RIM框架的扩展。我们凭经验评估我们的方法，该方法CatGAN针对合成数据以及图像分类任务，展示了学习分类器的鲁棒性。我们进一步定性地评估由对抗生成器生成的样本的保真度，对抗生成器由判别分类器学习而得，并将CatGAN目标和判别聚类算法之间进行连接。

# 1. Instroduction

Learning non-linear classifiers from unlabeled or only partially labeled data is a long standing problem in machine learning. The premise behind learning from unlabeled data is that the structure present in the training examples contains information that can be used to infer the unknown labels. That is, in unsupervised learning we assume that the input distribution p(x) contains information about p(y |x)…. By utilizing both labeled and unlabeled examples from the data distribution one hopes to learn a representation that captures this shared structure. Such a representation might, subsequently, help classifiers trained using only a few labeled examples to generalize to parts of the data distribution that it would otherwise have no information about. Additionally, unsupervised categorization of data is an often sought-after tool for discovering groups in datasets with unknown class structure.

在机器学习中，从未标记或仅部分标记的数据学习非线性分类器是一个长期存在的问题。

从未标记数据中学习的前提是，训练样本中存在的结构可用于推断未知标签的信息。

也就是说…p(x)包含了p(y|x)的信息

通过利用标记和未标记的样本，人们希望通过学习来获得这种共享结构的表示。

这样的表示可使用少部分已标记的样本来生成数据的部分分布，用于生成训练器。

此外，无监督的数据分类是用于发现未知标签的数据集的组的常用工具

This task has traditionally been formalized as a cluster assignment problem, for which a large number of well studied algorithms can be employed. These can be separated into two types: (1) generative clustering methods such as Gaussian mixture models, k-means, and density estimation algorithms, which directly try to model the data distribution p(x) (or its geometric properties); (2) discriminative clustering methods such as maximum margin clustering (MMC) (Xu et al., 2005) or regularized information maximization (RIM) (Krause et al., 2010), which aim to directly group the unlabeled data into well separated categories through some classification mechanism without explicitly modeling p(x). While the latter methods more directly correspond to our goal of learning class separations (rather than class exemplars or centroids), they can easily overfit to spurious correlations in the data; especially when combined with powerful non-linear classifiers such as neural networks.

传统上，该任务为聚类问题，针对该问题可以使用大量经过充分研究的算法。这些算法可以分为两种类型：

1.生成聚类方法，如高斯混合模型，k-means和密度估计算法，他们直接尝试对数据分布p(x)或其余几何性质进行建模

2.判别聚类方法，如MMC(2005)，RIM(2010),其目的是通过一些分类机制将未标记的数据直接分组到已区分好的类别，而无需明确地对p(x)进行建模。

虽然后一种方法更加直接地对应于学习类别分离的目标(而不是类别样本或中心)，但是它们更容易过拟合数据中的虚假相关性，特别是与非线性分类器(如神经网络)结合使用时。

More recently, the neural networks community has explored a large variety of methods for unsupervised and semi-supervised learning tasks. These methods typically involve either training a generative model – parameterized, for example, by deep Boltzmann machines (e.g. Salakhutdinov & Hinton (2009), Goodfellow et al. (2013)) or by feed-forward neural networks (e.g. (2014), Kingma et al. (2014)) –, or training autoencoder networks (e.g. Hinton & Salakhutdinov (2006), Vincent et al. (2008)). Because they model the data distribution explicitly through reconstruction of input examples, all of these models are related to generative clustering methods, and are typically only used for pre-training a classification network. One problem with such reconstruction based learning methods is that, by construction, they try to learn representations which preserve all information present in the input examples. This goal of perfect reconstruction is often directly opposed to the goal of learning a classifier which is to model p(yjx) and hence to only preserve information necessary to predict the class label (and become invariant to unimportant details)

最近，神经网络社区已经探索了用于无监督和半监督学习任务的各种方法。这些方法涉及到训练参数化的生成模型，如通过deep Boltzmann machines或通过前馈神经网络或通过自编码器网络。因为这些方法模拟数据分布通过重建输入样本，所有的模型都与生成聚类方法相关，并且只用于预训练分类网络。

这种基于重建学习的方法的一个问题是，通过构造，他们试图学习保留输入样本中存在的所有信息的表示。这种完美重建的目标通常与学习分类器的目标(模拟p(y|x))相反,因此只保留预测标签的必需信息(并对不重要的细节具有鲁棒性)。

The idea of the categorical generative adversarial networks (CatGAN) framework that we develop

in this paper then is to combine both the generative and the discriminative perspective. In particular, we learn discriminative neural network classifiers D that maximize mutual information between the inputs x and the labels y (as predicted through the conditional distribution p(yjx;D)) for a number of K unknown categories. To aid these classifiers in their task of discovering categories that generalize well to unseen data, we enforce robustness of the classifier to examples produced by an adversarial generative model, which tries to trick the classifier into accepting bogus input examples.

CatGAN的框架是结合生成和判别的角度。特别是，我们学习判别神经网络分类器D，其对于K个未知标签最大化输入x和标签y的互信息量(如通过p(y|x,D)预测的)。为了帮助这些分类器更好的发现潜在数据的分类，我们将分类器的鲁棒性施加在对抗生成网络产生的样本上，这些样本试图去欺骗分类器接受虚假输入样本。

The rest of the paper is organized as follows: Before introducing our new objective, we briefly

review the generative adversarial networks framework in Section 2. We then derive the CatGAN

objective as a direct extension of the GAN framework, followed by experiments on synthetic data,

MNIST (LeCun et al., 1989) and CIFAR-10 (Krizhevsky & Hinton, 2009).

本文的其余部分安排如下：在介绍新方法之前，将简要回顾一节中的生成对抗网络框架(GAN)；然后，将CatGAN的目标推导为GAN框架的扩展，然后在MNIST和CIFAR-10数据集上进行实验。

# 2.Generative Adversarial Networks

# 3.Categorical Generative Adversarial Networks

Building on the foundations from Section 2 we will now derive the categorical generative adversarial networks (CatGAN) objective for unsupervised and semi-supervised learning. For the derivation we first restrict ourselves to the unsupervised setting, which can be obtained by generalizing the GAN framework to multiple classes – a limitation that we remove by considering semi-supervised learning in Section 3.3. It should be noted that we could have equivalently derived the CatGAN model starting from the perspective of regularized information maximization (RIM) – as described in the appendix – with an equivalent outcome.

以第2章为基础，我们现在推导用于无监督学习或半监督学习的CatGAN。我们首先将局限于无监督的设置，这个可以通过将GAN推广到多个类来获得，局限性可以通过3.3节中的半监督学习来消除。值得注意的是，从RIM的角度出发，可以等效地推导出CatGAN模型(在附录中有描述)，且可得出相同结果。

3.1 Problem Setting

<https://blog.csdn.net/u014625530/article/details/82393414>

我们考虑从未标记的数据集X无监督地学习判别分类器D的问题，使得D将数据划分至先验选择的类别K。此外，我们要求D(x)产生类别的条件概率分布。学习的目标是训练概率分类器D，其类别分配满足拟合优度的标准。

值得注意的是，因为真实的类别分布是未知的，我们必须求助于中间度量来判断分类器性能，来替代仅仅最小化(如negative log liklihood)。具体来说，我们将总是优先选择D，对于某一特定样本x的p(y|x,D)有更高的确定性，并且对于所有的k，p(y|D)与先验分布P(y)更接近。

在下文中，我们始终假设类别的分布为均匀分布，即我们期待每个类别的样本数是相同的。