交叉熵损失函数

信息论中的基本想法是一个小概率发生的事情要比一个大概率事件发生能提供更多的信息。

假设随机变量x,且服从先验概率分布p(x)，此时我们用I(x)表示事件的信息量。

当两个不相关的事件x和y发生，同时发生的信息量应该等于事件各自发生时获得的信息之和

即I(x,y)=I(x)+I(y)

当两个事件是独立不相关，有p(x,y)=p(x)p(y)

所以我们将I(x)=-log(p(x)),负号保证信息量为正数或者0，因为p(x)的范围为(0,1]

信息熵即所有信息量的期望来量化整个概率分布中的不确定性总量：

H(X)=EX~P(I(x))=-EX~P(log(P(x)))

如果对于X的随机变量有两个单独的概率分布P(x)和Q(x)，可以用KL散度来衡量这两个分布的差异：

DKL(P||Q)=-EX~P(log(P(x))-log(Q(x)))

在机器学习中，我们不会直接使用KL散度，而是使用其替代形式---交叉熵用H(P,Q)

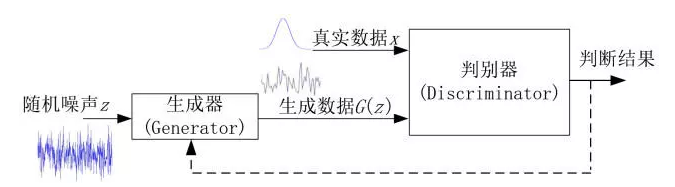
H(P,Q)=H(P)+DKL(P||Q)

H(P,Q)=-EX~P(log(Q(x)))=-sigmaP(x)logQ(x)

因此我们可以将交叉熵理解为概率分布Q对概率分布P估计的准确程度。

GAN网络（2014年NIPS）

GAN模型的基本结构包含两个网络结构-生成模型Generator以及判别模型Discriminator,G网络尽可能生成以假乱真的假样本，而D网络尽可能辨别出真假样本，类似于二分类器，在这个博弈过程中，两种网络性能都越来越好。GAN模型结构如下：

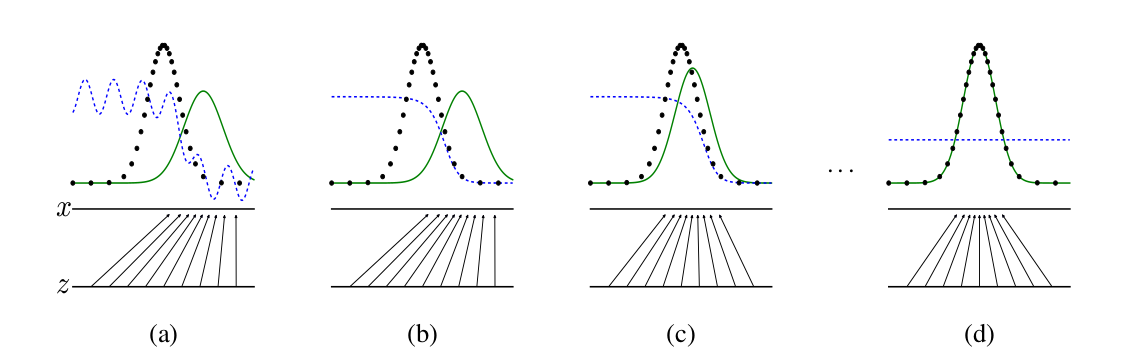


在该论文中，使用了MNIST,TFD,CIFAR-10数据集，以图像为例。

G网络，接收随机变量z的噪声，生成图片，记为G(z)，理想状态为判别器无法区分生成器的

D网络，接收真实数据x,也就是真实图像，输出D(x)为图片x为真实图片的概率，理想状态，如果为1，则表示为真实图片，输出0，表示生成器产生的假图片。

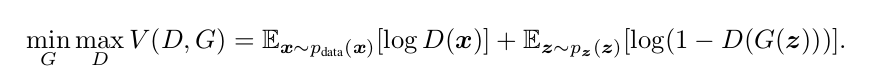
最后博弈结果应该为D(G(z))=0.5



论文中，展示了GAN模型的运行示意图，如上图所示，黑色点构成的离散曲线为真样本的分布，绿色曲线是由噪声z映射生成的假样本的分布，我们需要学习的就是蓝色曲线，表示为

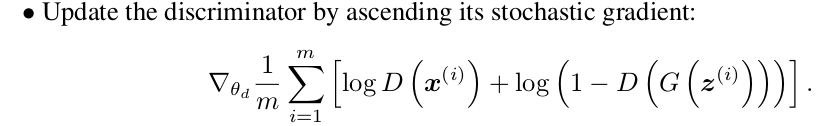
D\*(x)=，当达到理想状态G和D一致时，，D\*(x)=0.5

转化为数学公式为：



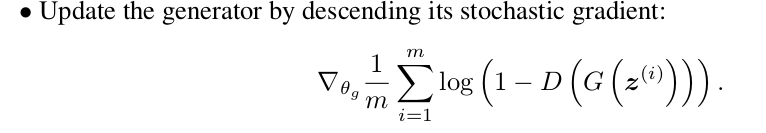
算法分两步：

1. 给定G的条件下,先优化D



公式的第一项，使得真样本x输入的时候，得到的结果越大越好，maxD(x)因为真样本的预测结果越接近1越好，对于假样本，希望预测结果越小越好min(D(G(x)))，为了统一写成max,可以将min(D(G(x)))改写成max(1-D(G(x)))，因此第一步需要对梯度进行上升，获得最大值

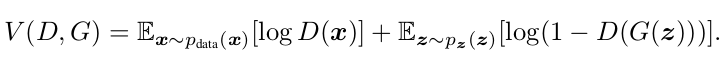
1. 将D网络的参数误差传给G网络，从而实现优化G网络

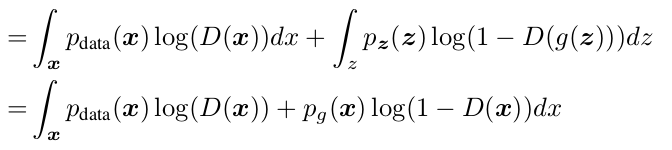


在优化G网络的时候，并不需要真样本，所以论文中也是直接把第一步的第一项直接去掉，对于假样本而言，希望判别器能判别为1，因此预测结果越大越好，由于之前为了统一形式，优化表达式定义为：

故此对D网络优化时应为min G，因此为梯度下降的。

下面就要证明GAN网络如何能得到：





对于给定x，最大化D\*，

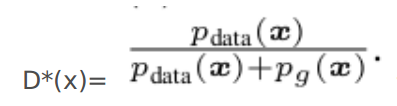


a D b D

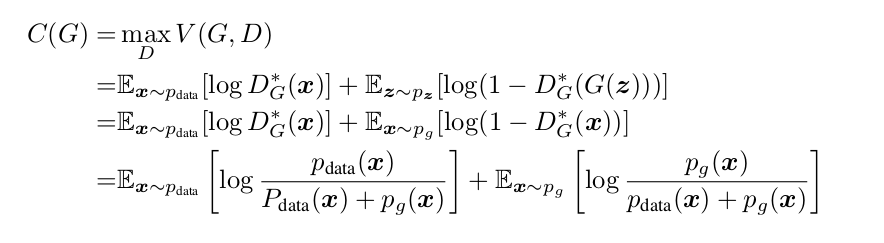
令f(D)=alog(D)+blog(1-D)

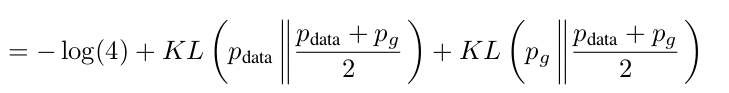
求导：

D\*=a/(a+b)



把最优的D=D\*带入有





表示G和D两个分布之间的差异，最小为-2log2，最大为0

因此时，G是最优的。