OUT-OF-DISTRIBUTION GENERALIZATION WITH MAXIMAL INVARIANT PREDICTOR

OOD问题的定义

> Out-of-Distribution (OOD) generalization is a problem of seeking the predictor function whose performance in the worst environment is optimal.

problem

1. 绝大部分机器学习方法的使用一般基于一个内在的假设:数据集中的每个样本都是independently,并且identically 从同一分布中采样得到的.这使得目前的模型无法应用与实际情况之中

background

针对机器学习模型受i.i.d假设的限制无法使用的问题,我们需要考虑所有可能的分布,并且找到在最差环境中表现最优的模型,即

$$\arg\min_{f}\max_{\epsilon\in\mathcal{A}}\mathbf{Cost}(f|\epsilon) \tag{1}$$

其中ϵ表示环境因子,$\mathcal{A}$表示所有环境的集合.$\mathbf{Cost}$表示某个模型f在环境ϵ下的损失函数值.但(1)中的模型f是无法解出来的,因为我们没办法获得$\mathcal{A}$

relation work

现有的解决OOD问题的方法主要包括IRM类的方法以及 distributional robustness-type methods.

针对目前的OOD问题的方法有两个比较核心的问题:

1. when does the invariant feature help us in solving OOD problem
2. what type of invaraint feature is useful in OOD problem

《Invariant risk minimization: An information theoretic view.》中针对第二个问题利用了信息论的理论提出了一下的目标函数:

$$\underset{P(Y|\Phi(X),\epsilon)=P(Y|\Phi(X))\forall\epsilon}{\arg\max} I(Y,\Phi(X))$$

即这篇论文指出在OOD问题中要发挥作用,不经要求特征$\Phi(X)$能够有invariant的能力并且,要与Y之间有最大的互信息

RELATED THEORETICAL WORKS

> Invariant risk minimization一文中,考虑了一种特殊情形即不变特征Φ与X之间是成线性关联的.
>
> Invariant rationalization一文中指出当隐藏的图模型满足某种特定的结构时,可以通过求解$\arg\max_{Y\perp\epsilon|M\odot X}I(Y;M\odot X)$来寻找OOD问题的解,其中M是一个二值的mask

这类方法一般假设invariant feature是X的一个子集或者说X的线性函数

contributions

- 给出了invariant Φ充分条件的证明,证明了使用最大不变预测器Maximal Invariant Predictor(MIP)来解决OOD问题的合理性。
- 利用Inter-Gradient-Alignment(IGA)来推到MIP

Notation and problem setting

令X,Y,ϵ分别代表输入,目标,环境三个随机变量.仅有变量X,Y是可观测的,并且按照ϵ来进行分组.使用函数D来描述预测器$f(X)$的表现.在某个环境ϵ中损失函数设计为:

$$L_\epsilon(f):=\mathbb{E}[D(f(X),Y)|\epsilon] \tag{2}$$

OOD问题可以表达为

$$\arg\min_f \max_{\epsilon\in\text{supp}(\mathcal{E})} L_\epsilon(f), \tag{3}$$

不变性假设表达为存在变量集合$\mathcal{I}$,使得下式成立.

$$\mathcal{I}:=\{\Phi\in\sigma(X);Y\perp\mathcal{E}|\Phi\}=\{\Phi\in\sigma(X);I(Y;\mathcal{E}|\Phi)=0\} \tag{4}$$

MAXIMAL INVARIANT PREDICTOR

Darmois的理论中指出当累计概率函数(cumulative distribution function)sufficiently regular时.对于任意的随机变量X, Y,可以找到一个噪声向量N_Y,满足$N_Y \perp X$,于是有Y=f(X,N_Y).关于sufficiently regular的定义chatgpt的解释如下,应该就是指函数的高阶导数存在.

In mathematics, a function or a curve is said to be "sufficiently regular" if it satisfies certain smoothness conditions. The precise definition of what constitutes sufficient regularity depends on the context and the particular application, but the concept is generally used in calculus, analysis, and geometry.

In calculus, a function is said to be sufficiently regular if it is differentiable a sufficient number of times. For example, a function that is twice differentiable (i.e., it has two continuous derivatives) is said to be "sufficiently regular" for many applications, as it can be used to model a wide variety of phenomena.

In analysis, a function is said to be sufficiently regular if it satisfies certain conditions of smoothness and continuity. For example, a function that is uniformly continuous or Lipschitz continuous is said to be sufficiently regular for many applications in analysis and optimization.

In geometry, a curve is said to be sufficiently regular if it is differentiable a sufficient number of times and has certain geometric properties such as curvature and torsion. For example, a curve that is twice differentiable and has nonzero curvature at every point is said to be "sufficiently regular" for many applications in differential geometry and curve theory.

The concept of sufficient regularity is important in many areas of mathematics and science, as it allows us to model and analyze a wide variety of phenomena using relatively simple mathematical structures.

那么将Darmois的理论应用与随机变量对(Φ, E)中,存在$\mathrm{E}_\psi \perp \Phi$,于是$\sigma(\Phi, \mathcal{E}) = \sigma(\Phi, \mathcal{E}_\psi)$.,存在$\mathcal{E}_\phi \perp \mathcal{E}_\psi$于是有$\sigma(\mathcal{E}, \mathcal{E}_\psi) = \sigma(\mathcal{E}_\phi, \mathcal{E}_\psi)$.

于是我们可以将环境因子E分解为$(\mathrm{E}_\phi, \mathrm{E}_\psi)$(分解为与$\Phi$相关与无关两部分):

- $\mathcal{E}_\psi \in \sigma(\mathcal{E})$, a part of $\mathcal{E}$ that is independent of Φ
- $\mathcal{E}_\phi = \mathcal{E}_\psi^c$; that is, $\mathcal{E}_\phi \perp \mathcal{E}_\psi$ and $\mathcal{E} \in \sigma(\mathcal{E}_\phi, \mathcal{E}_\psi)$.

考虑图像任务,若我们要根据照片X,来确定动物的分类Y.图像中的Φ显然是动物的物理外貌,那么会影响Φ的环境因子为E_Φ例如动物的物理物理状态以及性别特征等.不影响Φ的环境因子表示为E_ψ,例如摄影师

个人的摄影习惯等

maximal invariant predictior condition可以表达为:

$$\Phi^* \in \arg\max_{\Phi \in \mathcal{I}} I(Y;\Phi) = \arg\max_{I(Y;\mathcal{E}|\Phi)=0} I(Y;\Phi). \tag{5}$$

即Y与E之间的互信息当给定Φ时较小,但又要确保Y与Φ之间的互信息最大.若限制Φ的的空间为线性的,即$\{\Phi \in \mathcal{I}; \Phi = M \odot X\}$,其中M为二值的mask.那么这个结果就退化《Invariant Rationalization》一文中的结果.

INTER-ENVIRONMENTAL GRADIENT ALIGNMENT ALGORITHM

为了解决问题(5),作者提出了IGA算法,通过优化以下的目标函数来实现:

$$\arg\min_{\theta} \mathbb{E}[L_{\mathcal{E}}(\theta)] + \lambda \operatorname{trace}(\operatorname{Var}(\nabla_\theta L_{\mathcal{E}}(\theta))). \tag{6}$$

L_ϵ表示在环境ϵ在预测器输出的损失值,用$\mathcal{E}$作为下标表示要计算均值或者方差.使用(6)那就不需要单独计算Φ了.算法过程如下:

Algorithm 1 Inter-environmental Gradient Alignment Algorithm

Input: $q_\theta(y|x), \{D_\epsilon; \epsilon \in \mathcal{G}_{train}\}$
Return: q_θ

for each iteration **do**
  for ϵ_i in $\mathcal{G}_{train}$ **do**
    compute $L_{\epsilon_i}(\theta) = \hat{\mathbb{E}}[\log q_\theta(Y|X)|\epsilon_i]$
    compute $\nabla_\theta L_{\epsilon_i}(\theta)$
  end for
  compute $\hat{\mathbb{E}}[L_{\mathcal{E}}] := \frac{1}{|\mathcal{G}_{train}|} \sum_{\epsilon_i \in \mathcal{G}_{train}} L_{\epsilon_i}(\theta)$
  compute $\operatorname{trace}(\hat{\operatorname{Var}}(\nabla_\theta L_{\mathcal{E}}(\theta))) := \sum_{\epsilon_i \in \mathcal{G}_{train}} ||\nabla_\theta L_{\epsilon_i}(\theta) - \nabla_\theta \hat{\mathbb{E}}[L_{\mathcal{E}}]||^2$
  update θ by the gradient descent using the eq (6)
end for

Figure 2: IGA algorithm

其中$\mathcal{G}_{train} = \epsilon_1, \epsilon_2, ..., \epsilon_k \subset supp(\mathcal{E})$表示环境的集合,针对每个环境,用户可以专门搜集一个数据集$\{D_\epsilon; \epsilon \in \mathcal{G}_{train}\}$.在这里作者认为每个环境$\epsilon$对应的效应使用者是提前不知道的,所以$\epsilon$只是数据集的整数index即表示不同的数据集.

这篇论文是被拒稿的,审稿意见中大部分审稿人对其中涉及的定理也表示疑惑