Implicit Quantile Networks for Distributional Reinforcement Learning

Will Dabney [*1] Georg Ostrovski [*1] David Silver [1] Rémi Munos [1]

【强化学习 49】IQN

张楚珩
清华大学 交叉信息院博士在读

还是接着前两篇的工作，这里讲Implicit Quantile Network。

原文传送门

Dabney, Will, et al. "Implicit quantile networks for distributional reinforcement learning." arXiv preprint arXiv:1806.06923 (2018).

特色

神经网络的输入里面外加一个均匀分布采样的noise，使用神经网络来连续拟合整个分布，这样对于分布的表达能力更强，相应的实验效果也更好；distributional RL最关键的应用是做risk-aware的强化学习，IQN中可以直接对于相应的noise做变换，从而产生特定的风险偏好。

过程

1. 总览

Categorical DQN（C51，本专栏第47）使用均匀分布的多个格子上的概率密度来表示概率分布，使用KL散度（在该问题下即crossentropy）作为损失函数，在policy evaluation任务下，可以证明它在最小化cramer距离。

Quantile Regression（QR-DQN，本专栏第48）使用均匀分布的分位上的价值函数值来表示概率分布，使用quantile回归的损失函数，在policy evaluation任务下，可以证明它在最小化Wasserstein距离。

这里提出建立一个神经网络，输入为状态 s 和一个采样 $\tau \sim U[0,1]$ ，输出不同离散动作对应的价值函数分布的 τ 分位数。这样做有如下好处

- 可以通过调节神经网络的容量来决定对于分布的拟合精度，理论上能够以任意精度拟合价值函数的分布；
- 更充分地利用训练资源，训练资源多可以在计算时多采样 $\tau \sim U[0,1]$ ，获得更快的学习进度和更好的sample complexity；训练资源少也能够工作；
- 这样的表示方式能够便于后续对于学习到的分布的使用，比如能够容易得到具有特殊风险偏好的策略。

2. 强化学习中的风险如何表示

假设不同的事件发生 $\omega \in \Omega$ ，两个随机变量 X,Y 代表不同事件发生后带给你的效用（utility）。**风险偏好**的意思是对于有不同期望收益（均值）和不同风险（方差）的两个随机变量人们会更加倾向于选择哪个。

文中提到效用的独立性定理，有两个版本。

第一种是如果两个随机变量 X,Y 人们更偏向于前者，即 $X \succ Y$ ，那么对于任意的随机变量 Z 有 $\alpha F_X + (1+\alpha)F_Z \geq \alpha F_Y + (1+\alpha)F_Z$ ；这种假设产生的相应策略是

$$\pi(x) = \arg\max_a \mathop{\mathbb{E}}_{Z(x,a)}[U(z)].$$

其中 z 是从分布中采样出来的价值函数值， $U(\cdot)$ 代表效用函数，凸函数表示risk-seeking、凹函数表示risk-averse。

第二种是如果两个随机变量 X,Y 人们更偏向于前者，即 $X \succ Y$ ，那么对于任意的随机变量 Z 有 $\alpha F_X^{-1} + (1+\alpha)F_Z^{-1} \geq \alpha F_Y^{-1} + (1+\alpha)F_Z^{-1}$ ；这种假设产生的相应策略是

$$\pi(x) = \arg\max_a \int_{-\infty}^{\infty} z \frac{\partial}{\partial z}(h \circ F_{Z(x,a)})(z)\, dz.$$

其中 h 是distortion risk measure，通过不同的变换也可以实现不同的风险偏好。

3. IQN的结构

IQN结果主要是一个神经网络，输入为状态 x 和一个采样 $\tau \sim U[0,1]$ ，输出不同离散动作对应的价值函数分布的 τ 分位数 $Z_\tau(x,a) := F_Z^{-1}(\tau|x,a)$ 。

定义与风险倾向有关的价值函数

$$Q_\beta(x,a) := \mathop{\mathbb{E}}_{\tau \sim U([0,1])}\left[Z_{\beta(\tau)}(x,a)\right].$$

如果 $\beta(\cdot):[0,1] \to [0,1]$ 函数是单位映射，那么这个Q函数和之前定义的一样，是价值函数分布的期望；如果该函数为凸函数（或者在图像上都在单位映射下方），那么就等于往较差情况加了较大的权重，这就产生了risk-averse型的风险偏好；如果该函数为凹函数（或者在图像上都在单位映射上方），那么就等于往较好情况加了较大的权重，这就产生了risk-seeking型的风险偏好；该函数的具体选择后面再细讲。

可以定义在此价值函数下的贪心策略

$$\pi_\beta(x) = \arg\max_{a \in \mathcal{A}} Q_\beta(x,a).$$

实际计算中通过采样来得到该策略

$$\tilde{\pi}_\beta(x) = \arg\max_{a \in \mathcal{A}} \frac{1}{K} \sum_{k=1}^{K} Z_{\beta(\tilde{\tau}_k)}(x,a).$$

使用梯度下降优化如下损失函数

$$\mathcal{L}(x_t, a_t, r_t, x_{t+1}) = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa} \left(\delta_t^{\tau_i, \tau_j'} \right)$$

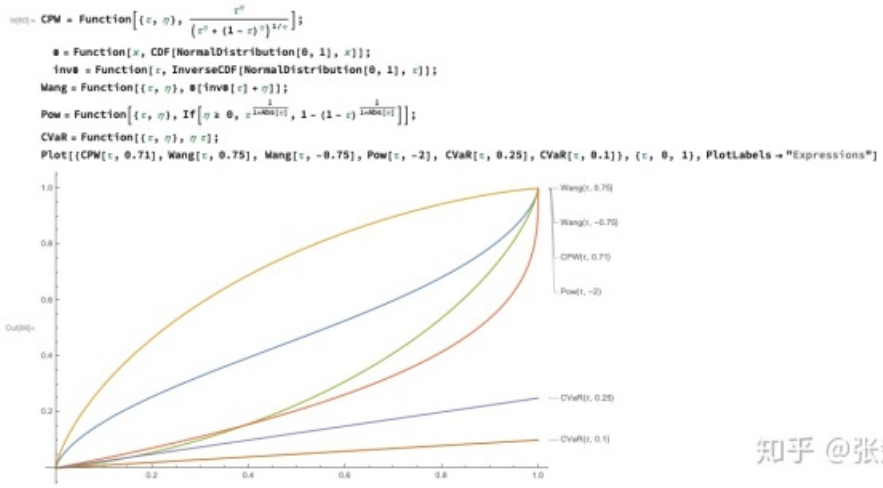
其中， $\rho(\cdot)$ 函数为前一讲提到的quantile regression的损失函数，TD误差为

$$\delta_t^{\tau,\tau'} = r_t + \gamma Z_{\tau'}(x_{t+1}, \pi_\beta(x_{t+1})) - Z_\tau(x_t, a_t)$$

该TD误差的前两项为目标，后一项为待优化的神经网络，即只对最后一项传播梯度。

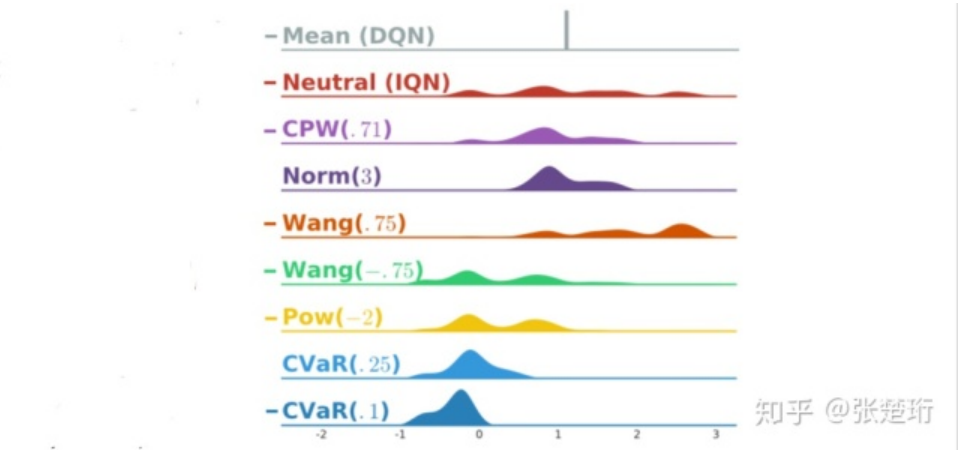
4. 风险调整函数 β 的选择

文章提供了以下若干种选择，不同的 $\beta(\cdot)$ 函数的函数图像如下图所示



如果原始的价值函数分布如下图Neutral所示，那么变换过之后的价值函数分布分别画在下图中。这个图可以这样理解，本来中性的价值函数分布是长Neutral这样的；在保守的人眼中，更关注这个分布中最差的情况是什么样，因此对于价值函数的判断更“悲观”，比如下面的CVaR；在激进的人眼中，更关注这个分布最最好的情况，即未来最优情形（有点像UCB，更注重探索），因此对于价值

函数的判断更“乐观”，比如下面的Wang(.75)。



相应的实验结果显示，使用risk-averse相关变换在Atari游戏上性能更好。不过在RL里面究竟应该使用risk-averse还是risk-seeking仍然是一个开放的问题，为什么Atari上面risk-averse性能更好也有待进一步探索。

5. 实现细节

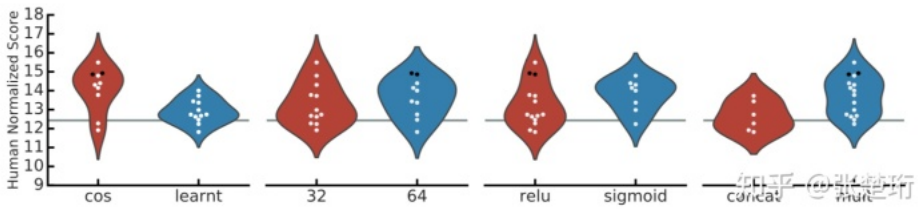
参数 K 影响不大，使用 $K=32$ ；参数 N 越大学习速率越快，Atari实验中 $N\geq 8$ 学习速率就比较合理；参数 N' 越大方差越小，实验中 $N'\geq 8$ 基本上饱和。

IQN可以写作

$$\mathrm{IQN}(x,\tau)=f(m(\psi(x),\phi(\tau))).$$

其中 $\psi:\mathcal{X}\to\mathbb{R}^d$ ， $\phi:[0,1]\to\mathbb{R}^d$ ， $f:\mathbb{R}^d\to\mathbb{R}^{|\mathcal{A}|}$ ， $m:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ 。

文章还研究了 ϕ 的表示（是按照余弦基函数加权还是MLP学习）、 ϕ 的隐含神经元个数、 f 中使用ReLU激活函数还是sigmoid激活函数、 m 是直接拼接还是两个向量相乘。结果如下图所示，图中有小黑点是是性能更好并且最后选择的方案。



实验结果

- 加上risk-averse调整之后，对于Atari游戏的性能有了一些提升
- 使用risk-neutral，相比于之前的QR-DQN有了很大的提升，比结合了包括C51在内的其他DQN技术的Rainbow差一些，但是IQN结合相关的其他技术应该性能上还会有较大的提升；
- 在本身AI玩的比人类水平差的游戏上提升较为明显

编辑于 2019-03-31

强化学习 (Reinforcement Learning)

文章被以下专栏收录