重要性采样(Importance Sampling)——TRPO与PPO的补充

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

上两篇博客已经介绍了信赖域策略优化(Trust Region Policy Optimization, TRPO)与近端策略优化算法(Proximal Policy Optimization Algorithms, PPO)，他们用到一个重要的技巧就是：重要性采样。但是都需要限制新旧策略使两者差异不能太大，TRPO通过添加新旧策略的KL约束项，而PPO是限制两者比率的变化范围，这究竟是为什么呢？不加这个约束会怎样？下面通过对重要性采样进行分析，来解答这个问题。更多强化学习内容，请看：随笔分类 - Reinforcement Learning。

1. 采样法(Sampling Method)/蒙特卡罗方法(Monte Carlo Method)

➢ 采样法(Sampling Method)/蒙特卡罗方法(Monte Carlo Method)

- 它是一种应用随机数来进行计算机模拟的方法。此方法对研究的系统进行随机观察抽样，通过对样本值的观察统计，求得所研究系统的某些参数。
- 给定一待推断的概率分布$p(x)$，并基于$p(x)$来计算函数$f(x)$的期望(以连续x为例)：

$$\mathbb{E}_{x\sim p}[f(x)]=\int_x p(x)f(x)dx$$

- 当比较复杂或难以精确推断时，可通过蒙特卡罗方法近似计算上述期望的解。
- 从给定概率密度函数$p(x)$中抽取出符合其概率分布的样本x_i (i=1, 2, …, N)，并将这些样本代入函数$f(x)$中。根据辛钦大数定律，当采样次数N足够大时，可以用平均值近似期望值，即

$$\int_x p(x)f(x)dx\approx\frac{1}{N}\sum_{i=1}^{N}f(x_i)$$

- 例如，定积分的定义是用曲边梯形的面积之和进行近似，此近似还可理解为蒙特卡罗方计算定积分(平均值法)

$$\int_a^b f(x)dx$$

- 设随机变量X服从(a, b)上的均匀分布，则$Y=f(X)$的数学期望为

$$\mathbb{E}[f(X)]=\int_a^b f(x)dx$$

- 由辛钦大数定律，可以用$f(X)$的观察值的平均估计其期望值。先用计算机产生N个(a, b)上均匀分布的随机数x_i (i=1, 2, …, N)，然后对每个x_i计算$f(x_i)$，最后得到上述定积分的估计值为：

$$\int_a^b f(x)dx\approx\frac{b-a}{N}\sum_{i=1}^{N}f(x_i)$$

茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程. 高等教育出版社, 2011.

2．重要性采样(Importance Sampling)

➢ 重要性采样(Importance Sampling)

- 有时候，直接从$p(x)$上是很难采样的，可以采用迂回战术，引入一个容易采样的分布$q(x)$，一般称为提议分布(Proposal Distribution)，则期望可以进一步写为：

$$\mathbb{E}_{x\sim p}[f(x)] = \int_x p(x)f(x)dx = \int_x \frac{p(x)}{q(x)}q(x)f(x)dx$$

$$= \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(x_i)}{q(x_i)}f(x_i)$$

- 这样，就可以从容易采样的分布$q(x)$中采样，并代入函数计算每个样本的$\frac{p(x_i)}{q(x_i)}f(x_i)$
- 与前面的蒙特卡罗方法的区别在于，$f(x_i)$前面多了一项 $\frac{p(x_i)}{q(x_i)}$ ，这一项称为重要性权重。
- 尽管期望相等，$\mathbb{E}_{x\sim p}[f(x)] = \mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right]$，但是方差呢？是否也相等呢？

➢ 重要性采样(Importance Sampling)

用到的公式

$$Var[X]=\mathbb{E}[X^2]-\left(\mathbb{E}[X]\right)^2$$

- 下面计算两者的方差。

$$Var_{x\sim p}[f(x)]=\mathbb{E}_{x\sim p}\left[f^2(x)\right]-\left(\mathbb{E}_{x\sim p}\left[f(x)\right]\right)^2$$

$$Var_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right]=\mathbb{E}_{x\sim q}\left[\left(\frac{p(x)}{q(x)}f(x)\right)^2\right]-\left(\mathbb{E}_{x\sim q}\left[\frac{p(x)}{q(x)}f(x)\right]\right)^2$$

$$=\int_x q(x)\frac{p^2(x)}{q^2(x)}f^2(x)dx-\left(\int_x q(x)\frac{p(x)}{q(x)}f(x)dx\right)^2=\int_x p(x)\frac{p(x)}{q(x)}f^2(x)dx-\left(\int_x p(x)f(x)dx\right)^2$$

$$=\mathbb{E}_{x\sim p}\left[\frac{p(x)}{q(x)}f^2(x)\right]-\left(\mathbb{E}_{x\sim p}\left[f(x)\right]\right)^2$$

- 可以看到两者的方差值第一项差了 $\frac{p(x)}{q(x)}$ 。只有当$p(x)=q(x)$时，两者的方差才会相等。因此，如果想用$q(x)$去采样近似替代$p(x)$，则应保证两个分布尽可能相似。

李宏毅, 强化学习课程, https://www.bilibili.com/video/BV1UE411G78S?spm_id_from=333.999.0.0, 2020.

3．重新思考TRPO与PPO

Importance Sampling

➢ 重新思考TRPO与PPO

- 保守策略迭代(Conservative Policy Iteration, CPI)

$$\max_{\theta} L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t\right]$$

- 信赖域策略优化(Trust Region Policy Optimization, TRPO)

$$\max_{\theta} L^{TRPO}(\theta) = \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t\right]$$

$$s.t.\ \hat{\mathbb{E}}_t\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t)\right)\right] \le \delta$$

- 近端策略优化(Proximal Policy Optimization, PPO)

Clipped PPO:

$$\max_{\theta} L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t, clip\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right]$$

- CPI目标函数就是通过重要性采样得到的，并未添加新旧策略约束项，在没有约束的情况下，最大化CPI目标函数将导致过大的策略更新。当新旧策略差别很大时，导致方差变化大，训练不稳定。TRPO与PPO这两中优化方法的思想都是尽量保证新旧策略之间差异不能过大，TRPO是通过添加约束项来约束新旧策略的KL散度小于一个值，而PPO则是通过裁剪函数限制新旧策略比率的变化范围。

李宏毅, 强化学习课程, https://www.bilibili.com/video/BV1UE411G78S?spm_id_from=333.999.0.0, 2020.

4. 参考文献

[1] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程. 高等教育出版社, 2011.
[2] 邱锡鹏，神经网络与深度学习，机械工业出版社，https://nndl.github.io/, 2020.
[3] 李宏毅, 强化学习课程, https://www.bilibili.com/video/BV1UE411G78S?spm_id_from=333.999.0.0, 2020.