6주차 PyTorch로 시작하는 강화학습 입문 Camp

출석 : 5명 (조교 2명 제외) ... 6명

커리큘럼

:: 금 주 진도

Part 1. Basics

> 2017년 이전 까지의 연구를 "기초"라 표현

- 강화학습 의 기반이 된 아이디어들 MDP, ML
- 가치 기반 방법론들 (Q-Learning, DQN)
- 정책 기반 방법론들(TRPO, PPO)

Part 2. Advanced (다양한 주제)

- Off-policy 정책기반 방법론들 (ACER, SAC)
- 탐험 기법들 (엔트로피 정규화, 내적 동기 부여)
- **강화학습과 불확실성 (Distributional DQN)**
    - Uncertainty Basic
    - Distributional RL

Part 3. Applications(적용의 확장)

- 모방학습, 역강화 학습 (GAIL, IRL)
- 다중 에이전트 강화 학습(Regret, MCTS)

피드백

- 강의는 5시 41분에 종료되었습니다.
- 추후 공부할 내용을 위해 라틴어도 알려주시는 친절함. 논문을 읽을 때 고려할 점을 중간중간 알려주시는게 참 좋습니다.
- 이번 강의에서도 질문이 많고 수업에 다들 참여 잘했습니다. 사람들이 적으니 더 열심히 하시는 분만 계시는 것 같습니다.
- 첫번째 시간에는 강의 내용이 널널 했고, 두번째 시간에는 강의 내용이 좀 많고 실습 설명까지 해서 40분 더 시간이 소요되었습니다. 강의 흐름 상 끊어져야했지만 수업시간이 많이 늦어지는게 자연스러운게 된거 같습니다.
- 강사님께서 질문을 많이 해도 된다고 해서 그런지 간단한 질문이라도 자주 해서 좋았습니다.

강의 내용

Uncertainty

: 학습이 끝나고 실제 활용 단계에서는 학습데이터와 다른 상황이 발생할 수 있다(1). 또는 데이터 자체의 변동이 큰 경우도 있다.(2) 그럼에도 에이전트가 행동하는 방식은 학습할 때와 동일하다. 이 두가지가 Uncertainty가 발생하는 큰 이유 2가지 이다.

- Model Uncertainty : Agent가 학습한 적 없는 데이터에 의한 불확실성 :: **Epistemic**
- Intrinsic Uncertainty : 학습 중에 봤더라도 여러 가능성이 있는 데이터 :: **Aleatoric**

Epsitemic Uncertainty

: 입력이 주어졌을 때, 모델이 결과에 대해서 얼마나 확신하는지를 고려하여 확신하지 못하는 만큼 다양한 결과를 만듦으로서 극복해볼 수 있다.

이를 위해 예측 때 마다 모델이 바뀌는 방법을 이용한다. (Dropout, BN)

기존의 학습, 결정론적인 가장 좋은 성능의 모델 하나를 찾는 것(**Point estimation**)에서 Dropout,BN을 사용하면 여러 모델을 가진 것(분포)이라고 생각할 수 있다. 이 때 학습의 목표는 모델이 데이터가 주어졌을 때의 Posterior 분포에 근사하면 된다. 아래 수식의 의미는 결과에 대한 불확실성은 고정된 모델에서 y가 나올 확률과 그 모델이 나올 확률을 고려한다는 것이다.

$$p(y|\mathcal{D}) = \int p(y|\theta)p(\theta|\mathcal{D})d\theta$$

이를 위한 방법 중 하나가 근사 추론(Approximation inference)를 사용하는 것이다.

근사 추론(Approximation inference)

1. Markov chain Monte-carlo(implicit) - 잘 사용되지 않는다.
2. Variational Inference(explicit) - 변분법적 추론이 잘 사용된다.

Variational Inference

우리가 다루기 쉬운 분포(approximation class) 중에서 사후분포와 가장 비슷한 분포를 찾는 것

여기서는 Dropout 또는 BN 등의 방법으로 학습한 모델의 분포가 사후분포와 가깝다는 것을 이용한다.

Aleatoric Uncertainty

관찰 할 수 있는 조건들이 같지만 결과가 다양한(Multi modal)한 경우 (동전던지기, 주사위던지기, 시험성적 등), 이런 경우는 데이터를 아무리 많이 모아도 확실한 모델링을 할 수가 없다.

Policy iteration -Policy gradient, Actor critic

Value iteration - Q-learning, DQN

다양성을 표현하는 방법

- 잠재변수 모델링 - VAE, GAN
- Categorical Parameterization(Softmax) - 변수가 1차원인 경우
- Cumulative Distribution Function (각 분위수에 해당하는 값을 예측) - 변수가 1차원인 경우

강화학습에서의 불확실성

에이전트에게 주어지는 불확실성은 : 행동을 결정한 뒤의 다음**상태**, 행동을 결정한 뒤의 **보상**

- 상태의 불확실성
    - 상태의 model uncertainty(Epistemic)에 관한 고민은 exploration에서 진행
    - 상태의 Intrinsic uncertainty(Aleatoric)에 관한 고민은 현재까지 **적다**

$\rightarrow$ 상태에 관한 연구는 model-based RL 맥락에서 진행된다고 한다.

- 보상의 불확실성
  - 보상의 model uncertainty에 관한 것 역시 exploration에서 진행
    - Bayesian RL
    - Thompson Sampling
  - 보상의 Intrinsic uncertainty에 관한 것
    - Distributional RL

	Model Uncertainty	Intrinsic Uncertainty
State	Intrinsic Motivation	??
Reward	Bayesian RL	Distributional RL

Value Distribution

하나의 보상은 전체 목적함수에 미치는 영향이 너무 작기 때문에 보상 대신 **수확**의 불확실성을 모델링한다.

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

수확에 영향을 주는 요소들

- 각 시점의 보상의 불확실성
- 각 시점의 보상에 영향을 주는 상태
- 각 시점의 상태에 영향을 주는 정책

가치 분포(Value Distribution)

가치함수와 같이 매 상태, 행동에 대해 정의

가치함수와 똑같이 이후 행동은 기존의 정책을 따른다 가정

또한 다음 상태는 전이 확률에 의해 결정된다고 가정

- 가치분포의 평균 = 행동 가치

$$Q^{\pi}(x,a) := \mathbb{E}Z^{\pi}(x,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)]$$

$x_t \sim P(\cdot|x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot|x_t), x_0 = x, a_0 = a$

이전에 했던 것 처럼 Policy iteration, Value Iteration 알고리즘을 적용

$$T^{\pi}Z(x,a) :\overset{D}{=} R(x,a) + \gamma P^{\pi}Z(x,a)$$

$P^{\pi}Z(x,a) :\overset{D}{=} Z(X',A')$
$X' \sim P(\cdot|x,a), A' \sim \pi(\cdot|X')$

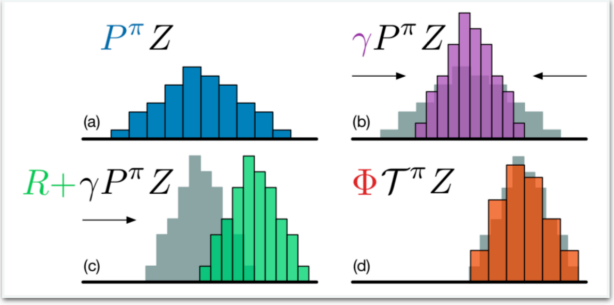
위 수식은 Target 분포는 reward의 분포에 그 다음 상태 행동의 분포에 γ 를 곱하여 더한다. (0~1사이 값을 곱하면 분포는 좀 더 좁아진다.)

Distributional Backup

Policy evaluation, Value Iteration에서 했던 것 처럼 좌변을 우변으로 대체. 다만 여기서는 **분포를 대체**

$$\mathcal{T}^\pi Z(x,a) :\overset{D}{=} R(x,a) + \gamma P^\pi Z(x,a).$$

$$P^\pi Z(x,a) :\overset{D}{=} Z(X',A')$$
$$X' \sim P(\cdot\,|\,x,a),\ A' \sim \pi(\cdot\,|\,X'),$$



고정된 정책에 대해서 Distributional Backup은 가치 분포를 찾는다.

- Policy Evaluation 은 가능하지만,
- Value Iteration처럼 Greedy Update를 바로 적용하는 것은 안된다. 이를 근사적으로 한 것이 Categorical DQN

Categorical DQN(C51)

1. 가치의 분포를 categorical 분포로 표현
    - Categorical 분포로 표현하는데 범위와 갯수가 필요하다.
        - 범위 : (최댓값, 최솟값), 각 경우마다 일어날 확률을 예측
2. 다음 상태의 Greedy action에 대해 행동가치분포를 가져오고
3. 다음 상태의 Greedy action에 대해 행동 가치 분포를 가져와서, Discount rate를 곱해주고
4. 보상만큼 더해준 것을 Target으로 만든다.
5. 이에 대한 Loss를 최소화 할 때, 두 분포의 범위가 달라서 Cross entropy계산은 불가능
6. 따라서, Categorical Projection을 해서 두 분포의 범위를 동일하게 맞춰준다.

몇 가지 가정을 추가하면 최적 정책의 가치 분포를 찾을 수 있기 때문에 어떤 상황에서는 Distributional Value Iteration을 해도 된다.

Quantile Regression DQN(QR-DQN)

CDF and Quantiles

Categorical DQN의 문제점

- Categorical 분포를 쓰기 때문에 최댓값, 최솟값, 이산화의 가정을 한다.
- 게임이라면 수확의 범위를 제한할 수 있어서 가능하지만 수확의 범위를 모르는 경우에는 가정이 깨진다.

누적분포함수(Cumulative Distribution Function)\

: 어떤 값보다 작거나 같을 확률을 나타내는 함수이고, 값이 커질 수록 확률이 커지기 때문에 증가 함수이다.

분위수 (Quantiles)

: 누적분포함수와는 반대로 확률을 넣으면 그보다 큰 누적확률을 가지는 값중에서 가장 작은 값을 알려준다. 거의 대부분의 경우에는 그 누적확률에 해당하는 값을 알려준다.

QR-DQN

가치분포는 연속적이고 가치분포의 Quantile들을 학습하게 되면 이산적인 부분만 표현가능하다.

- Target : 다음 상태 가치 분포의 Quantile들
- Prediction : 현재 상태 가치 분포의 Quantile들을 가지고 만든 현재 상태 가치 분포를 가장 잘 나타내는 이산분포

  그냥 이산 분포를 학습 시키는 것

Quantile Projection

: 어떤 분포를 가장 잘 나타내는 이산분포

- 그 분포의 Quantile 값들을 그대로 쓰는게 아니라 Quantile midpoint(이전 quantile과 다음 quantile을 평균 낸 값)을 Quantile로 가지는 이산분포가 가장 잘 근 사함

Quantile Regression Loss

Quantile Projection을 통해 얻은 이산 분포에게 주어지는 loss

Overestimation/Underestimation 여부에 따라 다르게 loss를 정의

$$\mathcal{L}_{QR}^{T}(\theta) := \mathbb{E}_{\hat{Z}\sim Z}[\rho_{\tau}(\hat{Z}-\theta)], where$$
$$\rho_{\tau}(u) = u(\tau - \delta_{(u<0)}), \forall u \in \mathbb{R}$$

:: Underestimation여부를 체크(일어나면1, 아니면 0)(예측분포가 밑에 있는지)

Implicit Quantile Networks(IQN)

연속적인 분포를 표현하기 위해서(Quantile Projection 부분 생략가능) Quantile도 입력으로 줘서 값들을 예측 (QR-DQN에서는 벡터 상의 위치가 Quantile을 의미했었음)

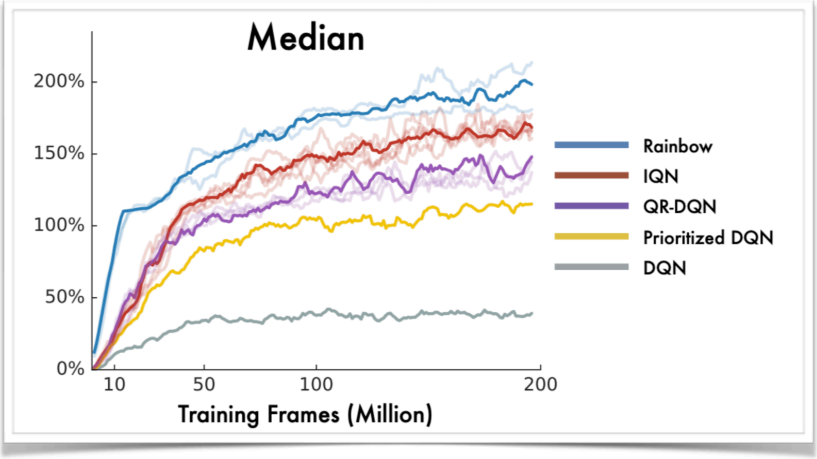
Quantile Encoding

: Convolution Layer을 통해 뽑은 Feature와 Quantile code를 곱하고 FFNN

: 최종적으로 예측하는 것은 input으로 주어진 Quantile에 대한 예측값이다.

$$IQN(x,\tau) = FFNN(m(\psi(x),\phi(\tau)))$$
$$\phi_j(\tau) := ReLU(\sum_{i=0}^{n-1}\cos(\pi i\tau)w_{ij}+b_j)$$

<각 방법론 성능 비교>



질문

MCBN - 색있는 부분의 면적은 실제로 무슨 의미?

- eval 시에 표준편차*1 는 짙은 색, 표준편차*2는 연한색으로 색칠

기존의 머신러닝들이 평균만 알려준다. 평균만 알려준다는 것의 의미는?

- 예를 들어, 주사위를 던지면 어떤 면이 나올까 질문 했을 때, 답으로 3.5라 대답하는 상황

기존 통계학에서는 모달이 여러개 있으면 여러개의 분포로 설명하려고 하는데 여기서는 꼭 왜 하나로 만들어서 하는가?

- 예로 가우시안믹스쳐 모델링 하더도 다양성을 확보하지 못한 경우가 많다. 그리고 머신러닝 입장에서는 하나의 모델을 사용할 수 있으니까 그렇다.

Distributed RL은 보상의 분포를 학습하는 것인가?

- value의 멀티모달 분포를 학습하는 것이다.

Uncetainty 유형에 따른 표에서 ??? 는?

- 아직 특별히 도미넌트한 연구가 없다. 굳이 보려면 model_based RL

상태를 고정시킨다. 근데 상태는 너무 많지 않나?

- 상태에서 행동을 했을 때 지금까지는 평균을 냈는데, 여기서는 각 상태 행동 쌍에 대해 분포를 보겠다는 것이다.

다구간 로스들의 평균을 내는 것인가?

- 그렇다.