Name: Yuxuan Zhang (yuxuanz8)

University of Illinois Spring 2020

CS 446/ECE 449 Machine Learning
Homework 10: REINFORCE

Due on Tuesday May 5 2020, noon Central Time

1. **[16 points]** REINFORCE

We are given a utility $U(\theta) = \mathbb{E}_{p_\theta}[R(y)] = \sum_{y\in\mathcal{Y}} p_\theta(y)R(y)$ which is the expected value of the non-differentiable reward $R(y)$ defined over a discrete domain $y \in \mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$. Our goal is to learn the parameters θ of a probability distribution $p_\theta(y)$ so as to obtain a high utility (high expected reward), *i.e.*, we want to find $\theta^* = \arg\max_\theta U(\theta)$. To this end we define the probability distribution to read

$$p_\theta(y) = \frac{\exp F_\theta(y)}{\sum_{\hat{y}\in\mathcal{Y}} \exp F_\theta(\hat{y})}. \tag{1}$$

(a) (3 points) If we are given an i.i.d. dataset $\mathcal{D} = \{(y)\}$ we can learn the parameters θ of a distribution via maximum likelihood, *i.e.*, by addressing

$$\max_\theta \sum_{y\in\mathcal{D}} \log p_\theta(y)$$

via gradient descent. State the cost function and its gradient when plugging the model specified in Eq. (1) into this program. When is this gradient zero?

Your answer:

$$\max_\theta \sum_{y\in\mathcal{D}} \log P_\theta(y) = \max_\theta \sum_{y\in\mathcal{D}} \log \frac{\exp F_\theta(y)}{\sum_{\hat{y}\in\mathcal{Y}} \exp F_\theta(\hat{y})}$$

$$\mathcal{L} = -\sum_{y\in\mathcal{D}} \left[F_\theta(y) - \log \sum_{\hat{y}\in\mathcal{Y}} \exp F_\theta(\hat{y})\right]$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{y\in\mathcal{D}} \left[\frac{\partial F_\theta(y)}{\partial \theta} - \frac{\sum_{\hat{y}\in\mathcal{Y}} \exp F_\theta(\hat{y}) \frac{\partial F_\theta(\hat{y})}{\partial \theta}}{\sum_{\hat{y}\in\mathcal{Y}} \exp F_\theta(\hat{y})}\right] = 0$$

$$\sum_{y\in\mathcal{D}} \left[\frac{\partial F_\theta(y)}{\partial \theta} - \sum_{\hat{y}} P_\theta(\hat{y}) \frac{\partial F_\theta(\hat{y})}{\partial \theta}\right] = 0$$

$$\sum_{y\in\mathcal{D}} \frac{\partial F_\theta(y)}{\partial \theta} - \sum_{y\in\mathcal{D}} |\mathcal{D}| P_\theta(\hat{y}) \frac{\partial F_\theta(y)}{\partial \theta} = 0.$$

$$\sum_{y\in\mathcal{D}} \frac{\partial F_\theta(y)}{\partial \theta} (1 - |\mathcal{D}| P_\theta(y)) = 0$$

$P_\theta(y) = \frac{1}{|\mathcal{D}|}$

(b) (2 points) If we aren't given a dataset but if we are instead given a reward function $R(y)$ we search for the parameters θ by maximizing the utility $U(\theta)$, *i.e.*, the expected reward. Explain how we can approximate the utility by sampling from the probability distribution $p_\theta(y)$.

Your answer:

$$\nabla_\theta U(\theta) \approx \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P_\theta(y) R(y).$$

We can randomly sample $\hat{y}$ from $P_\theta(y)$ and use average as an estimation.

Name: Yuxuan Zhang (yuxuanz8)

(c) (3 points) Using general notation, what is the gradient of the utility $U(\theta)$ w.r.t. θ, *i.e.*, what is $\nabla_\theta U(\theta)$. How can we approximate this value by sampling from $p_\theta(y)$? Make sure that you stated the gradient in the form which ensures that computation via sampling from $p_\theta(y)$ is possible.

Your answer: $U(\theta) = \sum_{y \in \mathcal{Y}} P_\theta(y) R(y)$

$\nabla U(\theta) = \sum_{y \in \mathcal{Y}} \nabla P_\theta(y) R(y)$

$= \sum_{y \in \mathcal{Y}} \frac{P_\theta(y)}{P_\theta(y)} \nabla P_\theta(y) R(y)$

$= \sum_{y \in \mathcal{Y}} P_\theta(y) \nabla \log P_\theta(y) R(y)$

$\nabla U(\theta) \approx \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P_\theta(y) R(y)$

(d) (5 points) Using the parametric probability distribution defined in Eq. (1), what is the approximated gradient of the utility? How is this gradient related to the result obtained in part (a)?

Your answer: $P_\theta = \exp F_\theta(y) / \sum_{\hat{y} \in \mathcal{Y}} \exp F_\theta(\hat{y})$

$\nabla U(\theta) \approx \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \nabla_\theta \log \frac{\exp F_\theta(y)}{\sum_{\hat{y}} \exp F_\theta(\hat{y})} \cdot R(y)$.

$= \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \left(\frac{\partial F_\theta(y)}{\partial \theta} - \frac{\sum_{\hat{y}} \exp F_\theta(\hat{y}) \frac{\partial F_\theta(\hat{y})}{\partial \theta}}{\sum_{\hat{y}} \exp F_\theta(\hat{y})} \right) R(y)$

(number of y in $\mathcal{Y}$)

This gradient is used $R(y)$ as the weight to $\nabla_\theta \log P_\theta(y)$.

(a) part just use 1 and compute the ~~mean~~ the sum of ~~$\nabla_\theta P_\theta(y)$~~. $\nabla_\theta \log P_\theta(y)$

(e) (3 points) In `A10_Reinforce.py` we compare the two forms of learning. Let the size of the domain $|\mathcal{Y}| = 6$, and let the groundtruth data distribution $p_{\text{GT}}(y) = 1/12$ for $y \in \{1, 6\}$, $p_{\text{GT}}(y) = 2/12$ for $y \in \{2, 5\}$, and $p_{\text{GT}}(y) = 3/12$ for $y = \{3, 4\}$. The dataset $\mathcal{D}$ contains $|\mathcal{D}| = 1000$ points sampled from this distribution. Further let $F_\theta(y) = [\theta]_y$, where $\theta \in \mathbb{R}^6$ and where $[a]_y$ returns the y-th entry of vector a. The reward function happens to equal the groundtruth distribution, *i.e.*, $R(y) = p_{\text{GT}}(y)$. What distribution p_θ is learned with the maximum likelihood approach? What distribution is learned with the REINFORCE approach? Explain why this is expected. Complete `A10_Reinforce.py` to answer these questions.

Your answer:

Maximum Likelihood. ~~[-0.5661, 0.0030, 0.4254, 0.4636, 0.0989, -0.6963]~~
[-0.4159, 0.1532, 0.5856, 0.6138, 0.2491, -0.5461]

Reinforcement Learning: [0.0966, 0.1995, 0.2640, 0.1278, 0.2297, 0.0823]

The ML method used gradient descending to compute the θ P_θ.
although the difference is descending, but it will get negative value.
It did not take ~~any~~ the performance of inner state θ into consideration.
just ~~the~~ use the mean to approximate.
The RL gave rewards to the distribution, if the approximated probability is close to true value, then it would stick to it next, otherwise, it will improve it.