Exercise 9

Kai Schneider

June 29, 2021

Task 1 Reinforce on the Cart-Pole

a.)

softmax: $\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,a,\theta)}}$

linear features: $h(s,a,\theta) = \theta_a^T s$

for the 2 actions space $a \in \{0,1\}$ of the cart pole this results in $\pi(a|s,\theta) = \frac{e^{\theta_a^T s}}{e^{\theta_0^T s} + e^{\theta_1^T s}}$ for the softmax function. With state $s = (p, \dot{p}, \alpha, \dot{\alpha})^T \in \mathbb{R}^4$ follows $\theta \in \mathbb{R}^4$
This results for the given actions in:

$$\pi(a=0|s,\theta) = \frac{e^{\theta_0^T s}}{e^{\theta_0^T s} + e^{\theta_1^T s}} \quad \text{and} \quad \pi(a=1|s,\theta) = \frac{e^{\theta_1^T s}}{e^{\theta_0^T s} + e^{\theta_1^T s}}$$

These can be rewritten as a sigmoid function:

$$\begin{aligned}
\pi(a=0|s,\theta) &= \frac{e^{\theta_0^T s}}{e^{\theta_0^T s} + e^{\theta_1^T s}} \\
&= \frac{e^{\theta_0^T s}}{e^{\theta_0^T s}\left(e^{\theta_1^T s - \theta_0^T s} + 1\right)} \\
&= \frac{1}{e^{\theta_1^T s - \theta_0^T s} + 1} \\
&= \sigma(\theta_0^T s - \theta_1^T s) \\
&= \sigma(s^T(\theta_0 - \theta_1))
\end{aligned}$$

similar for $\pi(a=1|s,\theta) = \sigma(s^T(\theta_1 - \theta_0))$
Also due to $\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$ we have

$$\begin{aligned}
\pi(a=0|s,\theta) &= 1 - \pi(a=1|s,\theta) \\
&= 1 - \sigma(s^T(\theta_1 - \theta_0))
\end{aligned}$$

and vice versa

$$\pi(a=1|s,\theta) = 1 - \sigma(s^T(\theta_0 - \theta_1))$$

With $\sigma'(z) = \sigma(z)\big(1-\sigma(z)\big)$ being the derivative of the sigma function we get:

$$\begin{aligned}
\nabla_\theta \pi(a=0|s,\theta) &= \sigma(s^T(\theta_0-\theta_1))\big(1-\sigma(s^T(\theta_0-\theta_1))\big)s \\
\nabla_\theta \pi(a=1|s,\theta) &= \sigma(s^T(\theta_1-\theta_0))\big(1-\sigma(s^T(\theta_1-\theta_0))\big)s
\end{aligned}$$

b.)

The derivate $\nabla_\theta \log \pi(a=0|s,\theta)$ can be calculated in a similar fashion:

$$\begin{aligned}
\nabla_\theta \log \pi(a=0|s,\theta) &= \frac{1}{\pi(a=0|s,\theta)}\sigma(s^T(\theta_0-\theta_1))\big(1-\sigma(s^T(\theta_0-\theta_1))\big)s \\
&= \frac{1}{\sigma(s^T(\theta_0-\theta_1))}\sigma(s^T(\theta_0-\theta_1))\big(1-\sigma(s^T(\theta_0-\theta_1))\big)s \\
&= \big(1-\sigma(s^T(\theta_0-\theta_1))\big)s
\end{aligned}$$

and

$$\nabla_\theta \log \pi(a=1|s,\theta) = \big(1-\sigma(s^T(\theta_1-\theta_0))\big)s$$

This can again be rewritten with the results from **a.)**:

$$\begin{aligned}
\nabla_\theta \log \pi(a=0|s,\theta) &= \big(1-\sigma(s^T(\theta_0-\theta_1))\big)s \\
&= \sigma(s^T(\theta_1-\theta_0))s
\end{aligned}$$

and

$$\begin{aligned}
\nabla_\theta \log \pi(a=1|s,\theta) &= \big(1-\sigma(s^T(\theta_1-\theta_0))\big)s \\
&= \sigma(s^T(\theta_0-\theta_1))s
\end{aligned}$$