Artificial Neural Networks (Gerstner). Exercises for week 5

Policy gradient methods

Exercise 1. Single neuron as an actor[1]

Assume an agent with binary actions $Y \in \{0,1\}$. Action $y = 1$ is taken with a probability $\pi(Y=1|\vec{x};\vec{w}) = g(\vec{w}\cdot\vec{x})$, where $\vec{w}$ are a set of weights and $\vec{x}$ is the input signal that contains the state information. The function g is monotonically increasing and limited by the bounds $0 \leq g \leq 1$.

For each action, the agent receives a reward $R(Y,\vec{x})$.

a. Calculate the gradient of the mean reward $\mathbb{E}[R] = \sum_{Y,\vec{x}} R(Y,\vec{x})\pi(Y|\vec{x};\vec{w})P(\vec{x})$ with respect to the weight w_j.

   Hint: Insert the policy $\pi(Y=1|\vec{x};\vec{w}) = g(\sum_k w_k x_k)$ and $\pi(Y=0|\vec{x};\vec{w}) = 1 - g(\sum_k w_k x_k)$. Then take the gradient.

b. The rule derived in (a) is a batch rule. Can you transform this into an 'online rule'?

   Hint: Pay attention to the following question: what is the condition that we can simply 'drop the summation signs'?

Exercise 2. Policy gradient for binary actions

a. Find an online policy gradient rule for the weights $\vec{w}$ for the same setup as in Exercise 1 by calculating the gradient of the log-likelihood $\log\pi(Y|\vec{x};\vec{w})$ with respect to the weights.

   Hint: the policy π can be written as $\pi(Y|\vec{x};\vec{w}) = (1-\rho)^{1-Y}\rho^Y$ with $\rho = g(\vec{w}\cdot\vec{x})$.

b. Rewrite your update rule for weight w_j in the form

$$\Delta w_j = F(\vec{x},\vec{w},R)\,[Y - \mathbb{E}[Y]]\,x_j$$

   and give the expression for the function F.

   Hint: Take your result from part a, use $\mathbb{E}[y] = g(\vec{w}\cdot\vec{x})$ and pull out a factor $\frac{1}{g(1-g)}$.

Exercise 3. Policy gradient

a. **Other parameterizations of Exercise 2**: Consider your solution to Exercise 2. What happens to the policy gradient rule if the likelihood ρ of action 1 is parameterized not by the weights $\vec{w}$ but by other parameters: $\rho = \rho(\theta)$? Derive a learning rule for θ.

b. **Generalization to the natural exponential family**: The natural exponential family is a family of probability distributions that is widely used in statistics because of its favorable properties. These distributions can be written in the form

$$p(Y) = h(Y)\exp\left(\theta Y - A(\theta)\right).$$

   This family includes many of the standard probability distributions. The Bernoulli, the Poisson and the Gaussian distribution are all member of this family. A nice property of these distributions is that the mean can easily be calculated from the function $A(\theta)$:

$$\mathbb{E}[Y] = A'(\theta) := \frac{dA}{d\theta}(\theta).$$

[1]Will be started in class.

Assume that the policy $\pi(Y|\vec{x};\theta)$ is an element of the natural exponential family. Show that the online rule for the policy gradient has the shape:

$$\Delta\theta = R(Y - \mathbb{E}[Y])\,.$$

Can you give an intuitive interpretation of this learning rule?

c. **The Bernoulli distribution**: Apply your result from (b) to the case of Exercise 2.

Exercise 4. Subtracting the mean

You have two stochastic variables, x and y with means $\mathbb{E}[x]$ and $\mathbb{E}[y]$. Angles denote expectations. We are interested in the product $z = (x - b)(y - \mathbb{E}[y])$ with a fixed parameter b.

a. Show that $\mathbb{E}[z]$ is independent of the choice of the parameter b.

b. Show that $\mathbb{E}[z^2]$ is minimal if $b = \frac{\mathbb{E}[xf(y)]}{\mathbb{E}[f(y)]}$, where $f(y) = (y - \mathbb{E}[y])^2$.

   Hint: write $\mathbb{E}[z^2] = F(b)$ and set $\frac{dF}{db} = 0$.

c. What is the optimal b, if x and $f(y)$ are approximately independent?

d. Make the connection to policy gradient rules.

   Hint: take $x = r$ (reward) and y the action taken in state s. Compare with the policy gradient formula of the simple 1-neuron actor. What can you conclude for the best value of b? Consider different states s. Why should b depend on s?

Exercise 5. Computer exercises: Environment 2 (part 1)[1]

Download the Jupyter notebook of the 2nd computer exercise and complete it until the end of Section 1.3.4 (Reinforce with Baseline).

[1] Start this exercise in the second exercise session of week 5.