Assignment #3

BMI 539: Introduction to Reinforcement Learning

Fall 2024

A Bernoulli-logistic unit is a stochastic neuron-like unit used in some artificial neural networks. Its input at time t is a feature vector $\mathbf{x}(S_t)$; its output, A_t, is a random variable having two values, 0 and 1 , with $\Pr\{A_t = 1\} = P_t$ and $\Pr\{A_t = 0\} = 1 - P_t$ (the Bernoulli distribution).

Let $h(s, 0, \boldsymbol{\theta})$ and $h(s, 1, \boldsymbol{\theta})$ be the preferences in state s for the unit's two actions given policy parameter $\boldsymbol{\theta}$. Assume that the difference between the action preferences is given by a weighted sum of the unit's input vector, $h(s, 1, \boldsymbol{\theta}) - h(s, 0, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}(s)$, where $\boldsymbol{\theta}$ is the unit's weight vector.

(a) Show that if the exponential soft-max distribution $\pi(a \mid s, \boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}}$ is used to convert action preferences to policies, then $P_t = \pi(1 \mid S_t, \boldsymbol{\theta}_t) = 1/\left(1 + \exp\left(-\boldsymbol{\theta}_t^\top \mathbf{x}(S_t)\right)\right)$

(b) What is the Monte-Carlo REINFORCE update of $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ upon receipt of return G_t ?

(c) Express the eligibility $\nabla \ln \pi(a \mid s, \boldsymbol{\theta})$ for a Bernoulli-logistic unit, in terms of $a, \mathbf{x}(s)$, and $\pi(a \mid s, \theta)$ by calculating the gradient.

Hint for part (c): Define $P = \pi(1 \mid s, \boldsymbol{\theta})$ and compute the derivative of the logarithm, for each action, using the chain rule on P. Combine the two results into one expression that depends on a and P, and then use the chain rule again, this time on $\boldsymbol{\theta}^\top \mathbf{x}(s)$, noting that the derivative of the logistic function $f(x) = 1/(1 + e^{-x})$ is $f(x)(1 - f(x))$.