AIML 268 homework 1

Adam Guo 2020–07–23

1. Prove that $\mathbb{KL}(p \,||\, q) \geq 0$ with equality if and only if $p = q$.

   Solution: Recall Jensen's inequality, which states that for a convex function f,

   $$f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i),$$

   where $\lambda_i \geq 0$, $\sum_{i=1}^{n} \lambda_i = 1$.
   Let $A = support(p), \mathcal{X} = support(q)$ for probability density functions p and q. Then,

   $$\begin{aligned}
   -\mathbb{KL}(p \,||\, q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
   &= \sum_{x \in A} p(x)(\log \frac{q(x)}{p(x)}) \\
   \mathbb{KL}(p \,||\, q) &= \sum_{x \in A} p(x)(-\log \frac{q(x)}{p(x)}) \\
   &\geq -\log\left(\sum_{x \in A} p(x) \frac{q(x)}{p(x)}\right) && \text{via Jensen's inequality, since } -\log \text{ is convex} \\
   &\geq -\log\left(\sum_{x \in A} q(x)\right) \\
   &\geq -\log\left(\sum_{x \in \mathcal{X}} q(x)\right) && \text{since } q(x) = 0 \text{ for all } x \in A - \mathcal{X} \\
   &\geq -\log(1) \\
   &\geq 0
   \end{aligned}$$

   Jensen's inequality also states that

   $$f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i)$$

   iff $x_1 = x_2 = \cdots = x_n$. Thus,

   $$\mathbb{KL}(p \,||\, q) = -\log\left(\sum_{x \in A} p(x) \frac{q(x)}{p(x)}\right) = 0 \text{ iff } \frac{q(x_1)}{p(x_1)} = \frac{q(x_2)}{p(x_2)} = \cdots = \frac{q(x_n)}{p(x_n)}$$

   Since $\sum_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} q(x_i) = 1$, $p(x_i) = q(x_i)$ for all $i = 1, \ldots, n$. Thus, $\mathbb{KL}(p \,||\, q) = 0$ iff $p = q$.

2. Find a paper with code that uses the KL-divergence, run the code on your computer, and discuss the code and results below.

   Solution: Agarwal, Liang, Schuurmans, Norouzi: Learning to Generalize from Sparse and Underspecified Rewards `https://arxiv.org/pdf/1902.07198v4.pdf`

   The authors consider the general problem of designing an agent that, given a natural language question, attempts to learn the correct behaviour that answers that question, while only having binary feedback

(true or false). This feedback is considered "underspecified" since it does not indicate whether the agent identified the correct pattern — it only tells the agent whether its current guess happens to be correct.

KL divergence is used to define the objective that the agent strives toward when optimising its answering policy. Let x be the question given to the agent, let α be the agent's response, and let $R(\alpha \mid x, y) \in \{0, 1\}$ denote whether or not α represents a success or failure in the context of x, with the goal specification y (for instance, y could be the answer to the question x, and R checks whether $\alpha = y$). The aim is to optimise a stochastic policy $pi(\alpha \mid x)$ that maximises the agent's success rate.

Let π^* be an optimal policy. The KL divergence between the optimal policy and parametric policy, i.e. $\mathbb{KL}(\pi^* \mid\mid \pi)$, provides one objective that promotes "mode covering behaviour" (which encourages the agent to explore all possible correct answers α with equal probability), while $\mathbb{KL}(\pi \mid\mid \pi^*)$ provides another objective that promotes "mode seeking behaviour" which encourages the agent to explore trajectories with higher marginal probability of success. The paper defines a new objective that it optimises toward based on these existing metrics, termed "meta reward-learning".

The code evaluates this approach on an agent exploring a maze with traps, attempting to escape without landing on any traps. With no prior information, the agent is given a natural language instruction as x outlining the optimal path it should take (e.g. "left right up"), and it is expected to take the corresponding path to exit the maze. With underspecified rewards (i.e. mazes with several possible trajectories that are not the optimal path), the agent achieves 69.8% success, while addition of the meta reward-learning algorithm improves accuracy to 74.5%.