TTIC 31020: Introduction to Machine Learning
Autumn 2022

Problem Set #5

(Collaborated with Sam Zhong, Haichuan Wang)

1 Information and learning

Here we will try to get a closer understanding of the relationship between information theoretic measures we have discussed in class and the objectives in machine learning.

Problem 1 **[17 points]**
Let $X \in [m]$ be a random variable, taking one of m according to some distribution $P \in \mathbb{S}^m$. I.e., P is a point on the m-*simplex*, in other words, P is an m-dimensional vectors whose m elements are all between 0 and 1, and sum to 1.

Show that the P maximizing the entropy $H(X)$ is the uniform distribution over $[m]$, i.e., $P(i) = 1/m$ for all $i \in [m]$.

Advice: There are at least two ways to do this problem. One is to set up an optimization problem, use Lagrange multipliers, and solve. Another is to let U be the uniform distribution over $[m]$, calculate $D_{KL}(U \,||\, P)$ for any arbitrary distribution P, and use this to conclude the desired result. The second method is cleaner.

End of problem 1

- Let P be any arbitrary distribution $[p_1, \ldots, p_m]$.
  Let U be the uniform distribution $[\frac{1}{m}, \ldots, \frac{1}{m}]$.

- $D_{KL}(P||U) = -\sum_{i=1}^{m} p_i \log \frac{1}{m} + \sum_{i=1}^{m} p_i \log p_i$

  $= -\log \frac{1}{m} + \sum_{i=1}^{m} p_i \log p_i$

- As $H(x) = -\sum_{i=1}^{m} p_i \log p_i$ is the entropy of an arbitrary distribution, we want

  $\max\{H(x)\} = \max\{-D_{KL}(P||U) - \log \frac{1}{m}\}$

  $= \min\{D_{KL}(P||U)\} - \log \frac{1}{m}$

- As $D_{KL}(P||U)$ achieves a minimum of zero at $P = U$, we obtain that the entropy maximizing distribution is U.

Problem 2 **[17 points]**

Consider a parameter estimation problem where we wish to maximize conditional log-likelihood of a predictive model on a labeled (classification) data set. That is, we're given the population $\{(\mathbf{x}_i, y_i)\}$, and a parametric likelihood function $p(y|\mathbf{x};\theta)$. show that the objective of *maximizing* the conditional log-likelihood of the model on the training data is equivalent to the objective of *minimizing* the KL divergence between the empirical distribution of the data $P_{\text{data}}(\mathbf{x}, y)$ and the posterior distribution predicted by the model:

$$\underset{\theta}{\operatorname{argmax}} \sum_i \log p(y_i|\mathbf{x}_i;\theta) \;=\; \underset{\theta}{\operatorname{argmin}}\, D_{KL}\left(P_{\text{data}}(y|\mathbf{x}) \,||\, p(y|\mathbf{x};\theta)\right). \tag{1}$$

Advice: Recall that the empirical data distribution is defined by the specific data set:

$$P_{\text{data}}(\mathbf{x}, y) \;=\; P^{x}_{\text{data}}(\mathbf{x}) P^{y}_{\text{data}}(y|\mathbf{x})$$

where P^{x}_{data} assigns probability of $1/N$ on every $\mathbf{x}_i$, and conditional probability P^{y}_{data} assigns probability 1 to y_i given $\mathbf{x}_i$ (and 0 to other values of y).

Also, note that the KL divergence between the conditional distributions, like the one that appears in (1), mean the expected KL divergence between the conditional distributions, with the expectation taken over the conditioning variable. That is, suppose we have joint distributions $p(a, b)$ and $q(a, b)$ over random variables a and b, from which we can derive the conditional $p(b|a), q(b|a)$ and marginal $p(a), q(a)$. Then

$$D_{KL}\left(p(b|a) \,||\, q(b|a)\right) \;=\; E_{\bar{a}\sim p(a)}\left[D_{KL}\left(p(b|\bar{a}) \,||\, q(b|\bar{a})\right)\right]$$

End of problem 2

- RHS $D_{KL}\left(P_{data}(y|x) \,||\, p(y|x;\theta)\right)$

$$= \underset{\bar{x}_i \sim P_{data}(x_i)}{E}\left[D_{KL}\left(P_{data}(y|\bar{x}) \,||\, p(y|\bar{x};\theta)\right)\right]$$

$$= \underset{\bar{x}_i \sim P_{data}(x_i)}{E}\left[\sum_{j=1}^{N} P_{data}(y_j|\bar{x}_j) \cdot \log \frac{P_{data}(y_j|\bar{x}_i)}{p(y_j|x_i;\theta)}\right]$$

$$= \sum_{i=1}^{N} \frac{1}{N} \cdot \left[\underbrace{\sum_{j=1}^{N} P_{data}(y_j|\bar{x}_i) \cdot \log \frac{P_{data}(y_j|\bar{x}_i)}{p(y_j|x_i;\theta)}}\right]$$

$P_{data}(y_j|x_i) = 0$ if $j \neq i$

$P_{data}(y_j|x_i) = 1$ if $j = i$

$$= \sum_{i=1}^{N} \frac{1}{N} \cdot \left(-\log p(y_i|x_i;\theta)\right).$$

- Therefore, $\underset{\theta}{\arg\min}\, D_{KL}\left(P_{data}(y|x) \,||\, p(y|x;\theta)\right)$

$$= \underset{\theta}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(-\log p(y_i|x_i;\theta)\right)\right\}$$

$$= \underset{\theta}{\arg\max}\left\{\sum_{i=1}^{N} \log p(y_i|x_i;\theta)\right\}.$$