Maximum Likelihood Estimation and KL Divergence

Let $\hat{p}(x, y)$ denote the empirical data distribution over a space of inputs $x \in \mathcal{X}$ and outputs $y \in \mathcal{Y}$. For example, in an image recognition task, x can be an image and y can be whether the image contains a cat or not. Let $p_\theta(\, y \mid x\,)$ be a probabilistic classifier parameterized by θ, e.g., a logistic regression classifier with coefficients θ. Show that the following equivalence holds:

$$arg\ max_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x,y)}\, [\ \log p_\theta\ (\,y\ \mid x\,)\,] = arg\ min_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x)} \left[D_{KL}\big(\hat{p}(\,y \mid x\,) \big|\big| p_\theta(\,y \mid x\,)\big)\right] \quad (1)$$

Where D_{KL} denotes the KL-Divergence:

$$D_{KL}\ (p(x)\ ||\ q(x)) = \ \mathbb{E}_{x\,\sim\,p(x)}\, [\log\ p(x)\ - \log q(x)] \quad (2)$$

To help you get started consider this: We rely on the known property that if ψ is a strictly monotonically decreasing function, then the following two problems are equivalent:

$$\max_{\theta} f(\theta) = \min_{\theta} \psi\big(f(\theta)\big)$$

Expanding the KL-Divergence term on the RHS of (1), we get:

$$D_{KL}\ (\hat{p}(\,y \mid x\,)\ ||\ p_\theta(\,y \mid x\,)) = \ \mathbb{E}_{\mathrm{y}\,\sim\,\hat{p}(\,y|x\,)}\, [\log \hat{p}(\,y \mid x\,) - \log p_\theta(\,y \mid x\,)]$$

substituting the term in (1)

$$arg\ max_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x,y)}\, [\ \log p_\theta\ (\,y\ \mid x\,)\,] = arg\ min_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x)} \left[\mathbb{E}_{\mathrm{y}\,\sim\,\hat{p}(\,y|x\,)}\, [\log \hat{p}(\,y \mid x\,) - \log p_\theta(\,y \mid x\,)]\right]$$

$$= arg\ min_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x,\mathrm{y})}[\log \hat{p}(\,y \mid x\,) - \log p_\theta(\,y \mid x\,)]$$

Since $\log \hat{p}(\,y \mid x\,)$ does not depend on θ, we can remove it from the optimization,

$$= arg\ min_{\theta \in \Theta}\ - \mathbb{E}_{\hat{p}(x,\mathrm{y})}[\log p_\theta(\,y \mid x\,)]$$

Since $-\mathbb{E}_{\hat{p}(x,\mathrm{y})}[\log p_\theta(\,y \mid x\,)]$ is strictly monotonically decreasing, we can use the hint, and have:

$$= arg\ max_{\theta \in \Theta}\ \mathbb{E}_{\hat{p}(x,\mathrm{y})}[\log p_\theta(\,y \mid x\,)]$$

Completing the proof.

Conditional Independence and Parameterization

Consider a collection of n discrete random variables $\{X_i\}_{i=1}^{n}$, where the number of outcomes for X_i is $|\operatorname{val}(X_i)| = k_i$.

(a) [1.50 points (Written)] Without any conditional independence assumptions, what is the total number of independent parameters needed to describe the joint distribution over $(X_1, \dots, X_n)$?

(b) [1.50 points (Written)] Under what independence assumptions is it possible to represent the joint distribution $(X_1, \dots, X_n)$ with $\sum_{i=1}^{n}(k_i - 1)$ total number of independent parameters?

(c) [2 points (Written)]

Let $1, 2, \dots, n$ denote the topological sort for a Bayesian network for the random variables $X_1, X_2, \dots, X_n$. Let m be a positive integer in $\{1,2,\dots,n-1\}$. Suppose, for every $i > m$, the random variable X_i is conditionally independent of all ancestors given the previous m ancestors in the topological ordering.

Mathematically, we impose the independence assumptions

$$p(X_i \mid X_{i-1}, X_{i-2}, \dots, X_2, X_1) = p(X_i \mid X_{i-1}, X_{i-2}, \dots, X_{i-m}) \qquad (7)$$

for $i > m$. For $i \leq m$, we impose no conditional independence of X_i with respect to its ancestors. Derive the total number of independent parameters to specify the joint distribution over $(X_1, \dots, X_n)$. You can express the answer using summation and product symbols.

(a) Without conditional independence assumptions, we need to specify the probability of every possible combination of outcomes for all variables. However, since the probabilities must sum to 1, the last probability can be determined from the others, so, the total number of independent parameters required is:

$$\left(\prod_{i=1}^{n} k_i\right) - 1$$

(b) When all variables are mutually independent, the joint distribution can be factored into the product of individual marginal distributions:

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2)\dots p(x_n)$$

For each variable X_i, we need $k_i - 1$ parameters to specify its marginal distribution since the last probability for each variable will be determined by the sum to 1 constraint. So, the total number of parameters required under the mutually independent assumption is:

$$\sum_{i=1}^{n}(k_i - 1)$$

(c) Given the conditional independence assumptions

$$p(X_i \mid X_{i-1}, X_{i-2}, \dots, X_2, X_1) = p(X_i \mid X_{i-1}, X_{i-2}, \dots, X_{i-m}) \qquad (7)$$

for $i > m$, the total number of independent parameters required to specific the joint distribution over $(X_1, \dots, X_n)$ is:

$$\left(\prod_{j=1}^{i} k_j\right) - 1$$

For $i \leq m$, we only need to specify the conditional distribution of X_i with respect to its previous m ancestors. The parameters required is:

$$\left(\prod_{j=i-m}^{i} k_j\right) - 1$$

Hence the total number of independent parameters required to specify the joint distribution over $(X_1, \ldots, X_n)$ is:

$$\sum_{i=1}^{m}\left(\prod_{j=1}^{i} k_j - 1\right) + \sum_{i=m+1}^{n}\left(\prod_{j=i-m}^{i} k_j - 1\right)$$

Monte Carlo Integration

A latent variable generative model specifies a joint probability distribution $p(x, z)$ between a set of observed variables x ∈ $\mathcal{X}$ and a set of latent variables $z \in \mathcal{Z}$. From the definition of conditional probability, we can express the joint distribution as $p(x, z) = p(z)p(x|z)$. Here, $p(z)$ is referred to as the prior distribution over z and $p(x|z)$ is the likelihood of the observed data given the latent variables. One natural objective for learning a latent variable model is to maximize the marginal likelihood of the observed data given by:

$$p(x) = \int_z p(x, z)dz \tag{12}$$

When z is high dimensional, evaluation of the marginal likelihood is computationally intractable even if we can tractably evaluate the prior and the conditional likelihood for any given x and z**.** We can however use Monte Carlo to estimate the above integral. To do so, we sample k samples from the prior $p(z)$ and our estimate is given as:

$$A\left(z^{(1)}, \ldots, z^{(k)}\right) = \frac{1}{k}\sum_{i=1}^{k} p\left(x \mid z^{(i)} \right), \text{where } z^{(i)} \sim p(z) \tag{13}$$

(a) [1 point (Written)] An estimator $\hat{\theta}$ is an unbiased estimator of θ if and only if $\mathbb{E}\left[\hat{\theta}\right] = \theta$. Show that A is an unbiased estimator of $p(x)$.

(b) [1 point (Written)] Is $log\ A$ an unbiased estimator of $log\ p(x)$? Prove why or why not.
Hint: The proof is short, estimator of $p(x)$ using the definition of an unbiased estimator and Jensen's Inequality

(a) From (13), we have

$$A\left(z^{(1)}, \ldots, z^{(k)}\right) = \frac{1}{k}\sum_{i=1}^{k} p\left(x \mid z^{(i)} \right), \text{where } z^{(i)} \sim p(z)$$

Taking the expectation:

$$\mathbb{E}[A] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} p\left(x \mid z^{(i)} \right)\right]$$

By the linearity of expectation:

$$\mathbb{E}[A] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[p\left(x \mid z^{(i)} \right)\right]$$

Since each $z^{(i)}$ is sampled from $p(z)$, for any i:

$$\mathbb{E}\left[p\left(x \mid z^{(i)} \right)\right] = \int_z \mathrm{p(x|z)p(z)dz} = \int_z \mathrm{p(x, z)dz} = \mathrm{p(x)}$$

Therefore,

$$\mathbb{E}[A] = \frac{1}{k}\sum_{i=1}^{k} \mathrm{p(x)} = p(x) \tag{14}$$

(b) No. $log\ A$ is not an unbiased estimator of $log\ p(x)$. Jensen' inequality for a concave function states that $\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$, since logarithm is a concave function: $\mathbb{E}[\log \mathrm{A}] \leq \log(\mathbb{E}[\mathrm{A}])$

From (14) above, we know $\mathbb{E}[A] = \mathrm{p(x)}$, so:

$$\mathbb{E}[\log \mathrm{A}] \leq \log(\mathrm{p(x)})$$

$Log\ A$ is not an unbiased estimator of $log\ p(x)$; it is tends to underestimate $log\ p(x)$.

Programming Assignment

(a) [1 point (Written)]

Suppose we wish to find an efficient bit representation for the 50257 tokens. That is, every token is represented as $(a_1, a_2, \ldots, a_n), where\ a_i \in \{0,1\}, \forall i = 1,2, \ldots, n$. What is the minimal n that we can use?

(b) [1 point (Written)]

If the number of possible tokens increases from 50257 to 60000, what is the increase in the number of parameters? Give an exact number and explain your answer.

Hint: The number of parameters in the GPT-2 module in Fig. 1 does not change.

(a) The minimal value of n is $n = \lceil log_2(50257) \rceil = 16$

(b) The number of parameters increase $= \ (60000 - 50257) \ * \ 768 \ = \ 7482624$, since we add a new 768-dimensional embedding vector for each token.