Modern Computational Statistics lec02 Notes

Part II

This is based on PKU course: Modern Computational Statistics. Thanks to Prof. Cheng Zhang, this is a very interesting course.

This lecture will introduce some of entropy maximization. See Stanford lecture notes for STAT 311 for more details.

1 Entropy Maximization

1.1 Introduction

The maximum entropy principle is, given some constraints about a distribution P, we consider all probability distributions satisfying said constraints. To encode our prior information while being as objective or agnostic as possible, we should choose the distribution P satisfying the constraints to maximize the Shannon entropy.

1.2 Shannon Entropy

Definition 1 *Let μ be a base measure on X and assume P has density p with respect to μ. Then the Shannon entropy of P is*

$$H(P) = -\int p(x) log p(x) d\mu(x)$$

1.3 The Maximum Entropy Problem

Here we will consider two ways of discription of this problem. One comes from Stanford STAT 311 and the other comes from PKU's mcs course.

Example 1 *Consider a linear constraints on our distributions. Then given a function $\phi : X \to R^d$, it should satisfy:$E_P[\phi(X)] = \alpha$ To maximize entropy:*

$$maximize \quad H(P)$$

$$subject \quad to \quad E_p[\phi(X)] = \alpha$$

Using the Shannon entropy, we get:

$$maximize \quad -\int p(x) log p(x) d\mu(x)$$

$$subject \quad to \int p(x)\phi_i(x) d\mu(x) = \alpha_i$$

$$p(x) \geq 0 \quad for \quad x \in X, \int p(x) d\mu(x) = 1$$

Another example is about discrete probability distribution.

Example 2 *The discrete probability distribution entropy maximization problem can be expressed as:*

$$minimie \quad f_0(x) = \sum_{i=1}^{n} x_i \log x_i$$
$$subject \quad to \quad -x_i \leq 0, i = 1, ..., n$$
$$\sum_{i=1}^{n} x_i = 1$$

1.4 Solutions to this Problem

First see problem one, the linear problem.

Recall the exponential family and log-partition-function.

Definition 2 *The exponential family associated with the function ϕ and base measure μ is defined as the set of distributions with desities p_θ with respect to μ, where*

$$p_\theta(x) = exp(< \theta, \phi(x) > -A(\theta))$$
$$A(\theta) := log \int_X exp(< \theta, \phi(x) >) d\mu(x)$$

A is the log-partition-function.

Recall the exponential family properties, we notice that the log-partition function is infinitely differentialble on its open domain, and moreover, A is convex.

Here we give the following theorem:

Theorem 1 *For $\theta \in R^d$, let P_θ have density*

$$p_\theta(x) = exp(< \theta, \phi(x) > -A(\theta)), \quad A(\theta) = log \int (< \theta, \phi(x) >) d\mu(x)$$

with respect to the measure μ. If $E_{P_\theta}[\phi(X)] = \alpha$, then P_θ maximizes $H(P)$ over P_α^{lin}; moreover, the distribution P_θ is unique.

A heuristic derivation of this theorem is very easy and interesting, let's prove it:

Proof 1 *Introducing Lagrange multipliers $\lambda(x) \geq 0$ for $p(x) \geq 0$, and θ_0 for the normalization constraint $P(X) = 1$, and θ_i for $E_P[\phi_i(X)] = \alpha_i$, we get:*

$$L(p, \theta, \theta_0, \lambda) = \int p(x) \log p(x) d\mu(x) + \sum_{i=1}^{d} \theta_i (\alpha_i - \int p(x) \phi_i(x) d\mu(x)) + \theta_0 (\int p(x) d\mu(x) - 1) - \int \lambda(x) p(x) d\mu(x)$$

Calculate the integral, and take derivatives:

$$\frac{\partial}{\partial p(x)} L(p, \theta, \theta_0, \lambda) = 1 + \log p(x) - \sum_{i=1}^{d} \theta_i \phi_i(x) + \theta_0 - \lambda(x) = 1 + \log p(x) - < \theta, \phi(x) > + \theta_0 - \lambda(x)$$

set this equal to zero:

$$p(x) = exp(< \theta, \phi(x) > -1 - \theta_0 - \lambda(x))$$

With this setting, we always have $p(x) > 0$*, and the constraint* $p(x) \geq 0$ *is not neccessary, so* $\lambda(x) = 0$*. Taking* $\theta_0 = -1 + A(\theta)$*:*

$$p_\theta(x) = exp(< \theta, \phi(x) > -A(\theta))$$

Q.E.D.

Now let's see the solution to the second problem. We will use lagrangian duality to solve this problem. Before we begin, let's recall some properties.

Consider a generalized lagrange function:

$$L(x, \alpha, \beta) = f(x) + \sum_{i=0}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)$$

To maximize this function, i.e, $\max_{\alpha,\beta} L$, set

$$\theta_p(x) = \max_{\alpha,\beta} L(x, \alpha, \beta)$$

Consider two conditions for $\theta_P(x)$, i.e, x satisfies or not satisfies the primal constraints. If x satisfies, then $f(x) = \theta_P(x)$, if not, then $+\infty = \theta_P(x)$. So to minimize $\theta_p(x)$ is just to minimize $f(x)$, and it equals to $\min_x \max_{\alpha,\beta} L(x, \alpha, \beta)$. The primal problem can be set as $\min_x \theta_P(x)$.

Now recall the dual problem. The primal problem:

$$\min_x \theta_P(x) = \min_x \max_{\alpha,\beta} L(x, \alpha, \beta)$$

Define a new function:

$$\theta_D(\alpha, \beta) = \min_x L(x, \alpha, \beta)$$

The dual problem is:

$$\max_{\alpha,\beta} \theta_D = \max_{\alpha,\beta} \min_x L(x, \alpha, \beta)$$

Now let's prove the theorem.

Proof 2 *Take the lagrangian of the primal problem:*

$$L(x, \lambda, v) = \sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} \lambda_i x_i + v(\sum_{i=1}^{n} x_i - 1)$$

The objection is to minimize ,so we take the derivative and setting it to zero.

$$\log x_i + 1 - \lambda_i + v = 0 \Rightarrow x_i = exp(\lambda_i - v - 1)$$

substitute x_i with $exp(\lambda_i - v - 1)$, and we obtain:

$$g(\lambda.v) = -\sum_{i=1}^{n} exp(\lambda_i - v - 1) - v$$

The dual problem is:

$$maximize \quad g(\lambda, v) = -exp(-v - 1)\sum_{i=1}^{n} exp(\lambda_i) - v, \lambda \geq 0$$

Take the derivative subject to v, we can obtain the dual optimal:

$$\lambda_i^* = 0, v^* = -1 + \log n$$

Put this value into the primal problem, and we get the distribution:

$$x_i^* = \frac{1}{n}$$

This is uniform distribution. So we know that the discrete probability distribution that has maximum entropy is the uniform distribution.

2 KKT Conditions

Suppose that functions $f_0, f_1, .., f_m, h_1, ..., h_p$ are all differentiable, $x^*$and (λ^*, v^*) are primal and dual optimal points with zero duality gap. Since x^* minimize $L(x, \lambda*, v*)$,the gradient vanishes at x^*:

$$\nabla f_0(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{p} v_I^* \nabla h_j(x^*) = 0$$

The Karush-Kuhn-Tucker(KKT) conditions can be expressed as:

$$\begin{aligned} f_i(x^*) &\geq 0, i = 1, ..., m \\ h_j(x^*) &= 0, j = 1, ..., p \\ \lambda_i^* &\geq 0, i = 1, ..., m \\ \lambda_I^* f_i(x^*) &= 0, i = 1, ...m \end{aligned}$$

Under KKT conditions, the x is primal feasible and minimizes $L(x, \lambda, v)$. This can help us solve he primal problem very quickly.