EE/Stats 376A: Information Theory **Winter 2017**

Lecture 16 — March 7

Lecturer: David Tse *Scribe: M. Tefagh, A. Thomas, N. Grimwood, F. Farnia*

16.1 Introduction

Starting this lecture, we will talk about the maximum entropy principle and its applications to machine learning and statistics. In the previous lecture, we proved that the Gaussian distribution maximizes differential entropy subject to the second moment constraint. In this lecture, we generalize that simple example to more general settings. A short read of Section 12.1 and 12.2 from the Cover & Thomas textbook will be helpful to understand our discussion on the maximum entropy principle.

16.2 Information Theory for Machine Learning

We have seen the fundamental role of entropy and mutual information in solving two important problems: compression and communication over noisy channels. Here we use these tools from information theory for different applications in machine learning and statistics.

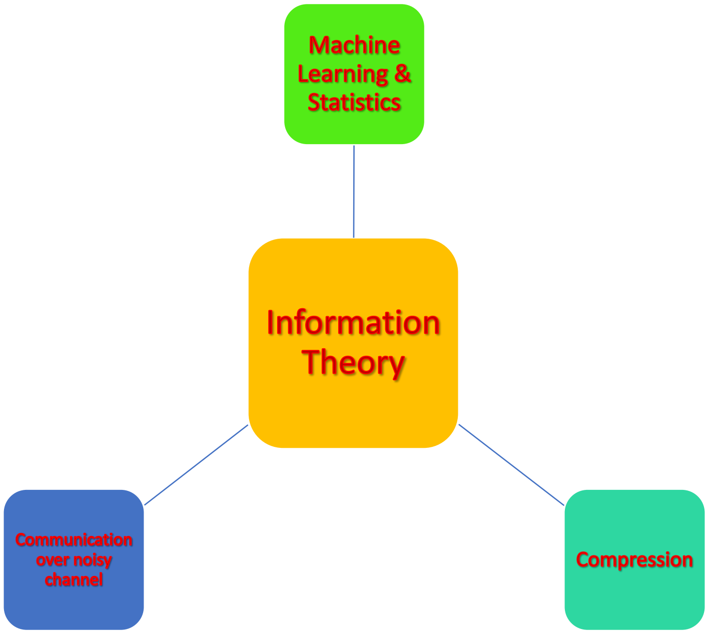
Machine Learning & Statistics
Information Theory
Communication over noisy channel
Compression

Figure 16.1: Applications of Information Theory reviewed in this course

16.3 Maximum Entropy Principle

The *principle of maximum entropy* states that given some constraints on the underlying distribution we should base our decision on the probability distribution with the largest entropy.

We can use this principle for both supervised and unsupervised learning tasks. In this lecture our main focus is on the unsupervised case, where given some data for X we want to learn the probability distribution which best represents the observed data.

However, we usually access to only very few samples, which is not enough to completely infer a very high-dimensional probability distribution of X. Hence, a good idea is to broaden our focus to a set Γ including some candidate distributions. Then, the principle of maximum entropy states that the desired probability distribution p^* is the solution to the following optimization problem:

$$\max_{p_X \in \Gamma} H(X)$$

16.3.1 Exponential Family distributions

Suppose that Γ is described by some moment constraints, the value of which can come from the observed data, i.e.

$$\Gamma = \Big\{ p_X : \ \mathbb{E}\big[r_i(X)\big] = \alpha_i, \ i = 1, \ldots, I \Big\} \tag{16.1}$$

Let p^* be the maximum entropy distribution over Γ. To find the structure of p^*, consider q as an arbitrary probability distribution in Γ. We have

$$\begin{aligned} 0 &\le D(q \,\|\, p^*) \\ &= \mathbb{E}\Big[\log \frac{q(X)}{p^*(X)}\Big] \quad \text{where } \ X \sim q \\ &= -H(X) + \mathbb{E}\Big[\log \frac{1}{p^*(X)}\Big] \end{aligned} \tag{16.2}$$

Note that, in case $\mathbb{E}\big[\log(1/p^*(X))\big] = \sum_x q(x)\log(1/p^*(x)) = H(X^*)$ where $X^* \sim p^*$ holds for all $q \in \Gamma$, the above calculation implies the optimality of p^*.

Example 1. We have seen that if the second order moment of X is set to be 1, in which case $\Gamma = \big\{p_X \mid \mathbb{E}\big[X^2\big] = 1\big\}$, then $X^* \sim \mathcal{N}(0,1)$ gives the entropy-maximizing distribution. To prove this claim, we used the same technique as described above in (16.2) and hence chose $p^*(x) \propto e^{\lambda x^2}$.

We can apply the same technique for a general Γ in (16.1) by choosing:

$$p^*(x) \propto \exp\left(\sum_i \lambda_i r_i(x)\right)$$

Then we can find a coefficient λ_0 to say

$$\Rightarrow\ p^*(x) = \exp\left(\lambda_0 + \sum_{i=1}^{n} \lambda_i r_i(x)\right)$$

$$\Rightarrow\ \log\frac{1}{p^*(x)} = -\lambda_0 - \sum_{i=1}^{n} \lambda_i r_i(x)$$

$$\Rightarrow\ \mathbb{E}\Big[\log\frac{1}{p^*(X)}\Big] = -\lambda_0 - \sum_{i=1}^{n} \lambda_i \mathbb{E}\big[r_i(X)\big] = H(X^*) \quad \forall q \in \Gamma \text{ where}\, X \sim q,\ X^* \sim p^*.$$

Example 2. Let

$$\Gamma = \big\{p_X :\ X \geq 0,\ \mathbb{E}\big[X\big] = \mu\big\}.$$

Thus, the maximum entropy distribution should have the form $f(x) \propto e^{\lambda x}$. This implies that X has an exponential distribution,

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}.$$

Theorem 1. If for coefficients $\lambda_0, \lambda_1, \ldots, \lambda_I$ the following probability distribution

$$p^*(x) = e^{\lambda_0 + \sum_{i=1}^{n} \lambda_i r_i(x)}$$

satisfies the constraints given by Γ (16.1) ($\mathbb{E}\big[r_i(X)\big] = \alpha_i$ for $\ i = 1, \ldots, I$), then $p^* \in \Gamma$ is the maximum entropy distribution, i.e.

$$p^* = \underset{p \in \Gamma}{\operatorname{argmax}}\, H(X).$$

In general, probability distributions of the form $p^*(x) = e^{\lambda_0 + \sum_{i=1}^{n} \lambda_i r_i(x)}$ are called *exponential family* distributions, and our maximum entropy discussion is an important reason to use exponential family distributions for various applications.

16.3.2 Lagrange Duality

Note that the Lagrangian for the discussed entropy maximization problem is:

$$\mathcal{L} = \left(\sum_x p(x) \log\frac{1}{p(x)}\right) - \sum_i \lambda_i \left(\sum_x p(x) r_i(x) - \alpha_i\right) - \lambda_0 \left(\sum_x p(x) - 1\right)$$

Therefore, λ_i's are indeed the Lagrange multipliers corresponding to each constraint in the dual optimization problem. Under some mild conditions excluding some degenerate cases, one can prove that strong convex-duality holds, and hence the converse of Theorem 1 holds as well.

16.4 Multiple Variables

Similar to the scalar case, the maximum entropy principle for a random vector $\mathbf{X} = (X_1, \ldots, X_p)$, is to select a joint distribution $p^*_{X_1,\ldots,X_p}$ maximizing

$$\max_{p_{\mathbf{X}} \in \Gamma} H(\mathbf{X}),$$

where

$$\Gamma = \Big\{ p_{\mathbf{X}} : \mathbb{E}\big[r_i(X_1, \ldots, X_p)\big] = \alpha_i,\ i = 1, \ldots, I \Big\}.$$

Example 3. Let $r_{ij}(x_i, x_j) = x_i x_j$ and $\Gamma = \{p_{\mathbf{X}} : \mathbb{E}\big[r_{i,j}(X_i, X_j)\big] = K_{i,j},\ i, j = 1, \ldots, I\}$.. By the same arguments as 16.2,

$$\begin{aligned} p^*(\mathbf{x}) &= \exp\left(\lambda_0 + \sum_{ij} \lambda_{ij} r_{ij}(x_i x_j)\right) \\ &= \exp\left(\lambda_0 + \sum_{ij} \lambda_{ij} x_i x_j\right) \\ &= \exp\left(\lambda_0 + \mathbf{x}^T \Lambda \mathbf{x}\right), \end{aligned} \tag{16.3}$$

where the ijth entry of matrix Λ is λ_{ij}. One can easily show the covariance matrix $\mathbf{K}$ for this probability distribution satisfies:

$$\Lambda = -\frac{1}{2}\mathbf{K}^{-1} \ \Rightarrow\ \mathbf{X}^* \sim \mathcal{N}(0, \mathbf{K})$$

As a result, similar to the scalar case, if we constrain all the second-order moments, then the maximum entropy distribution is a multivariate Gaussian distribution with zero-mean and covariance matrix $\mathbf{K}$.

Suppose that we deal with a high-dimensional dataset where random vector $\mathbf{X}$ includes millions of features. Then, because of high computational and statistical complexity, we cannot measure all p^2 K_{ij}'s. Instead, we restrict our attention to the second-order moments coming from a subset of interest $E \subseteq \{1, \ldots, n\}^2$. Hence, the resulted set of distributions is:

$$\Gamma = \Big\{ p_{\mathbf{X}} : \mathbb{E}\big[X_i X_j\big] = K_{ij},\ \forall (i,j) \in E \Big\}.$$

To solve the maximum entropy problem for the above Γ, we apply the same previous technique to get

$$\begin{aligned} p^*(\mathbf{x}) &= \exp\left(\lambda_0 + \sum_{(i,j)\in E} \lambda_{ij} r_{ij}(x_i x_j)\right) \\ &= \exp\left(\lambda_0 + \sum_{(i,j)\in E} \lambda_{ij} x_i x_j\right) \\ &= \exp\left(\lambda_0 + \sum_{ij} \tilde{\lambda}_{ij} x_i x_j\right) \\ &= \exp\left(\lambda_0 + \mathbf{x}^T \tilde{\Lambda} \mathbf{x}\right). \end{aligned}$$

Here

$$\tilde{\lambda}_{ij} = \begin{cases} \lambda_{ij}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise,} \end{cases}$$

and $\tilde{\Lambda} = [\tilde{\lambda}_{ij}]_{i,j=1}^{n}$. Using the same argument, we have $\tilde{\Lambda} = -\frac{1}{2}\mathbf{K}^{-1}$ where $\mathbf{K}$ denotes the covariance matrix of random vector $\mathbf{X}$. Here, we assign zero value to any entry of the inverse covariance matrix whose corresponding pair has not been included in E. Therefore,

$$\mathbf{X}^* \sim \mathcal{N}(0, \mathbf{K})$$

has a multivariate Gaussian distribution similar to the previous example. This model is known as *Gaussian Graphical Model* with E describing the edge set of that graphical model.