COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Moritz Hardt

Lecture # 18
April 14, 2008

Summary. In this lecture we study the problem of estimating a probability density function from random (unlabeled) samples distributed according to this density. This task is known as *probability modeling* or *density estimation*. We will introduce and relate two standard methods called *maximum likelihood* and *maximum entropy*.

1 Conditional Density Estimation

Although it will not be our focus for this lecture, let us briefly discuss what is known as *conditional density estimation*. Here, we are given random samples (x, y) distributed according to some unknown distribution and our goal is to estimate the conditional probability $\Pr[y|x]$. There are two approaches towards this problem:

- **Discriminative approach** in which we use tools from learning theory to compute a hypothesis that models the conditional probability distribution $\Pr[y|x]$ up to a small error.
- **Generative approach** in which we estimate the probability distribution $\Pr[x|y]$ separately for every y. For instance, y could represent the attribute "gender" and x could represent the attribute "height". In this case, we would learn the distribution of heights separately for women and men.

Recall Bayes' Rule,

$$\Pr[y|x] = \frac{\Pr[x, y]}{\Pr[x]} = \frac{\Pr[x|y]\Pr[y]}{\Pr[x]}.$$

It implies that these two approaches are in principle equivalent. The term $\Pr[x]$ can be ignored since it is constant with respect to y. The probabilities $\Pr[y]$ can be estimated easily by the marginal distributions derived from the samples. However, the two approaches do have different strengths in practice.

2 Maximum Likelihood

Suppose we are given examples $x_1, x_2 \ldots, x_m$ drawn from a probability distribution D over some discrete space[1] X. In the end, our goal is to estimate D by finding a model which fits the data, but is not too complex. As a first step, we need to be able to measure the quality of our model. This is where we introduce the notion of *maximum likelihood*.

To motivate this notion suppose D is distributed according to one out of two possible density functions q_1 and q_2. Intuitively, if we observe that $q_1(x_i)$ is typically much larger than $q_2(x_i)$, we will tend to conclude that D is distributed according to q_1.

[1] Even though what we discuss generalizes straightforwardly to the continuous setting.

In general, we consider a (possibly infinite) set of density functions $\mathcal{Q}$. For a particular $q \in \mathcal{Q}$, we call

$$\prod_{i=1}^{m} q(x_i) \tag{1}$$

the *likelihood of* $x_1, \ldots, x_m$ *under* q. Notice, if the examples x_i are independent, then this term is precisely the probability of generating the sequence $x_1, \ldots, x_m$.

Since the logarithm is strictly monotonically increasing and

$$\log \prod_{i=1}^{m} q(x_i) = \sum_{i} \log(q(x_i)), \tag{2}$$

we know that maximizing (1) is equivalent to maximizing (2) which in turn is equivalent to *minimizing*

$$\sum_{i} -\log(q(x_i)). \tag{3}$$

We think of (3) as a "loss function" that we call the *log loss of* q *on* $x_1, \ldots, x_m$. Let us also introduce the *true risk of* q as

$$\mathop{\mathrm{E}}_{x \sim D}[-\log q(x)] = -\sum_{x \in X} D(x) \log q(x). \tag{4}$$

The last term is a quantity sometimes called the *cross entropy* of D and q. It only differs by an additive constant from the *relative entropy* of D and q. Hence, it is minimized when $D = q$ (as we showed using Lagrange multipliers). Indeed,

$$\begin{aligned} -\sum_{x \in X} D(x) \log q(x) &= \sum_{x \in X} D(x) \log \frac{D(x)}{q(x)} - \sum_{x \in X} D(x) \log D(x) \\ &= \mathrm{RE}(D||q) + \mathrm{H}(D), \end{aligned}$$

where RE denotes relative entropy and H the Shannon entropy.

Example 1. Suppose we want to estimate the bias of a coin from a sequence of m coin tosses. In this case, $\mathcal{Q}$ is the set of all probability distributions supported on {HEADS, TAILS}. If we observe HEADS h times, then the likelihood of the sequence under a probability distribution q is equal to $q^h(1-q)^{m-h}$ where we identified q with the probability of HEADS. This term is maximized for $q = \frac{h}{m}$.

Example 2. Suppose a biologist wants to derive a probabilistic model of where on a given map X a particular species lives. The biologist is given (a) presence records $x_1, \ldots, x_m$ of the species according to the population distribution, and (b) environmental variables $f_1, \ldots, f_n$ describing the map with attributes such as "altitude", "average rain fall" and so forth. To model these additional variables we extend our formal setup as follows.

Formal Setup. Let us consider a large but finite set X of cardinality N. Our observation is $x_1, \ldots, x_m \sim D$ where D is some unknown distribution supported on X. Furthermore, we are given features $f_1, \ldots, f_n$ where each $f_j \colon X \to \mathbb{R}$ is a function. Our goal is to estimate D.

The first way to do this would be to maximize likelihood. To do this, we need to fix a domain $\mathcal{Q}$. One common choice is to let $\mathcal{Q}$ be the set of all density functions q of the form

$$q(x) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_{\boldsymbol{\lambda}}}, \tag{5}$$

where $\lambda_j \in \mathbb{R}$ and $Z_{\boldsymbol{\lambda}}$ is a normalization constant that depends on $\boldsymbol{\lambda}$. This family of density functions is often called the *exponential* family. We will refer to distributions of this form as *Gibbs distributions*. Now, the *maximum likelihood* is

$$\begin{aligned} \textbf{Maximize} \quad & \sum_{i=1}^{n} \log q(x_i) \\ \textbf{subject to} \quad & q \in \bar{\mathcal{Q}}, \end{aligned} \tag{6}$$

where $\bar{\mathcal{Q}}$ denotes the *closure* of $\mathcal{Q}$ (that is, $\mathcal{Q}$ together with all its limit points).

3 Maximum Entropy

Another way of approaching the above problem is to use the method of *maximum entropy*. Here, we start with the fact that we can approximate the true expectation $\mathrm{E}_D[f_j]$ of each feature f_j by its empirical average taken over the given samples, i.e.,

$$\hat{\mathrm{E}}[f_j] = \frac{1}{m} \sum_i f_j(x_i). \tag{7}$$

That is, we expect $\mathrm{E}_D[f_j] \approx \hat{\mathrm{E}}[f_j]$ for all j. This leads to the idea of finding a distribution p which satisfies the constraint

$$\mathrm{E}_p[f_j] = \hat{\mathrm{E}}[f_j] \tag{8}$$

for every j. There are typically many distributions satisfying these constraints. Among all such distributions we choose the one which minimizes its distance from the uniform distribution U in terms of relative entropy,

$$\begin{aligned} \mathrm{RE}(p||U) &= \sum_x p(x) \log \frac{p(x)}{1/N} \\ &= \log N + \sum_x p(x) \log p(x) \\ &= \log N - \mathrm{H}(p). \end{aligned}$$

Since $\log N$ is just a constant, we are looking for the distribution p which maximizes the Shannon entropy $\mathrm{H}(p)$. Once we define

$$\mathcal{P} = \{p \mid \forall j \colon \mathrm{E}_p[f_j] = \hat{\mathrm{E}}[f_j]\},$$

the *maximum entropy* can be written as

$$\begin{aligned} \textbf{Maximize} \quad & H(p) \\ \textbf{subject to} \quad & p \in \mathcal{P}. \end{aligned} \tag{9}$$

4 Duality between Maximum Entropy and Likelihood

For the setup we have chosen previously, one can use convex programming duality to argue that (6) and (9) do have unique optima that coincide.

Theorem 1. *Let q^* be a probability distribution. Then, the following are equivalent:*

1. $q^* = \arg\max_{p \in \mathcal{P}} H(p)$,
2. $q^* = \arg\max_{q \in \bar{\mathcal{Q}}} \sum_i \log q(x_i)$,
3. $q^* \in \mathcal{P} \cap \bar{\mathcal{Q}}$.

Furthermore, any of these statements uniquely determines q^*.

We will not prove this theorem. However, we quickly provide some intuition for why it is true. For this purpose, let us consider the Lagrangian relaxation of (9), i.e.,

$$\mathcal{L} = \sum_{x \in X} q(x) \log q(x) + \sum_{j=1}^{n} \lambda_j \left(\hat{\mathrm{E}}[f_j] - \sum_{x \in X} q(x) f_j(x) \right) + \gamma \left(\sum_{x \in X} q(x) - 1 \right). \tag{10}$$

The primal variables are $q(x)$ for $x \in X$, while the dual variables are λ_j and γ. Setting

$$0 = \frac{\partial \mathcal{L}}{\partial q(x)} = 1 + \log q(x) - \sum_j \lambda_j f_j(x) + \gamma$$

gives us

$$q(x) = \frac{\exp(\sum_j \lambda_j f_j(x))}{e^{\gamma+1}}. \tag{11}$$

Here, $e^{\gamma+1}$ takes the place of the normalization constant Z in (5). So, we recognize that the optimal solution to (9) must in fact define a Gibbs distribution.

On the other hand, upon substituting (11) back into (10), we get

$$\begin{aligned}
\mathcal{L} &= \sum_{x \in X} q(x) \left(\sum_{j=1}^{n} \lambda_j f_j(x) - \log Z \right) - \sum_{j=1}^{n} \lambda_j \sum_{x \in X} q(x) f_j(x) + \sum_{j=1}^{n} \lambda_j \hat{\mathrm{E}}[f_j] \\
&= -\log Z + \frac{1}{m} \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{m} f_j(x_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \lambda_j f_j(x) - \log Z \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \log q(x_i),
\end{aligned}$$

where q is as in (11). In other words, at optimality the Lagrangian simplifies to the objective function in (6) up to a constant multiple.