COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Bianca M. Dumitrascu

Lecture #19
April 15, 2014

1 Introduction to Probability Density Estimation

1.1 Summary

In this lecture we consider the framework of **Probability Density Estimation**. In our previous approaches to learning (Classification, Regression) we are given m labeled points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ according to a random distribution $\mathcal{P}$, and, disregarding this distribution, we want to predict with minimal error the label of a new point x. Such approaches belong to the class of **Discriminative Approaches**. The name comes from the fact that, from a probabilistic perspective, we attempt to find the conditional distribution $\Pr(\mathbf{y}|\mathbf{x})$ which helps us *discriminate* between different label values. An alternative is provided by **Generative Approaches** which constitutes the focus of our current discussion.

1.2 Intuition

In the generative approach setting, when presented with m labeled training samples (x_1, y_1), (x_2, y_2), …, (x_m, y_m), we assume the data points x are generated from an unknown distribution $\mathcal{P}$. We aim to model this distribution using the conditional density estimation of the quantity $\Pr(\mathbf{x}|\mathbf{y})$. To illustrate this, consider a dataset where the training samples give information about the distribution of height in a population. The training samples take the form (height, gender) where the labels y represent the gender. Given the height of a person we might want to guess the person's gender. Previously, we might have modeled this problem using a discriminative approach by looking for a threshold or for a separating hyperplane that would give us a decision rule according to which the height of interest belongs to a man or to a woman. In the new set-up, we are interested in estimating the distribution of heights separately for women and men such that we can calculate how *likely* it is for a new sample to have been generated from either of these distributions. In general, for the rest of this lecture we assume there is a probability distribution over the values x (in the example above, we assume there is a distribution of heights for women that is different from the height distribution of men), and our goal is to model this distribution for the purpose of learning and inference.

2 Maximum Likelihood

To formalize the above learning setting, consider we are given m samples $x_1, x_2, \ldots, x_m$ drawn from a probability distribution $\mathcal{P}$ over a finite domain $\mathcal{X}$ (generalizable to the continuous setting, the assumption over the domain is merely for making calculations easy). The goal is to estimate $\mathcal{P}$ by finding a model that while not too complex, is able to fit the data. To this end, let $\mathcal{Q}$ be a family of possibly infinitely many density functions q. It is among these functions that we will search for the best model to explain our data.

Definition. Let $x_1, x_2, ...x_m$ be m points sampled *iid.* from a distribution $\mathcal{P}$ and let $q \in \mathcal{Q}$ be a density function. The *likelihood* of $x_1, x_2, \ldots, x_m$ under q is the quantity:

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

Notice that this quantity is exactly the probability of generating the m independent samples given that the underlying model is given by the probability density q.

Example (Coin Toss) Consider flipping a coin with probability q of landing *heads.* Consider random variables x which take the value 1 if the coin lands heads, and 0 otherwise. Then given m coin flips, let the sequence $x_1, x_2, \ldots, x_m$ record the sequence of tosses and notice that the number of heads is nothing else but $h = \sum_{i=1}^{m} x_i$. With this notation the *likelihood* of the data under q is equal to $q^h(1-q)^{m-h}$.

Naturally, we are interested in the probability function q that performs closest to the real probability distribution $\mathcal{P}$ and which makes the likelihood quantity *more likely.* As we want to maximize the probability that the sequence $x_1, x_2, \ldots, x_m$ is generated from q we notice that:

$$\max_{q \in \mathcal{Q}} \prod_i q(x_i) = \max \log \prod_i q(x_i) = \max \sum_i \log(q(x_i)) = \min \frac{1}{m} \sum_i [-\log q(x_i)] \tag{2}$$

where $-\log q(x)$ is the log loss of q on x and where $\frac{1}{m}\sum_i[-\log q(x_i)]$ is the empirical risk, or the average log loss. We use this empirical risk as a proxy for the real expectation. The true value of the loss or true risk is given by the quantity $E_{x\sim\mathcal{P}}[-\log q(x)]$ which can be iteratively written as:

$$\begin{aligned}
&E_{x\sim\mathcal{P}}[-\log q(x)] \\
&= -\sum_{x\in\mathcal{X}} \mathcal{P}(x) \log q(x) \\
&= -\sum_{x\in\mathcal{X}} \mathcal{P}(x) \log q(x) + \sum_{x\in\mathcal{X}} \mathcal{P}(x) \log \mathcal{P}(x) - \sum_{x\in\mathcal{X}} \mathcal{P}(x) \log \mathcal{P}(x) \\
&= -\sum_{x\in\mathcal{X}} \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{q(x)} - \sum_{x\in\mathcal{X}} \mathcal{P}(x) \log \mathcal{P}(x) \\
&= RE(\mathcal{P}||q) + H(\mathcal{P})
\end{aligned}$$

where RE denotes the relative entropy and H is the Shannon entropy.
Equation (3) justifies the intuition that minimizing (2) would give us the probability density function closest to the true distribution, where the chosen metric is relative entropy. Going back to the **coin toss** example, estimating the bias of the coin from the sequence of tosses is simply finding the term that maximizes the log likelihood, namely $q = \frac{h}{m}$.

3 Maximum Entropy Formulation

Consider now the related problem of modeling the distribution of interest given multiple constraints, or features of the data. As before, we are given the samples $x_1, x_2, \ldots, x_m$ generated from some unknown distribution $\mathcal{P}$ over a finite set $\mathcal{X}$ of cardinality N, with $N \gg m$. In addition, for each such x we are given a set of n functions $f_1, f_2, \ldots, f_n$ where $f_j : \mathcal{X} \rightarrow \mathbb{R}$. We call these functions features and we think of them as constraints over the distribution $\mathcal{P}$. The goal is the same as before: estimating $\mathcal{P}$ subject to the constraints induced by the features.

A natural way to approach the problem is to use the additional information encoded into the feature. To this end, notice that we can approximate the expectations $E_{\mathcal{P}}[f_j]$ over the true distribution for features f_j using the empirical average taken over the given samples:

$$E_{\mathcal{P}}[f_j] \approx \widehat{E}[f_j] = \frac{1}{m}\sum_{j=1}^{m} f_j(x_i) \tag{3}$$

The problem can be recast as the problem of finding q such that for all i from 1 to n, $E_q[f_j] = \widehat{E}[f_j]$. As there could be many probability density functions satisfying this constraint, we will pick the one that minimizes $\mathrm{RE}(q||U)$ where U is the uniform distribution over X:

$$\begin{aligned} q = argmin_q RE(q||U) &= \sum_{x \in X} q(x) \ln \frac{q(x)}{1/N} \\ &= \ln N + \sum_{x \in X} q(x) \ln(q(x)) \\ &= \ln N - H(q) \end{aligned} \tag{4}$$

With these in mind let $\mathcal{P}$ be the set of probability densities constrained by their features:

$$\mathcal{P} = \{q \mid E_q[f_j] = \widehat{E}[f_j], \forall j\}. \tag{5}$$

$$\tag{6}$$

The probability of interest is therefore the solution to the following formulation, called *maximum entropy*:

$$\begin{aligned} &\textbf{maximize } H(q) \\ &\textbf{subject to } q \in \mathcal{P} \end{aligned} \tag{7}$$

Notice however that we can also think about the problem using the maximum likelihood framework developed in the previous section. To solve this in practice we need to make the search tractable by restricting the set of density functions $\mathcal{Q}$ over which we maximize (or minimize if we talk about the log loss). One solution to this is to consider the set of probability distributions q:

$$q(x) \propto \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right) \tag{8}$$

where $\lambda_j \in \mathbb{R}$. This family of distribution functions is an example of what is often referred to as an *exponential family*. For this particular family we will use the name *Gibbs distribution*. The maximum likelihood problem becomes:

$$\begin{aligned} &\textbf{maximize } \sum_{i=1}^{m} \log q(x_i) \\ &\textbf{subject to } q \in \bar{\mathcal{Q}} \end{aligned} \tag{9}$$

where $\bar{\mathcal{Q}}$ is the closure of $\mathcal{Q}$.

The following theorem brings together the two approaches stating that they are, in fact, equivalent.

Theorem 1. Duality between Maximum Entropy and Maximum Likelihood.
Let q^* be a probability distribution. Then, the following identities are equivalent:
1. $q^* = \arg\ max_{q \in \mathcal{P}} H(q)$,
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 properties uniquely determine q^*

Without proving this result, let us consider the Lagrangian form of the likelihood function, obtaining the following identity in which $q(x)$ for $x \in \mathcal{X}$ are primal variables and λ_j and γ are dual variables or Lagrange multipliers:

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

$$\tag{11}$$

Setting $\frac{\partial L}{\partial q(x)}$ to zero, that is

$$0 = 1 + \log q(x) + \sum_j \lambda_j f_j(x) + \gamma, \tag{12}$$

we obtain a closed form expression for $q(x)$ that resembles the form of the optimal solution given by the Gibbs distribution, namely:

$$q(x) = \frac{\exp\big(\sum_{j=1}^{n} \lambda_j f_j(x)\big)}{e^{\gamma+1}}. \tag{13}$$

By letting $Z = e^{\gamma+1}$ and plugging it back into L, we notice that the expression simplifies to be the log likelihood of q:

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

which suggests that the dual problem is to maximize the log likelihood of q as a function of the Lagrange multipliers λ.