COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire | Lecture 19
Scribe: Richard M. Price II | April 16, 2013

1 Density Estimation

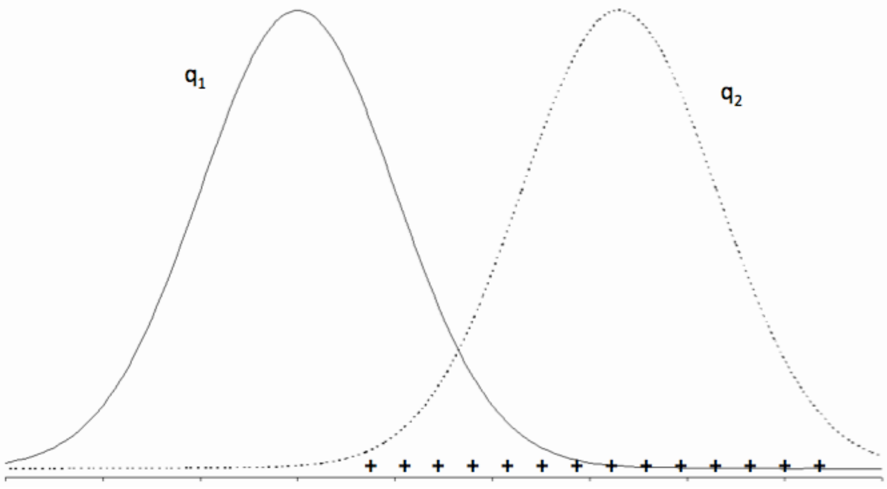
We are interested in modeling an unknown distribution D for which we have samples. This is known as "probability modeling" or "density estimation". Stating the problem in mathematical terms, given $(x, y) \sim D$, we wish to find $\Pr[y|x]$.

For example, consider applying this to medical diagnosis, where y is the underlying disease, and x represents symptoms, test results, etc. We hope to find the probability of a patient having disease y given the external observations represented by x.
We can also model $\Pr[x|y]$, which gives the probability of certain symptoms or external observations given the presence of disease y. If we can solve for one of these probabilities, we can then use Bayes' Rule to find the other.

$$\Pr[y|x] = \frac{\Pr[x|y] \cdot \Pr[y]}{\Pr[x]} \tag{1}$$

$\Pr[y]$ is usually easy to estimate, and the denominator serves as a normalization constant that can largely be ignored.
One can also model the joint distribution, finding $\Pr[y|x]$=$\Pr[x, y]/\Pr[x]$. There are various approaches to solving this problem, and a natural first step is just to do density estimation.

Given $x_1, \ldots, x_m$, chosen i.i.d from distribution P, our unknown distribution. Note that we do not have direct access to P, and want to build an estimate of P given our samples. Just as we usually attempt to select the best hypothesis in a hypothesis space, here we are trying to select the best distribution in a class of distributions, or possible models. What should be our criteria?
Let Q be a class of distributions. How do we pick the best $q \in Q$ given certain data? Consider $q_1, q_2 \in Q$ as drawn below, with a number of samples represented by the + signs.



If q_1 is the true distribution, then it is highly unlikely that we would have selected this particular sample, which seems more typical for q_2, which is clearly a better fit. This gives us a simple, yet powerful criterion for density estimation, namely looking at the probability of drawing the given sample under a particular distribution. You then choose the distribution that makes this sample more likely. This can be expressed as

$$\text{Likelihood of } q = \Pr_q[x_1, \ldots, x_m] = \prod_{i=1}^{m} q(x_i)$$

To find the *maximum likelihood* we choose the q that maximizes the above expression for likelihood.

Consider the following example of a biased coin, with all distributions having bias between 0 and 1, so $Q = [0, 1]$. Note the following definitions:

$$x = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

$$q(x_i) = \begin{cases} q & \text{if } x_i = 1 \\ 1 - q & \text{if } x_i = 0 \end{cases}$$

- Given $x_1, \ldots, x_m$
- $h = \sum_{i=1}^{m} x_i$, or the number of 1's

$$\begin{aligned} \text{Likelihood of } q &= \prod_{i}^{m} q(x_i) \\ &= q^h (1 - q)^{m-h} \end{aligned}$$

Differentiating and setting to 0, we get that $q = h/m$ is the distribution that satisfies the principle of maximum likelihood.

Going back to the goal of maximizing likelihood, dealing with products can get ugly. We can modify the expression for maximum likelihood as follows and still get the same solution.

$$\begin{aligned} \max_{q \in Q} \prod_i q(x_i) &\leftrightarrow \max \log(\prod_i q(x_i)) \\ &\leftrightarrow \max \sum_i \log(q(x_i)) \\ &\leftrightarrow \min \frac{1}{m} \sum_{i=1}^{m} (-\log(q(x_i)) \end{aligned}$$

We started out with maximum likelihood, and ended up with a function of the model and the data. We are really taking the sum of the log loss function. We can thus view maximum likelihood as minimizing log loss.

The way we have it written here, we are minimizing an empirical average, or empirical risk. We can think of empirical risk as a proxy for true risk. We will thus try to minimize empirical risk, selecting a distribution q, with the choice being entirely unconstrained, that does so as a first exercise. The choice q usually does have constraints. How do we choose this q, given access to p?

$$\begin{aligned}(\text{Risk of } q) &= E_{x\sim p}[-\log q(x)] \\ &= -\sum_x p(x)\cdot\log q(x) \quad \text{(at times called cross entropy)}\end{aligned}$$

We are trying to minimize the above expression such that $\sum_x q(x) = 1$. As we saw in the homework, the expression is minimized when $q = p$. This is good because if we can minimize the actual risk, not having constraints on q, then we will get the right answer. Further,

$$-\sum_x p(x)\cdot\ln q(x) = \sum_x p(x)\cdot\ln\left(\frac{p(x)}{q(x)}\right) - \sum_x p(x)\cdot\ln p(x)$$

The first term on the right is our old friend relative entropy, and the second term, including the minus sign, is the entropy of p, a nonnegative measure of how spread out a distribution is. So by minimizing the risk, we are finding a distribution q that is as close as possible to p.

2 Butterfly Hunting example

We can now look at a specific problem, namely that of modeling the population distribution of a particular butterfly species. Where can we find it? We can go into the field to look for it, plotting our results on a map of the United States. We are usually interested in rare species, so the number of sightings would likely be small. We can also study the temperature, altitude, amount of rainfall, etc. that the butterfly thrives in. Let us also assume that we are given the altitude, temperature, and average rainfall levels at every point in the United States. Given all this, our goal is to model the butterfly's distribution, essentially by figuring out the kind of habitat in which it is likely to live.

Let us call D the true distribution of the species population. Our goal is to estimate D, assuming that our observations all come from this distribution. This is like a density estimation problem, but with additional known features.

We are given a map, or set of possible locations, X that has been divided into a large but finite grid of points. We also have a set of butterfly sightings $x_1, \ldots, x_n$ i.i.d $\sim D$, and a set of features $f_1, \ldots, f_n$; $f_j : X \to \mathbb{R}$ providing altitude, average temperature, etc for every point on the map. Our goal is to estimate D.

2.1 First Idea: Maximum Entropy

Start with what we know how to estimate, namely the expected value of a random variable.

$$\mathbb{E}_D[f_j] \approx \hat{\mathbb{E}}[f_j] = \frac{1}{m}\sum_{i=1}^{m} f_j(x_i)$$

We expect this to give us a rough property that we can directly measure from our training set. How to find a distribution q that is an estimate of D with the quality that
$\mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j] \quad$ for all $j = 1, \ldots, n$?

Many distributions satisfy this constraint. Let us then naively find the distribution q closest to the uniform distribution.

$$\text{Minimize: } RE(q \parallel unif) = \sum_x q(x) \cdot \ln\left(\frac{q(x)}{1/N}\right) = \ln N + \sum_x q(x) \cdot \ln q(x)$$
$$= \ln N - H(q)$$

This is the same as maximizing the entropy H. So to summarize this first strategy, we are choosing q that maximizes the entropy $H(q)$ according to the constraints
$P = \{q : \mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j] \quad \forall j\}$

2.2 Second Idea

We want to use maximum likelihood. To do so, we must pick a parametric form, described by a set of real-valued parameters, to estimate unknown distribution D, then choosing the distribution in that parametric form that maximizes the likelihood of the data. In other words, we are going to guess a form with parameters that we think is natural for describing D, then finding the parameters that serve our purpose.
q is our estimate of D. We need a formula to express the probability of q over x. We can work with the given features, taking a linear combination thereof.

$$q(x) = \sum_{j=1}^{n} \lambda_j f_j(x)$$

This isn't optimal, as the feature values could be negative and not add up to 1. We can thus do the following

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

Note the normalization factor Z_λ. We can now choose the features λ by maximizing the likelihood of the training examples we have observed. More explicitly, Q is the set of all distributions q of the form given immediately above. These are often referred to as Gibbs distributions. To pick the best q, we choose the q of this form that maximizes the log-likelihood of the data:

$$\arg\max_{q \in \bar{Q}} \sum_{i=1}^{m} \ln q(x_i)$$

This maximum might not exist unless we take the "closure" of Q. The closure of a set is when you take all the points in a set as well as points that can be the results of a constructed sequence that converges to them. This is expressed by the bar above the Q in the above expression.

3 Theorem and Motivations

Note that the second idea leads to the same solution as the first idea. This leads us to a theorem.

Theorem 3.1 *The following statements are equivalent:*

- $q^* = \arg\max_{q \in P} H(q)$
- $q^* = \arg\max_{q \in \bar{Q}} \sum_{i=1}^{m} \ln q(x_i)$
- $q^* \in P \cap \bar{Q}$

Furthermore, any one of the above uniquely determines q^*.

3.1 Motivating argument for equivalence of bullet points 1 and 2

We will begin with the first bullet and see that it is equivalent to the second bullet by duality, using the method of Lagrange multipliers. We are trying to maximize the expression

$$-\sum_x q(x) \cdot \ln q(x)$$

such that for all j, $\sum_x q(x) \cdot f_j(x) = \hat{\mathbb{E}}[f_j]$ and $\sum_x q(x) = 1$.

This can be formulated as a constrained optimization problem, for which we can use Lagrange multipliers. Therefore, we wish to find

$$\min \sum_x q(x) \ln q(x)$$

such that for all j, $-\sum_x q(x) f_j(x) = \hat{\mathbb{E}}[f_j]$ and $\sum_x q(x) = 1$. We will now form the Lagrangian, with the λ_j terms and γ as Lagrange multipliers.

$$L = \sum_x q(x) \cdot \ln q(x) + \sum_{j=1}^{n} \lambda_j (\hat{\mathbb{E}}[f_j] - \sum_x q(x) f_j(x)) + \gamma(\sum_x q(x) - 1)$$

We will now optimize the Lagrangian over the $q(x)$ terms, or primal variables, setting the derivative to 0.

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

We can now solve for $q(x)$.

$$\begin{aligned} q(x) &= \exp(\sum_j \lambda_j f_j(x) - \gamma - 1) \\ &= \frac{\exp(\sum_j \lambda_j f_j(x))}{Z} \end{aligned}$$

Note that Z is standing in for $e^{\gamma+1}$ which, in order to meet the necessary constraints, serves as a normalization factor. Plugging this expression $q \in Q$ into the Lagrangian formulation selectively, you get the following

$$\begin{aligned} L &= \sum_x q(x) \cdot (\sum_j \lambda_j f_j(x) - \ln Z) + \sum_j \lambda_j \hat{\mathbb{E}}[f_j] - \sum_j \lambda_j \sum_x q(x) f_j(x) \\ &= -\ln Z + \frac{1}{m} \sum_j \lambda_j \sum_i f_j(x_i) \\ &= \frac{1}{m} \sum_i (\sum_j \lambda_j f_j(x_i) - \ln Z) \end{aligned}$$

Note that the inner summation term is equivalent to $\ln q(x_i)$, meaning that the Lagrangian expression simplifies to the log-likelihood of the training data over q. Furthermore, we know that the solution of the Lagrangian occurs at a saddle point where it is minimized in the primal variables and maximized at the dual variables. In other words, the dual problem is equivalent to the 2nd bullet.