COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Qiang Huang

Lecture #17
April 8, 2003

1 Review of Bayes' Rule

In the previous lecture, we talked about modeling and estimating probability distributions. In this generative approach, we deal with probabilities in the form $Pr(x|y)$, where y is the class and x is the input data. Bayes' Rule is used to relate the $Pr(x|y)$ to $Pr(y|x)$ which is estimated by discriminative techniques. Bayes' Rule states that:

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

Here is a numerical example. Assume the probability of a certain disease is $Pr[D] = 0.01$. The probability of test positive given that a person is infected with the disease is $Pr[T|D] = 0.95$ and the probability of test positive given the person is not infected with the disease is $Pr[T|\bar{D}] = 0.05$. Then the probability of test positive is

$$\begin{aligned} Pr[T] &= Pr[T \wedge D] + Pr[T \wedge \bar{D}] \\ &= Pr[T|D]Pr[D] + Pr[T|\bar{D}]Pr[\bar{D}] \\ &= 0.95 \times 0.01 + 0.05 \times 0.99 \\ &= 0.059 \end{aligned}$$

Bayes' Rule is used to calculate the probability of being infected with the disease given that the test result is positive:

$$Pr[D|T] = \frac{Pr[T|D]Pr[D]}{Pr[T]} = \frac{0.95 \times 0.01}{0.059} \approx 0.16$$

2 Model a Probability Distribution Using Maximum Likelihood

Given m samples $x_1, x_2, \cdots, x_m$, where x_i is distributed according to some distribution D. We are trying to figure out how we can model this distribution D. Imagine that we have a class of distributions to model D; the question is which one to choose. If $D = q$, and we assume the data is independent to each other, we have

$$Pr[x_1, x_2, \cdots, x_m] = q(x_1)q(x_2)\cdots q(x_m) = \prod_{i=1}^{m} q(x_i) \tag{2}$$

which is the likelihood of the samples from distribution q. The principle here is to choose the distribution with the maximum likelihood, i.e.,

$$\max_q \prod_i q(x_i) \equiv \max_q \ln \prod_i q(x_i) \equiv \max_q \sum_i \ln q(x_i) \equiv \min_q \sum_i -\ln q(x_i) \tag{3}$$

where $-\ln q(x_i)$ is called log loss and the maximum likelihood principle is equivalent to minimizing the log loss. The expected log loss should be minimized under the true distribution D, i.e., $E_D[-\ln q(x)] = -\sum_x D(x) \ln q(x)$ is minimized when $q = D$. Therefore, the principle of minimizing the log loss is to choose

$$\min_q \; -\sum_x D(x) \ln q(x) \quad \text{s.t.} \; \sum_x q(x) = 1 \tag{4}$$

One way to see this is to use Lagrange multipliers:

$$L = -\sum_x D(x) \ln q(x) + \lambda \left(\sum_x q(x) - 1 \right). \tag{5}$$

If we optimize this over q and λ by setting derivatives equal to zero, then we will get that $q = D$.

More generally, if we estimate D by q, then the expected log loss will be:

$$\begin{aligned} -\sum_x D(x) \ln q(x) &= \sum_x D(x) \ln \frac{D(x)}{q(x)} - \sum_x D(x) \ln D(x) & (6) \\ &= RE(D \parallel q) + H(D). & (7) \end{aligned}$$

In the above equation, $H(D)$ is the entropy of the true distribution D, thus does not depend on q, while $RE(D \parallel q)$ is the distance between the two distributions D and q. Therefore, the expected log loss only depends on the distance between distribution q and the true distribution D. When $q = D$, we have the expected log loss $E_D[-\ln q(x)] = -\sum_x D(x) \ln q(x)$ minimized.

3 An Example of Maximum Likelihood Modeling

Random variable X is Bernoulli distributed with parameter p, i.e., $X = 1$ with probability p and $X = 0$ with probability $1-p$, where p can take any value from $[0, 1]$. Given m examples $x_1, x_2, \cdots, x_m \in X$ and we want to estimate p. Let $h = \sum_i x_i$. Intuitively, we claim $\frac{h}{m}$ is the answer according to the maximum likelihood.

Now we calculate the likelihood for probability q: $L(q) = \prod_i q(x_i) = q^h(1-q)^{m-h}$. Let $\frac{\partial L(q)}{\partial q} = 0$ to maximize $L(q)$ over q, and we find $q = \frac{h}{m}$ is the answer.

4 A More Complex Example - Habitat Modeling Problem

In this habitat modeling problem, we are interested in estimating distribution of the places where a species of animals live. Take butterfly, for example, we are given information about their habitat places, and we can observe and collect information about test features at every possible location, where test features include altitude, average rainfall, average temperature, etc. What to estimate is the true distribution of the habitat.

Using mathematical symbols to represent this model, we are given:

- locations $x \in X$, $|X| = N$

- n features $f_1, f_2, \cdots, f_n$, where $f_i : X \to R$ (mapping locations of the habitat to real numbers)

- m samples $x_1, x_2, \cdots, x_m$. We assume the samples are iid from distribution D.

We are trying to estimate this D. Certain species prefer specific habitat of favorable environments, therefore, features usually determine the location distribution. We know how to estimate the expected value of the features:

$$E_D[f_j] \approx \hat{E}[f_j] = \frac{1}{m} \sum_i f_j(x_i) \tag{8}$$

where $\hat{E}[f_j] = \hat{E}_j$ is the empirical expected value of the features. So one idea is to find a distribution p such that

$$E_p[f_j] = \hat{E}[f_j], \quad \forall j \tag{9}$$

The problem is that a lot of distributions p can satisfy the above requirements. Among all these p's, in the absence of any other information, it is natural to choose the one that is closest to the uniform distribution, i.e., minimizing the relative entropy

$$RE(p \parallel \text{unif}) = \sum_x p(x) \ln \frac{p(x)}{1/N} = \ln N + \sum_x p(x) \ln p(x) = \ln N - H(p). \tag{10}$$

Thus, minimizing $RE(P \parallel \text{unif})$ is equivalent to maximizing $H(P)$, and we call this strategy the "Maximum Entropy Approach", in which we first find the set

$$\mathcal{P} = \{p : \ E_p[f_j] = \hat{E}[f_j] \ \ \forall j\} \tag{11}$$

and then choose

$$q^* = \arg \max_{p \in \mathcal{P}} H(p) \tag{12}$$

The second approach is to model the distribution $q(x)$ by a linear combination of the features, generating a "Gibbs" form of distribution:

$$q(x) = \frac{\exp\left(\sum_j \lambda_j f_j(x)\right)}{Z_\lambda} \tag{13}$$

where $\lambda_j \in R$ and Z_λ is a normalization factor. We can then derive a "Maximum Likelihood Approach", in which we choose the Gibbs distribution of maximum likelihood. Let

$$\mathcal{Q} = \{q \text{ of the Gibbs form}\}. \tag{14}$$

and let $\bar{\mathcal{Q}}$ be the closure of $\mathcal{Q}$. We then choose

$$q^* = \arg \max_{q \in \bar{\mathcal{Q}}} \sum_i \ln q(x_i) \tag{15}$$

An amazing result is that these two problems have identical solutions and moreover the intersection $\mathcal{P} \cap \bar{\mathcal{Q}}$ contains a single element which is the unique solution to both problems.

Theorem 1 *The following are equivalent:*

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

Moreover, q^* is uniquely determined by any one of these conditions.

The theorem states that the solution q^* is uniquely determined by any of the above equations and the "Maximum Entropy Approach" is equivalent to the "Maximum Likelihood Approach". We did not prove this theorem, but we can motivate the result using Lagrange multipliers.

For the first approach, the Lagrange multiplier is

$$L = \sum_x p(x) \ln p(x) + \sum_j \lambda_j (\hat{E}_j - \sum_x p(x) f_j(x)) + \gamma (\sum_x p(x) - 1) \tag{16}$$

$$\frac{\partial L}{\partial p(x)} = 1 + \ln p(x) - \sum_j \lambda_j f_j(x) + \gamma = 0 \tag{17}$$

$$\Rightarrow \quad p(x) = \frac{\exp(\sum_j \lambda_j f_j(x))}{e^{1+\gamma}} = \frac{\exp(\sum_j \lambda_j f_j(x))}{Z} \tag{18}$$

$$\text{where } Z = e^{1+\gamma} \text{ is the normalization factor to make } p(x) \text{ adds up to 1.} \tag{19}$$

We then conclude that the solution of the first approach has the same form as the second approach, i.e., generating a Gibbs distribution. Moreover, if we plug in this form of p into the Lagrangian L, we get the likelihood of the data which is what we need to maximize over the λ_j's. Thus, the two approaches are equivalent.

Next we introduce an iterative algorithm to solve λ:
In the first step, we set the initial value as λ_1.
At round $t+1$, we have

$$\lambda_{t+1} = \lambda_t + \alpha_t \tag{20}$$

The algorithm should make λ_t converge to the λ in $q^*(x)$. We assume $f_j(x) \geq 0$ and $\sum_j f_j(x) = 1$. (If $f_j(x)$ is sometimes negative, we can replace $f_j(x)$ by $f_j(x) + c$ for some constant c. Because of normalization, this does not change the corresponding Gibbs distribution. If $\sum_j f_j(x) < 1$, we can add a new feature $f_0(x) = 1 - \sum_j f_j(x)$. Since a linear combination over all the features including f_0 is the same as one over just the original features, this again does not change the problem or the Gibbs distributions that can be represented.)

Define

$$g_\lambda(x) = \sum_j \lambda_j f_j(x) \tag{21}$$

$$q_\lambda(x) = \frac{e^{g_\lambda(x)}}{Z_\lambda} \tag{22}$$

And the loss function is

$$L(\lambda) = -\sum_i \ln q_\lambda(x_i) \tag{23}$$

Then look at the difference $L(\lambda_{t+1}) - L(\lambda_t)$. Next class we will continue the algorithm to derive an approximation process for λ.