# COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Matt Hoffman

Lecture #18
April 18, 2006

## 1 Modeling Probability Distributions

Until now, we have mainly considered the problem of, given a set of examples of the form $(x, y) \sim D$, finding a function that maps a new example $x$ to a correct label $y$. We will now consider the problem of estimating the actual distribution $P$ from which a set of examples $X \sim P$ was drawn. Solving this problem has applications to areas such as language modeling, and can also be used to perform classification.

As an aside, two basic approaches to using conditional density estimation (the problem of finding $D$ given a series of examples of the form $(x, y) \sim D$) are:

- *The Discriminative Approach:* here we estimate $Pr[y|x]$ directly. The best choice for $y$ can be selected by simply choosing the most likely $y$ given $x$, but the probabilities themselves may also be useful if we wish to model the uncertainty in a process (e.g. for weather prediction, medical diagnoses...).

- *The Generative Approach:* here we instead estimate $Pr[x|y]$ for each $y$. We can then use Bayes's Rule ($Pr[y|x] = \frac{Pr[x|y]Pr[y]}{Pr[x]}$) to find $Pr[y|x]$.

## 2 Maximum Likelihood

We are given a set of examples $x_1, x_2, ..., x_m \sim P$ and a class $Q$ of probability distributions. Our goal is to find the $q \in Q$ that best estimates $P$. A natural approach is to choose the $q$ under which we would be most likely to see the examples $x_1, x_2, ..., x_m$. Under each $q$,

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

since the examples are independent. The product on the right is referred to as the "**likelihood of the data under** $q$," and it is this quantity we would like to maximize.

More formally:

$$\textbf{goal: } \text{choose } q \text{ to get } max \prod_i q(x_i) \equiv max \sum_i log(q(x_i)) \equiv min \sum_i \underbrace{-log(q(x_i))}_{\text{log loss}}$$

where $\equiv$ is used to imply that the same $q$ will maximize both expressions.

Why use the log loss instead of the regular loss? Well,

$$E_{x\sim P}[\text{log loss under q}] = \sum_x P(x)(-log(q(x))) = \underbrace{-\sum_x P(x)(log(q(x)))}_{\text{sometimes called cross entropy}}$$

which, as shown in the homework, is minimized when $q = P$. Furthermore:

$$-\sum_x P(x)log(q(x)) = -\sum_x P(x)log(q(x)) + \sum_x P(x)log(P(x)) - \sum_x P(x)log(P(x))$$

$$= \sum_x P(x) log\left(\frac{P(x)}{q(x)}\right) - \sum_x P(x) log(P(x)) = RE(P||q) - \underbrace{H(P)}_{\text{entropy of P}}$$

So, $E[log\ loss]$ is a function of the relative entropy of $P$ and $q$ and of the inherent variability of $P$ (as measured by its entropy $H(P)$).

An example:

Say

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

and $Q = [0,1]$. We observe examples $x_1, x_2, ..., x_m$. Let $h = \sum_i x_i$.

$$\text{The likelihood under } q = \prod_i q(x_i) = q^h(1-q)^{m-h} \text{ is maximized when } q = \frac{h}{m}$$

# 3 Application: Modeling Animal/Plant Habitats

Say we are given a small (around 20-100) set of points on an effectively two-dimensional map at which biologists have spotted individuals belonging to some species. Additionally, we are given information such as altitude, rainfall, average temperature, etc. at all points on the map. We assume there is some distribution generating the points we have been given which corresponds to the population distribution of the species being studied. Given only these positive examples, we will try to arrive at an approximation to the actual distribution of the population. More formally:

- Our locations $x$ come from a prediscretized, finite space of possible points $X$. $|X| < \infty$, but potentially quite large.
- We observe examples $x_1, x_2, ..., x_m \in X$, which we assume are i.i.d. according to the underlying distribution $D$. Our goal is to estimate $D$.
- Our features are the environmental variables at each point. So essentially, feature $f_j : X \to \mathbb{R}$

# 4 Maximum Entropy

The seemingly simplest option would be to estimate $D$ by the empirical distribution, i.e., the probability of each example point is $\frac{1}{m}$, and the probability at any other point is 0. Unsurprisingly, this estimate will perform terribly since it is concentrated on just a tiny fraction of points in a potentially huge space. However, it is worth noting that by Chernoff bounds,

$$\text{The expectation with respect to the empirical distribution } \hat{E}[f_j] = \frac{1}{m}\sum_i f_j(x_i) \approx E_D[f_j]$$

so our goal should be to find a distribution $P$ s.t. $\forall j\, E_p[f_j] = \hat{E}[f_j]$. Many choices of $P$, including the empirical distribution, satisfy this requirement. Since arguably the most "natural" distribution in the absence of any assumptions is the uniform distribution, let's see what happens when we attempt to find a distribution $P$ s.t. $\forall j\, E_p[f_j] = \hat{E}[f_j]$ and $P$ is as close as possible to uniform (i.e. $RE(P||\text{uniform})$ is minimized).

$$RE(P||\text{unif.}) = \sum_x P(x) log \frac{P(x)}{\frac{1}{N}} = \sum_x P(x) log P(x) + \sum_x P(x) log N$$

$$= \underbrace{\sum_x P(x) log P(x)}_{-H(P)} + log N$$

So the idea is to define $\mathcal{P} = \{p : E_p[f_j] = \hat{E}[f_j] \text{ and } \sum_x P(x) = 1\}$, and try to find the distribution $p \in \mathcal{P}$ of *maximum entropy*, i.e. $\max_{p \in \mathcal{P}} H(p)$

## 5 Maximum Likelihood Revisited

Consider distributions that are linear combinations of features: $q(x) = \sum_j \lambda_j f_j(x)$. (Note that $f_j$ need not be in $[0,1]$.) To turn this into a proper probability distribution, we need to ensure that it be positive and normalized to the range $[0,1]$. We do this by making the distribution look like so:

$$q(x) = \frac{\text{to make positive: } \exp(\sum_j \lambda_j f_j(x))}{\text{to normalize: } Z}$$

Probability distributions of this form are called **Gibbs Distributions**. Now the problem is to compute the unconstrained $\lambda_j$s with maximum likelihood: $\max \prod_i q_\lambda(x_i)$

$$Q = \{q : q \text{ is a Gibbs distribution}\}, \text{ so choose } \max_{q \in \bar{Q}} \sum_i log(q(x_i))$$

$\bar{Q}$ represents the closure of $Q$, the precise meaning of which is unimportant except insofar as it guarantees that a maximum always exists.

## 6 Duality theorem

As it turns out, maximum likelihood and maximum entropy are convex duals, and $\mathcal{P} \cap \bar{Q}$ contains only a single point, which will miraculously solve both problems.
Theorem: The following are equivalent:

- $q^* = \text{argmax}_{p \in \mathcal{P}} H(p)$ (which solves maximum entropy)
- $q^* = \text{argmax}_{q \in \bar{Q}} \sum_i log(q(x_i))$ (which solves maximum likelihood)
- $q^* \in \mathcal{P} \cap \bar{Q}$

And furthermore, any one of these properties *uniquely* defines $q^*$.

Not quite a proof: We can solve the maximum entropy problem using Lagrange multipliers:

$$\mathcal{L} = \sum_x q(x) \log(q(x)) + \sum_j \lambda_j [\hat{E}[f_j] - \sum_x q(x) f_j(x)] + \gamma(\sum_x q(x) - 1)$$

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

$$q(x) = \exp\left(\sum_j \lambda_j f_j(x) - \gamma - 1\right) = \frac{\exp(\sum_j \lambda_j f_j(x))}{Z = e^{\gamma+1}}$$

Now we plug this $q(x)$ selectively into $\mathcal{L}$.

$$\mathcal{L} = \sum_x q(x)[\sum_j \lambda_j f_j(x) - \log Z] - \sum_j \lambda_j \sum_x q(x) f_j(x) + \sum_j \lambda_j \hat{E}[f_j] + \overbrace{\gamma(\sum_x q(x) f_j(x))}^{=0}$$

$$= -\log Z + \frac{1}{m} \sum_j \lambda_j \sum_i f_j(x_i) = \frac{1}{m} \sum_i \underbrace{[\sum_j \lambda_j f_j(x_i) - \log Z]}_{=\log(q_\lambda(x_i))}$$

Thus, the solution of the maximum entropy problem occurs when $q$ is a Gibbs distribution maximizing the (log) likelihood. In other words, the maximum entropy and maximum likelihood problems are duals of each other.

Of course, this may look like a proof, but there are a bunch of cases that haven't been dealt with.