COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Guanchun (Arvid) Wang

Lecture #19
April 16, 2008

1 Recap

Last time we started talking about density estimation. The problem is established as following:

Given: a space $\mathcal{X}$, where $|\mathcal{X}| < \infty$
examples $x_1, x_2, \cdots, x_m \in \mathcal{X},\ x_i \sim \mathcal{D}$
a set of features: $f_1, f_2, \cdots, f_n,\ f_j : \mathcal{X} \to \mathbb{R}$

Goal: Estimate density distribution $\mathcal{D}$.

The solution using the *Maximum Likelihood* approach is to select a distribution from

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

Note that these are called Gibbs Distributions. Among all distributions of this form, we select the one of maximum likelihood:

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

Here $\bar{Q}$ means the closure of Q, denoted in this way for technical reasons.

The solution using the *Maximum Entropy* approach is to select from

$$P = \left\{ p : E_p[f_j] = \hat{E}[f_j], \forall j \right\}$$

the distribution of maximum entropy:

$$\max_{p \in P} H(p).$$

The Duality Theorem that we motivated (but didn't prove completely) last time states that the following are equivalent and any of the three uniquely defines q^*:

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

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

(3) $q^* \in P \cap \bar{Q}$ (roughly equivalent to KKT conditions).

2 How to find q^*

Now we need to develop some computational algorithm so as to find the solution q^*. From some observations, we can determine that (2) will be the most useful because it is an unconstrained optimization problem; technically we should be able to find the solution using calculus, while (1) is a constrained optimization problem and will be more complicated. We will use (3) to prove convergence later. (2) is equivalent to

$$\min_{q \in \bar{Q}} -\frac{1}{m} \sum_{i=1}^{m} \ln q_{\boldsymbol{\lambda}}(x_i)$$

where $q_{\boldsymbol{\lambda}}(x) = \frac{e^{g_{\boldsymbol{\lambda}}}}{Z_{\boldsymbol{\lambda}}}$ and $g_{\boldsymbol{\lambda}}(x) = \sum_{j=1}^{n} \lambda_j f_j(x)$. So we want to find $\boldsymbol{\lambda}$ to minimize

$$L(\boldsymbol{\lambda}) = -\frac{1}{m} \sum_{i=1}^{m} \ln q_{\boldsymbol{\lambda}}(x_i). \tag{1}$$

Note that Eq. 1 is called the empirical log loss function and is a convex function of $\boldsymbol{\lambda}$.

3 Algorithm

Even though the optimization is unconstrained and convex, it is still too hard to get the solution analytically (by taking derivatives and setting them to zero). Therefore, we will do so numerically. The roadmap is to get a sequence of $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ so that

$$\lim_{t \to \infty} (\boldsymbol{\lambda}_t) = \min_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}).$$

So our algorithm will look like below:

1. Choose $\boldsymbol{\lambda}_1$ arbitrarily (e.g. set to zero)
2. for $t = 1, 2, \cdots$ compute $\boldsymbol{\lambda}_{t+1}$ from $\boldsymbol{\lambda}_t$.

The algorithm we present is called "Iterative Scaling". We will first scale and translate features so that

$$f_j : \mathcal{X} \to [0, 1].$$

Then replace f_j by f_j/n so that:

$$\sum_{j=1}^{n} f_j(x) \leq 1.$$

Last we will add a feature $f_0 = 1 - \sum_{j=1}^{n} f_j$, so that we have the nice equality that

$$\sum_{j=0}^{n} f_j(x) = 1, \text{ for all } x \in \mathcal{X}.$$

This addition of f_0 isn't necessary but will make the math work better later.

Now we return to the problem how to compute $\boldsymbol{\lambda}_{t+1}$ from $\boldsymbol{\lambda}_t$. Recall our goal is to minimize the difference $\Delta L = L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t)$. So we will approximate this difference and minimize it.

4 Derivation of the Approximation

We let $\boldsymbol{\lambda}' = \boldsymbol{\lambda}_{t+1}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$. Note that$\lambda'_j$ is λ_j plus some small adjustment, which we can formalize as:

$$\lambda'_j = \lambda_j + \alpha_j.$$

So we have

$$\begin{aligned}
\Delta L &= L(\boldsymbol{\lambda}') - L(\boldsymbol{\lambda}) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\ln\left(\frac{e^{g_{\boldsymbol{\lambda}}(x_i)}}{Z_{\boldsymbol{\lambda}}}\right) - \ln\left(\frac{e^{g_{\boldsymbol{\lambda}'}(x_i)}}{Z_{\boldsymbol{\lambda}'}}\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[(g_{\boldsymbol{\lambda}}(x_i) - g_{\boldsymbol{\lambda}'}(x_i)) + \ln\left(\frac{Z_{\boldsymbol{\lambda}'}}{Z_{\boldsymbol{\lambda}}}\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}(g_{\boldsymbol{\lambda}}(x_i) - g_{\boldsymbol{\lambda}'}(x_i)) + \ln\left(\frac{Z_{\boldsymbol{\lambda}'}}{Z_{\boldsymbol{\lambda}}}\right). \qquad (2)
\end{aligned}$$

Recall that

$$g_{\boldsymbol{\lambda}}(x) = \sum_{j=1}^{n}\lambda_j f_j(x).$$

Therefore,

$$\begin{aligned}
(g_{\boldsymbol{\lambda}}(x_i) - g_{\boldsymbol{\lambda}'}(x_i)) &= \sum_{j=1}^{n}(\lambda_j f_j(x_i) - \lambda'_j f_j(x_i) \\
&= -\sum_{j=1}^{n}\alpha_j f_j(x_i).
\end{aligned}$$

So the first term of Eq. 2 becomes

$$\begin{aligned}
-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_j f_j(x_i) &= -\frac{1}{m}\sum_{j=1}^{n}\sum_{i=1}^{m}\alpha_j f_j(x_i) \text{ (interchange the summation order)} \\
&= -\frac{1}{m}\sum_{j=1}^{n}\alpha_j\sum_{i=1}^{m}f_j(x_i) \\
&= -\sum_{j=1}^{n}\alpha_j\hat{E}[f_j].
\end{aligned}$$

The second term that we need to deal with is

$$\begin{aligned}
\frac{Z_{\boldsymbol{\lambda}'}}{Z_{\boldsymbol{\lambda}}} &= \frac{\sum_{x\in\mathcal{X}} \exp\left(\sum_{j=1}^{n} \lambda'_j f_j(x)\right)}{Z_{\boldsymbol{\lambda}}} \\
&= \frac{\sum_{x\in\mathcal{X}} \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x) + \sum_{j=1}^{n} \alpha_j f_j(x)\right)}{Z_{\boldsymbol{\lambda}}} \\
&= \sum_{x\in\mathcal{X}} q_{\boldsymbol{\lambda}}(x) \exp\left(\sum_{j=1}^{n} \alpha_j f_j(x)\right) \\
&\leq \sum_{x\in\mathcal{X}} q_{\boldsymbol{\lambda}}(x) \sum_{j=1}^{n} f_j(x) \exp(\alpha_j) \text{ (Jensen's Inequality and since exp() is convex)} \\
&= \sum_{j=1}^{n} e^{\alpha_j} \sum_{x\in\mathcal{X}} q_{\boldsymbol{\lambda}}(x) f_j(x) \\
&= \sum_{j=1}^{n} e^{\alpha_j} E_{q_{\boldsymbol{\lambda}}}[f_j].
\end{aligned}$$

Now we can put the whole thing together and we have

$$\begin{aligned}
\Delta L &\leq -\sum_{j=1}^{n} \alpha_j \hat{E}[f_j] + \sum_{j=1}^{n} e^{\alpha_j} E_{q_{\boldsymbol{\lambda}}}[f_j] \\
&= -\sum_{j=1}^{n} \alpha_j \hat{E}_j + \ln\left(\sum_{j=1}^{n} e^{\alpha_j} E_j\right)
\end{aligned}$$

where we denote $\hat{E}_j = \hat{E}[f_j]$ and $E_j = E_{q_{\boldsymbol{\lambda}}}[f_j]$.

Now we have bounded the difference and can optimize it by taking the partial derivative with respect to α_j, which yields

$$\frac{\partial}{\partial \alpha_j} = -\hat{E}_j + \frac{E_j e^{\alpha_j}}{\sum_{j=1}^{n} E_j e^{\alpha_j}}. \tag{3}$$

Apparently solving for α_j by setting Eq. 3 to zero would be a lot easier if not for the denominator in the second term. But we also notice that if α'_j is a solution, then $\alpha_j = \alpha'_j + C$ is also a solution, for any constant C. So we can choose C so that

$$\sum_{j=1}^{n} E_j e^{\alpha_j} = 1.$$

Therefore, setting Eq. 3 to zero gives

$$\alpha_j = \ln\left(\frac{\hat{E}_j}{E_j}\right),$$

and we have finally marched to a method for approximately minimizing ΔL.

5 Iterative Scaling Algorithm

Now our algorithm will look like below:

choose $\boldsymbol{\lambda}_1 = 0$
for $t = 1, 2, \cdots$ until convergence
for all j: $\lambda_{t+1,j} = \lambda_{t,j} + \ln\left(\frac{\hat{E}[f_j]}{E_{q_{\lambda_t}}[f_j]}\right)$.

We can also interpret the distribution over the samples instead of λ's, if we let $p_t = q_{\lambda_t}$. Then we can write

$$p_{t+1}(x) \propto p_t(x) \prod_j \left(\frac{\hat{E}[f_j]}{E_{p_t}[f_j]}\right)^{f_j(x)}. \tag{4}$$

This alternative formulation makes intuitive sense because it has the effect of adjusting the distribution weight. In the end, we want to reach a point at which $E_{p_t}[f_j] = \hat{E}[f_j]$. For example, if $\hat{E}[f_j] > E_{p_t}[f_j]$, this means that we have underestimated the expected value of feature j over the samples and would therefore like to increase the distribution weight of those samples with high values for j. As in Eq. 4, the quotient of the expectations will be > 1, so the weight adjustment component corresponding to j will be proportional to the value of f_j for each sample.

6 Proof of Convergence

Now let's prove the convergence of the two probability distributions, i.e. $p_t \to q^*$. As a side note, we should be aware that it is not sufficient to show that non-negative L is strictly decreasing, because it can converge to some positive value. Here we will construct an auxiliary function A mapping probability distributions over $\mathcal{X}$ to real numbers, that has the following three properties:

(1) $L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t) \leq A(p_t) \leq 0$.

(2) A is continuous.

(3) $A(p) = 0 \Rightarrow E_p[f_j] = \hat{E}[f_j]$, $\forall j$.

If there exists such an auxiliary function, then we are done with the proof. Why? First, we know that $L \geq 0$ and never increasing (by property(1)), which implies

$$L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t) \to 0.$$

In addition, by property (1), $A(p_t)$ is squeezed between $L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t)$ and 0, and this implies

$$A(p_t) \to 0.$$

Suppose that $p_t \to p$. Why does this imply p is optimal? By property (2), since A is continuous,

$$A(p) = \lim_{t \to \infty} A(p_t) = 0$$

And by property (3), this implies that $p \in P$. On the other hand, since each $p_t \in Q$ is a Gibbs distribution, and $p_t \to p$, we have that

$$p \in \bar{Q}.$$

Therefore,

$$p \in P, p \in \bar{Q} \Rightarrow p \in P \cap \bar{Q} \Rightarrow p = q^*$$

(last equality by Duality Theorem).

This argument assumes that the p_t's have a limit, a fact which we need a little bit of analysis or topology to prove. Although we brushed over this in class, for those who are interested, here is how this can be proved. Suppose the sequence of p_t's does not converge to q^*. Then there must exist a neighborhood R around q^* such that an infinite number of p_t's lie outside of R. The p_t's lie in the space of all probability distributions over the finite set X. This is a compact space. Therefore, the infinite subset of p_t's outside of R must have a subsequence which converges to some point p (this is a property of compactness). By the same argument given above (slightly modified), p must be equal to q^*, a contradiction since all of the points are outside of the neighborhood R around q^*. Therefore, the p_t's converge to q^*.

Next we need to find such an auxiliary function, which upper-bounds ΔL. We can plug $\alpha_j = \ln\left(\frac{\hat{E}_j}{E_j}\right)$ back into Eq. 3 and we will have

$$\begin{aligned}
\Delta L &= L(\boldsymbol{\lambda}') - L(\boldsymbol{\lambda}) \\
&= -\sum_{j=1}^{n} \hat{E}[f_j] \ln\left(\frac{\hat{E}[f_j]}{E_{q_{\boldsymbol{\lambda}}}[f_j]}\right) + \ln\left(\underbrace{\sum_{j=1}^{n} e^{\alpha_j} E_j}_{=1}\right) \\
&= -\sum_{j=1}^{n} \hat{E}[f_j] \ln\left(\frac{\hat{E}[f_j]}{E_p[f_j]}\right) \quad (p = q_{\boldsymbol{\lambda}}) \\
&= -RE(\hat{E}[f_j] \mid\mid E_p[f_j]) = A(p).
\end{aligned}$$

Note that $\sum_{j=1}^{n} \hat{E}[f_j] = \hat{E}[\sum_{j=1}^{n} f_j] = 1$ and $\sum_{j=1}^{n} E_p[f_j] = 1$, i.e. both are valid probability distributions over the feature set, hence conforming to the definition of relative entropy. And because $RE \geq 0$ and when $RE = 0 \Rightarrow \hat{E}[f_j] = E_p[f_j]$, we can check that $A(p)$ satisfies the properties of being an auxiliary function.

7 General Comments

1. The above proof was written for the case that we have a distribution over examples; however it is common to have labelled data pairs (x, y) and the goal of estimating the conditional probability $\Pr[y|x]$. It turns out that one can apply similar ideas in this case, essentially trying to maximize the entropy of $Y|X$ given constraints derived from data. This approach is called **Logistic Regression**. Therefore, Logistic Regression is just a special case of *Maximum Entropy*.

2. If the true probability distribution is in the class of distributions that you are searching over, maximum likelihood will eventually converge to the true probability; however, ML can behave badly if the true distribution is not in that class and ill-defined. For example, consider distributions over $\{0, 1\}$ which are defined by the bias or probability of 1. Say the true distribution is given by $p = 0.98$, which we estimate using

distributions in $\{0.01, 1\}$. So intuitively $q = 1$ should be the better estimate. However ML will return 0.01, because the expected log loss is

$$-E[\ln q] = 0.98 \ln q - 0.02 \ln(1 - q),$$

and if $q = 1$, then the second term is ∞. Therefore we see an important caveat of ML due to the bad behavior of the log function at the end points.

8 Preview

Next lecture, we will talk about density estimation in an online setting. For example, imagine you are betting on horses at the track. You want to estimate the probability of each horse wining and translate these estimates into bets corresponding to the probability distribution over horses. Before each race, you would combine the probability estimates of the experts into a single aggregated distribution. One horse will win the race and then you'll move on and repeat the expert advice pooling for the next race. So with the probability estimates from a panel of experts, you want to perform as well as the best expert. Now the loss function will be different.

The problem can be formalized as below

for $t = 1, 2, \cdots, T$

each expert i chooses a distribution $p_{t,i}$ over$\mathcal{X}$

master combines into q_t

observe $x_t \in \mathcal{X}$

loss $= -\ln q_t(x_t)$

We want to minimize the accumulated loss relative to the loss of the best expert so that

$$-\sum_{t=1}^{T} \ln q_t(x_t) \leq \min \left[-\sum_{y} \ln p_{t,i}(x_t) \right] + \text{a small amount}$$

It turns out, as we will see next time, this problem is closely related to investment theory and coding theory.