COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Elena Sizikova

Lecture # 20
April 17, 2013

Last time, we began discussing how to learn a probability distribution D from samples. Suppose we are given:

1. a large but finite space X with $|X| = N$
2. D - an (unknown) distribution over X
3. $x_1, \ldots, x_m \sim D$ iid samples
4. $f_1, \ldots f_n$ features, where each feature is a function $f_j : X \to \mathbb{R}$

We would like to estimate the distribution D. Using the example of butterfly observance patterns from before, we could be estimating the distribution of butterfly positions from position features such as average temperature, average annual rainfall, altitude etc.

We have discussed two possible approaches. The first is to directly find a distribution q^* that maximizes the relative entropy:

$$P = \{q \ \ | \forall j : \ \ \mathbf{E}_q[f_j] = \hat{\mathbf{E}}[f_j]\}$$
$$q^* = \underset{q \in P}{\operatorname{argmax}} H(q) \qquad \text{where } H \text{ is the entropy function}$$

Above, $\mathbf{E}_q[f_j]$ and $\hat{\mathbf{E}}[f_j]$ are defined as before:

$$\mathbf{E}_q[f_j] = \mathbf{E}_{x \sim q}[f_j(x)] \quad \text{and} \quad \hat{\mathbf{E}}[f_j] = \frac{1}{m} \sum_{i=1}^{m} f_j(x_i)$$

Analytically, it is usually difficult to find a maximizer for H directly. The second approach is to use a parametric form, i.e. find a parameter $\boldsymbol{\lambda} = \langle \lambda_1, \ldots, \lambda_m \rangle$ in order to minimize:

$$L(\boldsymbol{\lambda}) = \underbrace{-\frac{1}{m} \sum_{i=1}^{m} \ln(q_{\boldsymbol{\lambda}}(x_i))}_{\text{form of log loss}} \tag{1}$$

Above, $q_{\boldsymbol{\lambda}}$ is defined to be:

$$q_{\boldsymbol{\lambda}}(x) = \frac{\exp\left(\sum_j \lambda_j f_j(x)\right)}{Z_\lambda}$$

where Z_λ is a normalization constant. We can rewrite q as:

$$q_{\boldsymbol{\lambda}}(x) = \frac{\exp(g_\lambda(x))}{Z_\lambda} \quad \text{where we use} \quad g_{\boldsymbol{\lambda}}(x) := \sum_{j=1}^{n} \lambda_j f_j(x)$$

The usual approach of setting the derivative of $q_{\boldsymbol{\lambda}}$ with respect to λ_j to 0 does not yield an easily solvable system of equations, so instead we will use an iterative method of finding

the minimum:

Choose $\boldsymbol{\lambda}_1$
For $t = 1, 2, \ldots$
    compute $\boldsymbol{\lambda}_{t+1}$ from $\boldsymbol{\lambda}_t$

i.e. we are trying to find a sequence of $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \ldots$ such that:

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

Above, we have assumed that the features are arbitrary functions, but in practice, we make the following assumptions (without loss of generality):

$$\forall x \forall j : f_j(x) \geq 0 \tag{2}$$

$$\forall x \quad : \sum_{j=1}^{n} f_j(x) = 1 \tag{3}$$

It is straightforward to justify why (2) and (3) come w.l.o.g. Adding a constant to the feature function does not affect distribution q, so we can assume (2). We can also scale the features to have range: $f_j : X \to \left[0, \frac{1}{n}\right]$, without affecting the distribution:

$$\sum_j f_j(x) \leq 1$$

Finally, we create a dummy feature f_0 defined by:

$$f_0(x) = 1 - \sum_{j=1}^{n} f_j(x) \quad \text{which again doesn't alter } q$$

we obtain (3) for the set of all features.

Consider the difference of loss functions after each iteration of the algorithm:

$$\Delta L = L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t)$$

We will derive a tractable approximation to ΔL and minimize it, since minimizing the loss at each step is equivalent to minimizing the overall loss. Let us focus on a particular round $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda'} = \boldsymbol{\lambda}_{t+1}$. We have:

$$\begin{aligned}
\Delta L &= L(\boldsymbol{\lambda'}) - L(\boldsymbol{\lambda}) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left[-\ln \left(\frac{\exp(g_{\lambda'}(x_i))}{Z_{\lambda'}} \right) + \ln \left(\frac{\exp(g_{\lambda}(x_i))}{Z_{\lambda}} \right) \right] \\
&= \frac{1}{m} \sum_i \left[g_{\lambda}(x_i) - g_{\lambda'}(x_i) \right] + \ln \left(\frac{Z_{\lambda'}}{Z_{\lambda}} \right)
\end{aligned} \tag{4}$$

For the first term in (4), write the update to λ as $\lambda'_j = \lambda_j + \alpha_j$. Then this term becomes:

$$\begin{aligned}
\frac{1}{m}\sum_i \left[g_\lambda(x_i) - g_{\lambda'}(x_i)\right] &= \frac{1}{m}\sum_i \sum_j \left(\lambda_j f_j(x_j) - \lambda'_j f_j(x_j)\right) \\
&= \frac{-1}{m}\sum_i \sum_j \alpha_j f_j(x_i) \\
&= -\sum_j \alpha_j \underbrace{\left(\frac{1}{m}\sum_i f_j(x_i)\right)}_{\text{empirical average of } f_j} \\
&= -\sum_j \alpha_j \hat{\mathbb{E}}[f_j]
\end{aligned}$$

Now, rewriting the second term:

$$\begin{aligned}
\frac{Z_{\lambda'}}{Z_\lambda} &= \frac{\sum_{x\in X} \exp\left(\sum_j \lambda'_j f_j(x)\right)}{Z_\lambda} \\
&= \sum_{x\in X} \underbrace{\frac{\exp\left(\sum_j \lambda_j f_j(x)\right)}{Z_\lambda}}_{q_\lambda(x)} \cdot \exp\left(\sum_j \alpha_j f_j(x)\right) \\
&= \sum_x q_\lambda(x) \exp\left(\sum_j \alpha_j f_j(x)\right)
\end{aligned}$$

Note that for each x, the feature values $f_1(x), \ldots, f_n(x)$ form a distribution by our assumptions (2) and (3). Also $\sum_j \alpha_j f_j(x)$ is a weighted average of the α_js. Using convexity of the exponential function, we have:

$$\begin{aligned}
\frac{Z_{\lambda'}}{Z_\lambda} &\le \sum_x q_\lambda(x) \sum_j f_j(x) e^{\alpha_j} \\
&= \sum_j e^{\alpha_j} \underbrace{\sum_x q_\lambda(x) f_j(x)}_{\mathbb{E}_{q_\lambda}[f_j]}
\end{aligned}$$

Finally, going back to (4), we have:

$$\begin{aligned}
\Delta L &= \frac{1}{m}\sum_i \left[g_\lambda(x_i) - g_{\lambda'}(x_i)\right] + \ln\left(\frac{Z_{\lambda'}}{Z_\lambda}\right) \\
&\le -\sum_j \alpha_j \hat{\mathbf{E}}[f_j] + \ln\left(\sum e^{\alpha_j} \mathbf{E}_{q_\lambda}(f_j)\right) \\
&= -\sum_j \alpha_j \hat{\mathbf{E}}_j + \ln\left(\sum_j e^{\alpha_j} \mathbf{E}_j\right) \qquad (5)
\end{aligned}$$

where we define $\hat{\mathbf{E}}_j := \hat{\mathbf{E}}[f_j]$ and $\mathbf{E}_j := \mathbf{E}_{q_\lambda}(f_j)$. Notice that we can now optimize the RHS of (5) directly, by taking partial derivatives:

$$0 = \frac{\partial}{\partial \alpha_j} = -\hat{\mathbf{E}}_j + \frac{\mathbf{E}_j e^{\alpha_j}}{\sum_j \mathbf{E}_j e^{\alpha_j}}$$

Notice that if α_j is a solution to the above, then so is $\alpha_j + c$ for a constant c, and so we choose c so that the denominator $\sum_j \mathbf{E}_j e^{\alpha_j}$ is equal to zero. We thus find a solution where $\alpha_j = \ln\left(\hat{\mathbf{E}}_j/\mathbf{E}_j\right)$. It follows that the algorithm's iterative update on round t is:

$$\lambda_{t+1,j} = \lambda_{t,j} + \ln\left(\frac{\hat{\mathbf{E}}[f_j]}{\mathbf{E}_{q_{\boldsymbol{\lambda}_t}}(f_j)}\right)$$

we hope that this process converges to the optimal value of $\boldsymbol{\lambda}$.

Thus, it remains to prove convergence. Define $p_t = q_{\lambda_t}$.
Definition. A function $A : \{$ probability distributions over $X\} \to \mathbb{R}$ is an *auxiliary* function if it satisfies the following requirements:

1. A is continuous
2. $L(\boldsymbol{\lambda_{t+1}}) - L(\boldsymbol{\lambda_t}) \leq A(p_t) \leq 0$. (We want ≤ 0 so that the loss is always decreasing.)
3. If for some distribution p, $A(p) = 0$, then $\mathbf{E}_p[f_j] = \hat{\mathbf{E}}[f_j]$ for all j. In other words, $p \in P$.

Theorem. $p_t \to q^*$.
We first prove that if an auxiliary function A exists, then the theorem statement holds.

Suppose A is an auxiliary function. We know that $L \geq 0$ by properties of $\ln(x)$ and definition of L. By the second property of auxiliary functions, the loss L is decreasing and bounded below by 0, so $L(\boldsymbol{\lambda_{t+1}}) - L(\boldsymbol{\lambda_t}) \to 0$, and thus $A(p_t) \to 0$ as $t \to \infty$.

Now, we consider what happens at the limit of t. Suppose $p = \lim_{t\to\infty} p_t$. Since for all t, $p_t \in \overline{Q}$, where $\overline{Q}$ is the closure of Q, we have that $p \in \overline{Q}$. Also, since A is continuous,

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

Thus, $p \in P$. But now we have proved that $p \in P$ and $p \in \overline{Q}$, so $p \in P \cap \overline{Q}$. As we have stated (without proof) a theorem that $P \cap \overline{Q} = \{q^*\}$, it follows that $p = q^*$.
(This assumes that the limit $\lim_{t\to\infty} p_t$ exists. If it does not exist, applying general results from real analysis (which are slightly beyond the scope of this class), we know that $\{p_t | t = 0, 1, \ldots\}$ belong to a compact subspace of $\mathbb{R}^n$, and so there is a convergent subsequence of p_t's. By the same proof just given, this subsequence must converge to q^*. Thus, the only limit point of this subsequence is q^*. Therefore, by general results from real analysis, the entire sequence p_t converges to q^*.)

We now have:

$$\begin{aligned}
\Delta L &\leq -\sum_j \alpha_j \hat{\mathbf{E}}_j + \ln\left(\sum_j e^{\alpha_j}\mathbf{E}_j\right) \\
&= -\sum_j \hat{\mathbf{E}}_j \ln\frac{\hat{\mathbf{E}}_j}{\mathbf{E}_j} + \ln\left(\sum_j \hat{\mathbf{E}}_j\right) \quad \text{using } \alpha_j = \ln\left(\hat{\mathbf{E}}_j/\mathbf{E}_j\right) \qquad (6)
\end{aligned}$$

Now,for any distribution q,

$$\sum_j \mathbf{E}_q[f_j] = \mathbf{E}_q\left[\sum_j f_j(x)\right] = \mathbf{E}_q[1] = 1$$

and therefore $\mathbf{E}_q[f_1], \ldots, \mathbf{E}_q[f_n]$ forms a distribution. In particular, this means that $\hat{\mathbf{E}}_j$ and $\mathbf{E}_j$ form distributions.

So in (6), we find that the second term simplifies to:

$$\ln\left(\sum_j \hat{\mathbf{E}}_j\right) = \ln 1 = 0$$

Hence we can rewrite (6) in terms of relative entropy:

$$\Delta L \leq -RE\left(\left\langle \hat{\mathbf{E}}[f_1], \ldots, \hat{\mathbf{E}}[f_n]\right\rangle \,||\, \langle \mathbf{E}_{p_t}(f_1), \ldots, \mathbf{E}_{p_t}(f_n)\rangle\right)$$

Now, define:

$$A(p) := -RE\left(\left\langle \hat{\mathbf{E}}[f_j]\right\rangle \,||\, \langle \mathbf{E}_p(f_j)\rangle\right)$$

where $\left\langle \hat{\mathbf{E}}[f_j]\right\rangle := \left\langle \hat{\mathbf{E}}[f_1], \ldots, \hat{\mathbf{E}}[f_n]\right\rangle$ and $\langle \mathbf{E}_p(f_j)\rangle := \langle \mathbf{E}_p(f_1), \ldots, \mathbf{E}_p(f_n)\rangle$.

It remains to verify that A is an auxiliary function. Clearly A satisfies properties 1 and 2 (continuity and non-positivity) of auxiliary functions, by properties of relative entropy. Now, relative entropy is zero iff two distributions are indentical, so $A(p) = 0$ implies $\hat{\mathbf{E}}[f_j] = \mathbf{E}_p(f_j)$ for all j, i.e. $p \in P$. $\square$

Observe that we have not addressed over how quickly does the given algorithm converge, but this is out of scope of the lecture.

Next: The above algorithm applies to the batch setting. The following is an outline of an analogous algorithm for the online setting, that we will explore next time:

For round $t = 1, \ldots, T$:
Each expert i chooses distribution $p_{t,i}$ over X
Master combines the distribution into its own distribution q_t
Observe $x_t \in X$
Evaluate loss $= -\ln q_t(x_t)$

We want:

$$\sum_{t=1}^{T} -\ln q_t(x_t) \leq \min_i \underbrace{\sum_{t=1}^{T} -\ln p_{t,i}(x_t)}_{\text{loss of expert } i} + \text{small regret}$$