COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Udaya Ghai

Lecture # 20
April 18, 2018

1 Introduction and Recap

In the previous lecture, we introduced a specific instance of learning a probability distribution. Recall, we wanted to model the probability distribution of butterfly observations as a function of known features of the area. Formally, we were given:

- A finite domain $\mathcal{X}$ with size $|\mathcal{X}| = N$
- A distribution $\mathcal{D}$ over $\mathcal{X}$
- An iid sample drawn from the distribution, $x_1, \ldots, x_m \sim \mathcal{D}$
- A set of features, $f_1, \ldots, f_n$, where each feature is a function $f_j : \mathcal{X} \to \mathbb{R}$

We define shorthands for the empirical average and the expectation of a feature f as

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

We will also use the notation $\Delta_{\mathcal{X}}$ to represent the probability distributions over $\mathcal{X}$.

We explored two approaches for tackling this problem. The first was to find the maximum entropy distribution such that the expectation of each feature matches its empirical average. Alternatively, we looked at maximizing likelihood amongst an exponential family of distributions. In the last class, we saw that these problems were equivalent to each other through the duality theorem:

Theorem 1 *Let*

$$\mathcal{P} = \{q \in \Delta_{\mathcal{X}} : \mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j] \ \forall j\}$$

and

$$\mathcal{Q} = \{q_\lambda : \boldsymbol{\lambda} \in \mathbb{R}^n\}, \text{ where } q_\lambda(x) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_\lambda}.$$

The following are equivalent and have unique solution:

$$q^* = \text{argmax}_{q\in\mathcal{P}}\, H(q) \tag{1}$$

$$q^* = \text{argmax}_{q\in\overline{\mathcal{Q}}} \sum_{i=1}^{m} \ln q(x_i) \tag{2}$$

$$q^* \in \mathcal{P} \cap \overline{\mathcal{Q}} \tag{3}$$

While this gives an interesting characterization of a solution, we still don't know how to find it. In this class we have mostly been concerned with learning, but here we will shift our focus towards optimization and provide an example of one particular approach.

2 Solving the Optimization Problem

To solve this problem, we first need to look at which equivalent notion to optimize. The maximum liklihood formulation seems like a good choice as it can be written as an unconstrained optimization problem. We rewrite this as a loss-minimization problem by negating:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^n} L(\boldsymbol{\lambda}) = -\frac{1}{m} \sum_{i=1}^{m} \ln q_\lambda(x_i)$$

$$g_\lambda(x) = \sum_{j=1}^{n} \lambda_j f_j(x) \quad \text{and} \quad q_\lambda(x) = \frac{e^{g_\lambda(x)}}{Z_\lambda}$$

At a very high level, we would like to find an iterative update that converges to a good parameterization $\boldsymbol{\lambda}$. This would be some algorithm of the form:

- Choose $\boldsymbol{\lambda}_1$
- For $t = 1 \dots T$
  - Compute $\boldsymbol{\lambda}_{t+1}$ from $\boldsymbol{\lambda}_t$

We would like the loss to be reduced numerically in each round. We also want the loss to converge to the minimal possible loss so $L(\boldsymbol{\lambda}_t) \to \inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$.

In order to illustrate a more general technique, we will go through an example of one way to do this. The way we will approach this optimization problem is to come up with an *approximation* of $L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t)$ and *exactly* minimize this approximation[1]. At each point we will have a new approximation to minimize. Before we proceed, we will make some simplifying assumptions.

2.1 Simplifying Assumptions

A useful assumption that we will make is that, for each $x \in \mathcal{X}$, the features lie in the probability simplex. That is, for each $x \in \mathcal{X}$ we have

$$\sum_{j=1}^{n} f_j(x) = 1 \quad \text{and} \quad \forall j f_j(x) \geq 0.$$

This assumption can be made without loss of generality. To see this, we can first transform the features such that they are all positive. For each feature f_j, we can find $c_j = \min_{x \in \mathcal{X}} f_j(x)$ and make the transformation $f'_j(x) = f_j(x) - c_j$ so the minimum value for all features will be exactly 0. We note that the probability distributions before and after this transformation are unchanged as

$$g'_\lambda(x) = \sum_{j=1}^{n} \lambda_j f'_j(x) = \sum_{j=1}^{n} \lambda_j f_j(x) + \sum_{j=1}^{n} \lambda_j c_j = g_\lambda(x) + \sum_{j=1}^{n} \lambda_j c_j$$

[1] If we didn't use an approximation, there would be no point to an iterative algorithm as we would immediately find ourselves at the global minimum loss.

Adding a constant to g_λ is just a scaling of q_λ so the change will be normalized away by Z_t. Now each feature can be restricted to $[0, 1/n]$ by scaling the features by

$$b_j = \frac{1}{n \max_{x \in \mathcal{X}} f_j(x)}.$$

Scaling does not change which distributions can be represented. This leaves us with nonnegative features f_j such that $\sum_{j=1}^n f_j(x) \leq 1$. We can add a dummy feature f_0 where $f_0(x) = 1 - \sum_{j=1}^n f_j(x)$ so the features all sum to 1. We note that this adds a term $\lambda_0(1 - \sum_{j=1}^n f_j(x))$ to $g_\lambda(x)$. This does not affect the functions that can be represented because λ_0 is just a constant term and for each j, $-\lambda_0 f_j(x)$ can be absorbed by the term $\lambda_j f_j(x)$.

2.2 Approximating the Change in Loss

We will work towards finding some upper bound on this difference that has a form that can be minimized directly. We will focus on a particular round t, so for notational convenience, we write $\boldsymbol{\lambda}$ for $\boldsymbol{\lambda}_t$, $\boldsymbol{\lambda}'$ for $\boldsymbol{\lambda}_{t+1}$, and we define the change α using $\lambda'_j = \lambda_j + \alpha_j$. We now define ΔL to be the change in loss:

$$\begin{aligned}
\Delta L &= L(\boldsymbol{\lambda}') - L(\boldsymbol{\lambda}) \\
&= -\frac{1}{m}\sum_{i=1}^m \ln q_{\lambda'}(x_i) + \frac{1}{m}\sum_{i=1}^m \ln q_\lambda(x_i) && \text{Expanding log-loss} \\
&= -\frac{1}{m}\sum_{i=1}^m \ln\left(\frac{e^{g_{\lambda'}(x_i)}}{Z_{\lambda'}}\right) + \frac{1}{m}\sum_{i=1}^m \ln\left(\frac{e^{g_\lambda(x_i)}}{Z_\lambda}\right) && \text{Plugging in definition of } q_\lambda \\
&= \frac{1}{m}\sum_{i=1}^m (g_\lambda(x_i) - g_{\lambda'}(x_i)) + \ln\frac{Z_{\lambda'}}{Z_\lambda} && \text{Straightforward algebra}
\end{aligned}$$

We will now handle both of these terms separately. Plugging in the definition of g_λ and rearranging the sums we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m (g_\lambda(x_i) - g_{\lambda'}(x_i)) &= \frac{1}{m}\sum_{i=1}^m \sum_{j=1}^n (\lambda_j f_j(x) - \lambda'_j f_j(x)) && \text{Applying definition of } g \\
&= -\frac{1}{m}\sum_{i=1}^m \sum_{j=1}^n \alpha_j f_j(x) && \text{Rewriting in terms of } \boldsymbol{\alpha} \\
&= -\sum_{j=1}^n \alpha_j \frac{1}{m}\sum_{i=1}^m f_j(x) && \text{Rearranging the sums} \\
&= -\sum_{j=1}^n \alpha_j \hat{\mathbb{E}}[f_j] && \text{Using definition of empirical mean}
\end{aligned}$$

Now we bound $\dfrac{Z_{\lambda'}}{Z_\lambda}$.

$$\frac{Z_{\lambda'}}{Z_\lambda} = \frac{\sum_{x\in\mathcal{X}} \exp\left(g_\lambda(x) + \sum_{j=1}^n \alpha_j f_j(x)\right)}{Z_\lambda} \qquad \text{Expanding } Z_{\lambda'}$$

$$= \frac{\sum_{x\in\mathcal{X}} \exp(g_\lambda(x)) \exp\left(\sum_{j=1}^n \alpha_j f_j(x)\right)}{Z_\lambda} \qquad \text{Simple algebra}$$

$$= \sum_{x\in\mathcal{X}} q_\lambda(x) \exp\left(\sum_{j=1}^n \alpha_j f_j(x)\right) \qquad \text{Applying the definition of } q_\lambda$$

We now note that because the exponential function is convex and $(f_1(x)\dots f_j(x))$ forms a probability distribution, we can use Jensen's inequality which gives us

$$\exp\left(\sum_{j=1}^n \alpha_j f_j(x)\right) \leq \sum_{j=1}^n f_j(x) e^{\alpha_j}.$$

Plugging this in and rearranging, we have

$$\frac{Z_{\lambda'}}{Z_\lambda} \leq \sum_{x\in\mathcal{X}} q_\lambda(x) \sum_{j=1}^n f_j(x) e^{\alpha_j}$$

$$= \sum_{j=1}^n e^{\alpha_j} \sum_{x\in\mathcal{X}} q_\lambda(x) f_j(x) = \sum_{j=1}^n e^{\alpha_j} \mathbb{E}_{q_\lambda}[f_j]$$

Putting everything back together, we have the following bound:

$$\Delta L \leq -\sum_{j=1}^n \alpha_j \underbrace{\hat{\mathbb{E}}[f_j]}_{\hat{E}_j} + \ln\left(\sum_{j=1}^n e^{\alpha_j} \underbrace{\mathbb{E}_{q_\lambda}[f_j]}_{E_j}\right)$$

2.3 Minimizing the Approximation

We now have $\Delta L \leq B(\boldsymbol{\alpha}) = -\sum_{j=1}^n \alpha_j \hat{E}_j + \ln\left(\sum_{j=1}^n e^{\alpha_j} E_j\right)$. We can minimize this approximation with calculus

$$\frac{\partial B}{\partial \alpha_j} = -\hat{E}_j + \frac{E_j e^{\alpha_j}}{\sum_{j=1}^n E_j e^{\alpha_j}} = 0$$

By the same reasoning as our simplifying assumptions, adding a constant shift to $\boldsymbol{\alpha}$ does not change $B(\boldsymbol{\alpha})$, so we can assume $\boldsymbol{\alpha}$ is normalized such that $\sum_{j=1}^n E_j e^{\alpha_j} = 1$. We end up with

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

2.4 The Algorithm and Intuition

What we've done so far brings us to the final iterative update on the dual variables $\boldsymbol{\lambda}$:

$$\lambda_{t+1,j} = \lambda_{t,j} + \ln\left(\frac{\hat{\mathbb{E}}[f_j]}{\mathbb{E}_{q_{\lambda_t}}[f_j]}\right).$$

This has a very simple form and is easy to compute[2]. Still, it's not yet clear how this is going to progress over time. To get some intuition, we can translate this update into the primal space of distributions by plugging in the definition of q_λ. Here if we let $p_t = q_{\lambda_t}$, we end up with the following update:

$$p_{t+1}(x) \propto p_t(x) \prod_{j=1}^{n} \left(\frac{\hat{\mathbb{E}}[f_j]}{\mathbb{E}_{p_t}[f_j]}\right)^{f_j(x)}$$

If we look at the duality theorem, then we know that we want to reach $q^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$. The constraints for $\mathcal{P}$ correspond to $\mathbb{E}_p[f_j] = \hat{\mathbb{E}}[f_j]$ for all j. If these are all satisfied, then we clearly are at a fixed point. On the other hand, suppose at time t $\mathbb{E}_{p_t}[f_j] < \hat{\mathbb{E}}[f_j]$. Intuitively, we want $\mathbb{E}_{p_{t+1}}[f_j]$ to be larger. This is encouraged as $\dfrac{\hat{\mathbb{E}}[f_j]}{\mathbb{E}_{p_t}[f_j]} > 1$. If $f_j(x)$ is big, it's probability is increased more, resulting in a higher expectation.

2.5 Proving Convergence

While this intuition is nice, we want to be able to prove that p_t converges[3] to q^*. We will do this in two parts. First, we will define properties on an approximation function A for the change in loss that will guarantee convergence. Then, we will prove that the approximation we have chosen to use satisfies these criteria.

Definition 2 *A function defined on the probability simplex, $A : \Delta_{\mathcal{X}} \to \mathbb{R}$ is an auxiliary function if:*

1. *A is continuous.*
2. *$L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t) \leq A(p_t) \leq 0$.*
3. *$A(p) = 0 \Rightarrow p \in \mathcal{P}$.*

Claim 3 *If an auxiliary function exists for the process $\boldsymbol{\lambda}_t$ then $p_t \to q^*$.*

Proof: The log-loss $L(\boldsymbol{\lambda})$ is bounded below by 0 because each term in the sum must be nonnegative. By property (2) of an auxiliary function, we know that the differences $L(\lambda_t)$ must be monotonically decreasing. Therefore, we know that the differences $L(\boldsymbol{\lambda}_{t+1}) - L(\boldsymbol{\lambda}_t)$ must converge to 0 (and so $A(p_t)$ must also converge to 0), otherwise $L(\lambda_t)$ would eventually

[2] As long as N is not gigantic.

[3] Ideally, we would like this convergence to be fast, but all we will prove here is the distributions must eventually converge to the solution.

decrease below 0.
Now we assume the limit of p_t exists[4]. We want to show that

$$p = \lim_{t\to\infty} p_t \in \mathcal{P} \cap \overline{\mathcal{Q}}$$

$p \in \overline{\mathcal{Q}}$ by definition of the closure of a set as each $p_t \in \mathcal{Q}$. Now by the continuity of A (property (1)), we have

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

Therefore, by property (3), $p \in \mathcal{P}$. It follows that $p \in \mathcal{P} \cap \overline{\mathcal{Q}} = q^*$ as desired. □

Theorem 4 $p_t \to q^*$

Proof: It suffices to show that an auxiliary function A exists. We consider the approximation we derived with

$$\Delta L \leq A(p_t) = -\sum_{j=1}^{n} \alpha_j \hat{\mathbb{E}}[f_j] + \ln\left(\sum_{j=1}^{n} e^{\alpha_j} \mathbb{E}_{p_t}[f_j]\right)$$

We first note that when we chose $\alpha_j = \ln\left(\frac{\hat{\mathbb{E}}[f_j]}{\mathbb{E}_{p_t}[f_j]}\right)$ in the minimization, we enforced that $\sum_{j=1}^{n} e^{\alpha_j}\mathbb{E}_{p_t}[f_j] = 1$. As a result, our approximation simplifies to

$$A(p_t) = -\sum_{j=1}^{n} \alpha_j \hat{\mathbb{E}}[f_j] = -\sum_{j=1}^{n} \hat{\mathbb{E}}[f_j] \ln\left(\frac{\hat{\mathbb{E}}[f_j]}{\mathbb{E}_{p_t}[f_j]}\right).$$

This is exactly the form of the negated relative entropy function.

By our simplifying assumptions, for each $x \in \mathcal{X}$ the vector $\mathbf{f}(x) = (f_1(x) \dots f_n(x))$ is a probability distribution. Both the vectors of empirical averages $\hat{\mathbb{E}}[\mathbf{f}] = (\hat{\mathbb{E}}[f_1] \dots \hat{\mathbb{E}}[f_n])$ and the expectation in respect to p_t, $\mathbb{E}_{p_t}[\mathbf{f}] = (\mathbb{E}_{p_t}[f_1] \dots \mathbb{E}_{p_t}[f_n])$ are convex combinations of $\mathbf{f}(x)$, so these are also probability vectors. Thus, we can rewrite our approximation function A as

$$A(p) = -\mathrm{RE}(\hat{\mathbb{E}}[\mathbf{f}] \parallel \mathbb{E}_p[\mathbf{f}]).$$

Now the three properties are easy to prove. The relative entropy function is continuous satisfying property (1). It's nonnegative, so it's negation is nonpositive, and by our derivation $\Delta L \leq A(p_t)$, so property (2) is satisfied. Finally, relative entropy is 0 only if the two distributions are the same, so $A(p) = \mathrm{RE}(\hat{\mathbb{E}}[\mathbf{f}] \parallel \mathbb{E}_p[\mathbf{f}]) = 0 \Rightarrow \hat{\mathbb{E}}[\mathbf{f}] = \mathbb{E}_p[\mathbf{f}]$. This says that the expectations of the features in respect to p match the empirical expectations, or in other words $p \in \mathcal{P}$, so property (3) is satisfied. □

[4]This assumption isn't actually necessary. Because the probability simplex is a compact space, we know that the p_ts belong to a compact space. Therefore, there must be a convergent subsequence, which must converge to q^* using the same proof as when the limit exists. Since the sequence can only have a single limit point (by the uniqueness of q^*), it can be argued that the entire sequence must converge to q^*.

3 Next: On-line Log-loss

We just worked out an algorithm for minimizing the log loss for the exponential family in the batch setting. The next thing we will do is to consider minimizing log loss in an online model. Consider betting on horse racing as an example. Experts will provide probabilities of different horses winning in rounds. In each round, a learner takes all these expert distributions and must play a single distribution. The learner will then observe a winning horse and observe a penalty $\ln 1/q$ if the learner assigns probability q to that horse. More generally, we have

- for $t = 1 \dots T$
  - Each expert i chooses a distribution $p_{t,i} \in \Delta_{\mathcal{X}}$
  - Learner chooses distribution $q_t \in \Delta_{\mathcal{X}}$
  - Observe $x_t \in \mathcal{X}$
  - Incur loss $-\ln q_t(x_t)$

We want to find algorithms that incur small regret in comparison to the best expert. Formally, we want

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