6.883: Online Methods in Machine Learning
Alexander Rakhlin

LECTURE 12

1. EXAMPLE: PREDICTION ON GRAPHS, CONTINUED

Let us continue the example of prediction on graphs. Recall that at each step, the prediction method needs to compute $\phi(\tilde{\boldsymbol{y}})$ where $\tilde{\boldsymbol{y}} = (y_{1:t-1}, +1, \epsilon_{t+1:n})$, or the version with a minus sign at the tth position. We proposed the following definition of ϕ:

$$F_K = \{f \in \{\pm 1\}^n : f^\mathsf{T} L f \le K\}, \qquad \phi(\boldsymbol{y}) = d_H(\boldsymbol{y}, F_K) + C_n$$

However, it might be difficult to compute the Hamming distance. Essentially it is asking for the number of changes one needs to make to the labeling $\boldsymbol{y}$ of vertices to bring it to the set with cut value at most K.

The idea is to enlarge F, thus decreasing the Hamming distance, but increasing the Rademacher averages C_n for this larger set. Let us illustrate this approach. Write

$$d_H(\boldsymbol{y}, F_K) = \min_{f \in F_K} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{f_i \ne y_i\} = \frac{1}{2} - \frac{1}{2n} \max_{f \in F_K} \langle f, \boldsymbol{y} \rangle \tag{1}$$

Further,

$$\max_{f \in F_K} \langle f, \boldsymbol{y} \rangle \le \max_{f \in [-1,1]^n, f^\mathsf{T} L f \le K} \langle f, \boldsymbol{y} \rangle \tag{2}$$

by going to the real-valued vectors. Thus, let us take

$$\phi'(\boldsymbol{y}) = \frac{1}{2} - \frac{1}{2n} \max_{f \in [-1,1]^n, f^\mathsf{T} L f \le K} \langle f, \boldsymbol{y} \rangle + C'_n \tag{3}$$

We need to check that ϕ' is smooth. We leave it as a homework exercise. We can now use ϕ' in our prediction algorithm, since its computation is a convex optimization problem. We will later provide an even better solution based on hinge loss (and one that works much better in practice).

The hope now is that C'_n is not too large (and, in particular, still $o(1)$). We have

$$C'_n = \frac{1}{2n} \mathbb{E} \max_{f \in [-1,1]^n, f^\mathsf{T} L f \le K} \langle f, \boldsymbol{\epsilon} \rangle . \tag{4}$$

Let us overbound the above expression to get a closed-form solution. Notice that

$$[-1,1]^n \subset \{f : f^\mathsf{T} I f \le n\}$$

and thus

$$[-1,1]^n \cap \{f : f^\mathsf{T} L f \le K\} \subset \left\{ f : f^\mathsf{T} \left(\frac{I}{2n} + \frac{L}{2} \right) f \le 1 \right\}$$

(prove this!) Hence,

$$C'_n = \frac{1}{2n} \mathbb{E} \sqrt{\boldsymbol{\epsilon}^\mathsf{T} M^{-1} \boldsymbol{\epsilon}} \tag{5}$$

for $M = \frac{I}{2n} + \frac{L}{2}$. This can now be analyzed via spectral properties of the graph G. Homework: show that (5) is upper bounded in terms of eigenvalues of L.

1.1 Model selection

In the previous lecture, we considered the star graph and argued that the cut value is $n-1$ (very large!) for the labeling $\boldsymbol{y}$ assigning a -1 to the center and $+1$ to the rest, yet this labeling is Hamming distance 1 from the constant labeling (all $+1$), and thus the number of mistakes on this sequence will be small. Let us now consider a different example which will motivate the question of model selection.

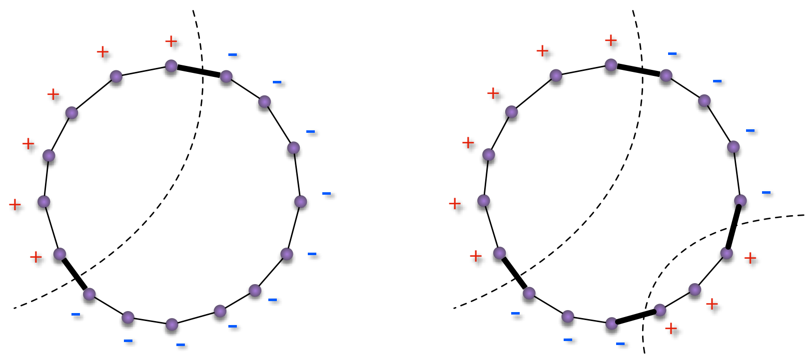

Figure 1: Labeling with cuts of size 2 and 4.

Consider a ring graph with n vertices. Suppose we choose $K = 2$. That is, we take F_2 to be the set of $\boldsymbol{y}$ that have either zero or two switches in sign. However, consider a labeling with cut value 4. The Hamming distance to F_2 may be $\Omega(n)$, and thus we obtain a very weak bound on the number of mistakes incurred by the associated prediction algorithm. The issue here is that K was not chosen "correctly", and the mistake bound is very sensitive to this choice. The question is whether one can choose the best K for the given sequence, as if it were known a priori. This is a model selection question, and we will see that it is possible!

2. BINARY PREDICTION WITH INDICATOR LOSS AND SIDE INFORMATION

Consider now a supervised learning scenario, where covariates $x_1, \ldots, x_n$ are drawn i.i.d. from some (unknown) marginal distribution P_X. The sequence $y_1, \ldots, y_n$ is still assumed to be arbitrary.

For $t = 1, \ldots, n$
  Observe an independent draw $x_t \sim P_X$
  Predict $\widehat{y}_t \in \{\pm 1\}$
  Observe outcome $y_t \in \{\pm 1\}$

What happens to the argument in the previous lecture? Let ϕ now be a function of two sequences: $\phi : \mathcal{X}^n \times \{\pm 1\}^n \to \mathbb{R}$ and suppose

$$|\phi(x_{1:n}, y_{1:t-1}, +1, y_{t+1:n}) - \phi(x_{1:n}, y_{1:t-1}, -1, y_{t+1:n})| \leq 1/n. \tag{6}$$

We will prove the following generalization of Cover's result.

Theorem 1. *Let $\phi : (\mathcal{X} \times \{\pm 1\})^n \to \mathbb{R}$ be such that (6) holds, and suppose that x_t's are i.i.d. Then there exists a prediction strategy (specified later in Algorithm 1) such that*

$$\forall y_{1:n}, \qquad \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\}\right] \leq \mathbb{E}\phi(x_{1:n}, y_{1:n}) \tag{7}$$

if and only if

$$\mathbb{E}\phi(x_{1:n}, \epsilon_{1:n}) \geq 1/2. \tag{8}$$

Above, the expectation on the left-hand side of (7) is over the randomization of the algorithm and the x's, while on the right-hand side is over x's. In (8), the expectation is both over the x's and over the independent Rademacher random variables.

Proof. Having observed $x_{1:n-1}, y_{1:n-1}$ and x_n at the present time step, we need to solve

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n}\mathbf{1}\{\widehat{y}_n \neq y_n\}\right] + \mathbf{Rel}(x_{1:n}, y_{1:n}) \right\} \tag{9}$$

For the last time step, take $\mathbf{Rel} = -\phi$. The same steps as before lead to the solution

$$q_n(x_{1:n}, y_{1:n-1}) = n(\phi(x_{1:n}, y_{1:n-1}, -1) - \phi(x_{1:n}, y_{1:n-1}, +1)), \tag{10}$$

We point out that q_n depends on x_n, as given by the protocol of the problem. Then (9) is upper bounded by

$$\mathbb{E}_{\epsilon_n}\mathbf{Rel}(x_{1:n}, y_{1:n-1}, \epsilon_n) + \frac{1}{2n} = -\mathbb{E}_{\epsilon_n}\phi(x_{1:n}, y_{1:n-1}, \epsilon_n) + \frac{1}{2n} \tag{11}$$

We now take expectation over x_n with respect to the unknown P_X on both sides:

$$\mathbb{E}_{x_n} \min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\}\right] + \mathbf{Rel}(x_{1:n}, y_{1:n}) \right\} \tag{12}$$

$$\leq \mathbb{E}_{x_n, \epsilon_n}\mathbf{Rel}(x_{1:n}, y_{1:n-1}, \epsilon_n) + \frac{1}{2n} \tag{13}$$

$$= -\mathbb{E}_{x_n, \epsilon_n}\phi(x_{1:n}, y_{1:n-1}, \epsilon_n) + \frac{1}{2n} \tag{14}$$

$$\triangleq \mathbf{Rel}(x_{1:n-1}, y_{1:n-1}) \tag{15}$$

It is not hard to see (verify this!) that the argument continues back to $t = 1$, with

$$\mathbf{Rel}(x_{1:t}, y_{1:t}) = -\mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}}\phi(x_{1:n}, y_{1:t}, \epsilon_{t+1:n}) + \frac{n-t}{2n} \tag{16}$$

and

$$\begin{aligned} &q_t(x_{1:t}, y_{1:t-1}) \\ &= n\left[\mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}}\phi(x_{1:n}, y_{1:t-1}, -1, \epsilon_{t+1:n}) - \mathbb{E}_{x_{t+1:n}, \epsilon_{t+1:n}}\phi(x_{1:n}, y_{1:t-1}, +1, \epsilon_{t+1:n})\right] \end{aligned} \tag{17}$$

just as in the previous lecture, and the initial condition $\mathbf{Rel}(\varnothing) \leq 0$ is

$$\mathbb{E}\phi(x_{1:n}, \epsilon_{1:n}) \geq \frac{1}{2}.$$

An attentive reader will notice, however, that the algorithm is not implementable: it requires the knowledge of P_X. However, all we need is to be able to sample $x_{t+1:n} \sim P_X$ and independent Rademacher $\epsilon_{t+1:n}$, and define

$$\widehat{q}_t(x_{1:n}, y_{1:t-1}, \epsilon_{t+1:n}) = n\left[\phi(x_{1:n}, y_{1:t-1}, -1, \epsilon_{t+1:n}) - \phi(x_{1:n}, y_{1:t-1}, +1, \epsilon_{t+1:n})\right]. \tag{18}$$

Regarding the required smoothness condition on ϕ, we see that it is simply that (17) is within the range $[-1, 1]$. In particular, it is implied by the assumed smoothness condition. $\square$

- In conclusion, we can solve the online classification problem with i.i.d. covariates if we have access to independent draws from the distribution. In particular, this step can be implemented with *unlabeled data* which is often available in practice.
- Importantly, the reason we were able to use "random playout" is because the solution q_t was in the form of an expectation. In examples we will study later in the course, q_t will not be in such a nice form, and the straightforward argument for random playout fails. However, there is a different argument that will be shown work.
- We also remark that the fact that $x_{1:n}$ are i.i.d. was not really used. All we require is that we are able to sample continuation of paths $P(x_{t+1}|x_{1:t})$ from the conditional distribution.

Perhaps, it's worth writing down the algorithm explicitly:

Algorithm 1 Online Supervised Binary Classification

Input: smooth potential function $\phi : (\mathcal{X} \times \{\pm 1\})^n \to \mathbb{R}$
for t=1,...,T **do**
  Observe x_t
  Draw $x_{t+1}, \ldots, x_n$ (e.g. as unlabeled data)
  Draw $\epsilon_{t+1}, \ldots, \epsilon_n$ independent Rademacher
  Compute

$$\widehat{q}_t(x_{1:n}, y_{1:t-1}, \epsilon_{t+1:n}) = n\Big[\phi(x_{1:n}, y_{1:t-1}, -1, \epsilon_{t+1:n}) - \phi(x_{1:n}, y_{1:t-1}, +1, \epsilon_{t+1:n})\Big]$$

  Predict by drawing $\widehat{y}_t$ from the distribution on $\{\pm 1\}$ with mean $\widehat{q}_t$
end for