Online-to-batch conversion

Daniel Hsu (COMS 4771)

Helmbold and Warmuth’s online-to-batch conversion

Consider an online learning algorithm that operates as follows on a data sequence \((x_1, y_1), (x_2, y_2), \ldots \) from \(\mathcal{X} \times \mathcal{Y}\):

- Initially, the learner picks a predictor \(\hat{f}_1 : \mathcal{X} \to \mathcal{Y}\).
- For each round \(t = 1, 2, \ldots\):
  1. The learner observes the context \(x_t \in \mathcal{X}\).
  2. The learner makes a prediction \(\hat{y}_t := \hat{f}_t(x_t) \in \mathcal{Y}\).
  3. The learner observes the outcome \(y_t \in \mathcal{Y}\).
  4. If \(\hat{y}_t \neq y_t\), the learner has made a mistake.
  5. The learner chooses a new predictor \(\hat{f}_{t+1} : \mathcal{X} \to \mathcal{Y}\).

Let \(\hat{f}_1, \ldots, \hat{f}_{n+1}\) be the predictors produced in the process of running the online learning algorithm. Now consider a randomized predictor that operates as follows. Upon input \(x \in \mathcal{X}\), we choose \(T\) uniformly at random from \(\{1, \ldots, n+1\}\) and then output \(\hat{f}_T(x)\). (In the next section, we shall consider a derandomized predictor based on majority vote.)

Let \((X, Y) \sim P\) for some distribution \(P\) over \(\mathcal{X} \times \mathcal{Y}\), and consider the randomized predictor \(\hat{f}_T\) based on \(\hat{f}_1, \ldots, \hat{f}_{n+1}\). Since our predictor \(\hat{f}_T\) is a randomized function, we shall consider its (zero-one loss) risk \(R(\hat{f}_T)\) to be the probability that \(\hat{f}_T(X) \neq Y\), where \((X, Y)\) and \(T\) are treated as independent random variables.

We study the expected value of \(R(\hat{f}_T)\), where the expectation is taken with respect to the data sequence \((X_1, Y_1), \ldots, (X_n, Y_n) \sim_{iid} P\) (which are independent of both \((X, Y)\) and \(T\)):

\[
E \left[ R(\hat{f}_T) \right] = \sum_{t=1}^{n+1} P(T = t) \cdot E \left[ R(\hat{f}_T) \mid T = t \right]
\]

\[
= \frac{1}{n+1} \sum_{t=1}^{n+1} E \left[ R(\hat{f}_t) \right]
\]

\[
= \frac{1}{n+1} \sum_{t=1}^{n+1} E \left[ 1\{\hat{f}_t(X) \neq Y_t\} \right]
\]

\[\overset{(*)}{=} \frac{1}{n+1} \sum_{t=1}^{n+1} E \left[ 1\{\hat{f}_t(X_t) \neq Y_t\} \right]
\]

\[
= \frac{1}{n+1} E \left[ \sum_{t=1}^{n+1} 1\{\hat{f}_t(X_t) \neq Y_t\} \right]
\]

\[
= \frac{E[M_{n+1}]}{n+1},
\]

where \((X_{n+1}, Y_{n+1}) := (X, Y)\) and \(M_{n+1} := \sum_{t=1}^{n+1} 1\{\hat{f}_t(X_t) \neq Y_t\}\) is the number of mistakes incurred by the online learning algorithm on a sequence of \(n+1\) iid random examples from \(P\). The fourth equality (*) above follows from the fact that \((\hat{f}_t, (X, Y))\) has the same joint probability distribution as \((\hat{f}_t, (X_t, Y_t))\).
Derandomizing $\hat{f}_T$

(In this section, we assume $\mathcal{Y} = \{-1,+1\}$.)

A randomized predictor may be undesirable for many reasons, and hence it is natural to ask if it is possible to derandomize $\hat{f}_T$. A simple alternative to $\hat{f}_T$ is the majority vote predictor, $\hat{f}_{\text{maj}}$, given by

$$\hat{f}_{\text{maj}}(x) := \begin{cases} +1 & \text{if at least } (n+1)/2 \text{ of the } \hat{f}_i(x) \text{ are } +1, \\ -1 & \text{otherwise,} \end{cases} x \in \mathcal{X}.$$  

How does $\hat{f}_{\text{maj}}$ compare to $\hat{f}_T$? (We shall henceforth condition on the data sequence $(X_1,Y_1), \ldots, (X_n,Y_n)$ that determine $\hat{f}_1, \ldots, \hat{f}_{n+1}$, and omit this conditioning in notation until further notice.)  

For a given $x \in \mathcal{X}$, let us examine the conditional (zero-one loss) risk of $\hat{f}_{\text{maj}}$ given $X = x$:

$$\mathbb{P}(\hat{f}_{\text{maj}}(X) \neq Y \mid X = x) = \mathbb{P}(\hat{f}_{\text{maj}}(x) = +1, Y = -1 \mid X = x) + \mathbb{P}(\hat{f}_{\text{maj}}(x) = -1, Y = +1 \mid X = x)$$

$$= 1\{\hat{f}_{\text{maj}}(x) = +1\} \mathbb{P}(Y = -1 \mid X = x) + 1\{\hat{f}_{\text{maj}}(x) = -1\} \mathbb{P}(Y = +1 \mid X = x).$$

Recall that $\hat{f}_{\text{maj}}(x) = +1$ only if $\mathbb{P}(\hat{f}_T(x) = +1) \geq 1/2$, and similarly, $\hat{f}_{\text{maj}}(x) = -1$ only if $\mathbb{P}(\hat{f}_T(x) = -1) \geq 1/2$. Therefore

$$\frac{1}{2} \cdot 1\{\hat{f}_{\text{maj}}(x) = +1\} \leq \mathbb{P}(\hat{f}_T(x) = +1), \quad \frac{1}{2} \cdot 1\{\hat{f}_{\text{maj}}(x) = -1\} \leq \mathbb{P}(\hat{f}_T(x) = -1).$$

This implies

$$1\{\hat{f}_{\text{maj}}(x) = +1\} \mathbb{P}(Y = -1 \mid X = x) \leq 2 \cdot \mathbb{P}(\hat{f}_T(x) = +1) \mathbb{P}(Y = -1 \mid X = x)$$

$$= 2 \cdot \mathbb{P}(\hat{f}_T(x) = +1, Y = -1 \mid X = x),$$

$$1\{\hat{f}_{\text{maj}}(x) = -1\} \mathbb{P}(Y = +1 \mid X = x) \leq 2 \cdot \mathbb{P}(\hat{f}_T(x) = -1) \mathbb{P}(Y = +1 \mid X = x)$$

$$= 2 \cdot \mathbb{P}(\hat{f}_T(x) = -1, Y = +1 \mid X = x),$$

and

$$\mathbb{P}(\hat{f}_{\text{maj}}(X) \neq Y \mid X = x) \leq 2 \cdot \mathbb{P}(\hat{f}_T(x) \neq Y \mid X = x).$$

Since this holds for all $x \in \mathcal{X}$, we conclude that

$$\mathbb{P}(\hat{f}_{\text{maj}}(X) \neq Y) \leq 2 \cdot \mathbb{P}(\hat{f}_T(X) \neq Y),$$

i.e., the risk of the majority vote predictor $\hat{f}_{\text{maj}}$ is at most twice the risk of the randomized predictor $\hat{f}_T$.

Combining with the results of the previous section (now taking expectation over all random variables, including the data sequence):

$$\mathbb{E} \left[ \mathcal{R}(\hat{f}_{\text{maj}}) \right] \leq \frac{2 \cdot \mathbb{E}[M_{n+1}]}{n + 1}.$$