Stat 928: Statistical Learning Theory **Lecture: 20**

The Perceptron for Generalized Linear Models and Single Index Models

Instructor: Sham Kakade

1 Learning Generalized Linear Models

Algorithm 1 GLM-tron

Input: function $u(\cdot)$
$w_1 := 0$;
for $t = 1, 2, \ldots$ **do**
  $\hat{y}_t := u(w_t \cdot x)$;
  $w_{t+1} := w_t + (y_t - \hat{y}_t)x_t$;
end for

To analyze the performance of the algorithm, we show that if we run the algorithm for sufficiently many iterations, one of the predictors h_t obtained must be nearly-optimal, compared to the Bayes-optimal predictor.

Theorem 1.1. *Suppose the sequence* $(x_1, y_1), (x_2, y_2), \ldots$ *satisfy, for all* t*:*

- $\|x_t\|^2 \leq 1$ *and* $y_t \in [0, 1]$
- $u : \mathbb{R} \to [0, 1]$ *is a known non-decreasing* 1*-Lipschitz function*
- *there exists a* w *such that* $y_t = u(w \cdot x_t)$

Then GLM-tron satisfies:

$$\sum_{t=1}^{\infty} (y_t - \hat{y}_t)^2 \leq \|w\|^2$$

The proof is based on the following lemma:

Lemma 1.2. *At iteration* t *in GLM-tron,*

$$\|w_t - w\|^2 - \|w_{t+1} - w\|^2 \geq (y_t - \hat{y}_t)^2$$

Proof. We have

$$\|w_t - w\|^2 - \|w_{t+1} - w\|^2 \ = \ 2(y_t - \hat{y}_t)(w \cdot x_t - w_t \cdot x_t) - \|(y_t - \hat{y}_t)x_t\|^2 . \tag{1}$$

Consider the first term above,

$$\frac{2}{m} \sum_{i=1}^{m} (y_t - \hat{y}_t)(w \cdot x_t - w_t \cdot x_t) \ = \ 2(u(w \cdot x_t) - u(w_t \cdot x_t))(w \cdot x_t - w_t \cdot x_t)$$

Using that u is non-decreasing and 1-Lipschitz, we have:

$$2(u(w \cdot x_t) - u(w_t \cdot x_t))(w \cdot x_t - w_t \cdot x_t) \geq 2(u(w \cdot x_t) - u(w_t \cdot x_t))^2 \; = \; 2(y_t - \hat{y}_t)^2. \tag{2}$$

To justify this step, consider the case where $w \cdot x_t > w_t \cdot x_t$. We then have (using that u is non-decreasing and 1-Lipschitz)

$$0 \leq u(w \cdot x_t) - u(w_t \cdot x_t) \leq |w \cdot x_t - w_t \cdot x_t| = w \cdot x_t - w_t \cdot x_t$$

The case where $w \cdot x_t > w_t \cdot x_t$ is identical.

For the second term in (4), we have

$$\|(y_t - \hat{y}_t)x_t\|^2 \; = \; (y_t - \hat{y}_t) \|x_t\|^2 \; \leq \; (y_t - \hat{y}_t)^2 \tag{3}$$

which completes the proof. □

Hence, we have that:

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \leq \left\|w^1 - w\right\|^2 - \|w_{T+1} - w\|^2 \leq \|w\|^2$$

which completes the proof.

2 Isotonic Regression and the PAV algorithm

The Pool Adjacent Violators (PAV) algorithm is finds the best monotonic one dimensional fit for $(\hat{z}_1, y_1), (\hat{z}_2, y_2), \ldots (\hat{z}_m, y_m)$, where the z_i and y_i's are real. Precisely,

$$\mathsf{PAV}((\hat{z}_1, y_1), (\hat{z}_2, y_2), \ldots (\hat{z}_m, y_m)) = \arg\min_{\text{nondecreasing functions } f} \; \frac{1}{m} \sum_{i=1}^{m} (y_i - f(z_i))^2$$

If u is returned by PAV, then it satisfies the following calibration property for any $y \in \mathbb{R}$

$$\sum_{i \text{ s.t. } u(z_i) = y} (y_i - z) = 0$$

In other words, wherever the function u is constant (say when u is y) then this constant must be the average of all y_i where $u(z_i) = y$). If this were not the case, note that we could slightly shift the function at y without breaking the monotonicity property so that the square error is decreased.

With this observation the PAV algorithm can be implemented in $O(m \log m)$ time. The algorithm first sorts the z_i's. Now the algorithm partitions the data into "pools", where the function value is constant in each pool. Initially, each point belongs to it's own pools. If the function is non-monotonic, then any two pools violating the monotonicity property can be merged (and the function value u is the average of the points within the pool).

3 (Batch) Learning of Single Linear Models

Now suppose that u is not known.

Here PAV is the isotonic regression algorithm (the "Pool Adjacent Violator" algorithm). It finds the best 1-dimensional non-decreasing function (with respect to the square loss).

Algorithm 2 Isotron

Input: data $\langle(x_i, y_i)\rangle_{i=1}^m$.
$w^1 := 0$;
for $t = 1, 2, \ldots$ **do**
  $u_t :=$ PAV $((w_t \cdot x_1, y_1), \ldots, (w_t \cdot x_m, y_m))$
  For all i, set $\hat{y}_{t,i} := u_t(w_t \cdot x_i)$
  $w_{t+1} := w_t + \frac{1}{m}\sum_{i=1}^m (y_i - \hat{y}_{t,i})x_i$
end for

Theorem 3.1. *(Isotron algorithm for unknown u) Define the loss on the dataset as:*

$$\hat{L}(u_t, w_t) = \frac{1}{m}\sum_{i=1}^m (y_i - u_t(w_t \cdot x_i))^2$$

Suppose the dataset data $\langle(x_i, y_i)\rangle_{i=1}^m$ satisfy for all t:

- $\|x_i\|^2 \le 1$ *and* $y_i \in [0, 1]$ *for* $i = 1, 2, \ldots m$.
- $u : \mathbb{R} \to [0, 1]$ *is a non-decreasing* 1*-Lipschitz function.*
- *There exists a w such that $y_i = u(w \cdot x_i)$ for $i = 1, 2, \ldots m$.*

Then Isotron *satisfies:*

$$\sum_{t=1}^{\infty} \hat{L}(u_t, w_t) \le \|w\|^2$$

The following corollary shows how this results in a batch optimization algorithm.

Corollary 3.2. *(Optimization) For any iteration T, we have:*

$$\frac{1}{T}\sum_{t=1}^{\infty} \hat{L}(u_t, w_t) \le \frac{\|w\|^2}{T}$$

So there exists a $t \le T$ such that:

$$\hat{L}(u_t, w_t) \le \frac{\|w\|^2}{T}$$

(and this hypothesis can be found by explicitly computing $\hat{L}(u_t, w_t)$ for each $t \le T$).

The following lemma is useful

Lemma 3.3. *At iteration t in* Isotron*,*

$$\|w_t - w\|^2 - \|w_{t+1} - w\|^2 \ge \hat{L}(u_t, w_t)$$

Proof. Let v be any inverse of u (this v may not be unique and we choose one arbitrarily).

We have

$$\|w_t - w\|^2 - \|w_{t+1} - w\|^2 = \frac{2}{m}\sum_{i=1}^m (y_i - \hat{y}_{t,i})(w \cdot x_i - w_t \cdot x_i) - \left\|\frac{1}{m}\sum_{i=1}^m (y_i - \hat{y}_{t,i})x_i\right\|^2. \quad (4)$$

Consider the first term above,

$$\frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})(w\cdot x_i - w_t\cdot x_i) \;=\; \frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})(w\cdot x_i - v(\hat{y}_{t,i})) + \frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})(v(\hat{y}_{t,i}) - w_t\cdot x_i)$$

By the same argument as in the proof of GLM-tron (using that $u = v^{-1}(\cdot)$ is non-decreasing and 1-Lipschitz), we have that for the first term above:

$$\begin{aligned}\frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})(w\cdot x_i - v(\hat{y}_{t,i})) &\geq \frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})(u(w\cdot x_i) - u(v(\hat{y}_{t,i})))\\ &= \frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})^2\\ &= 2\hat{L}(u_t, w_t)\end{aligned}$$

We also have that:

$$\frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})v(\hat{y}_{t,i}) = 0 \tag{5}$$

and

$$\frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})w_t\cdot x_i \leq 0 \tag{6}$$

Equation 5 follows from the calibration property. To see this, consider those i for which $\hat{y}_{t,i} = y$ (for some arbitrary y). The sum over these i is 0. Hence, the sum over all i is 0. For Equation 6, recall that $u_t(\cdot)$ is the output of the isotonic regression, e.g. $u_t = \mathsf{PAV}\left((w_t\cdot x_1, y_1), \ldots, (w_t\cdot x_m, y_m)\right)$. Note that $u_t(\cdot) + \alpha \mathrm{I}(\cdot)$ is also an increasing function when $\alpha > 0$ and $\mathrm{I}(\cdot)$ is the identity function. Equation 6 is just the first derivative condition that for $\alpha > 0$ — note that $u_t(\cdot) + \alpha\mathrm{I}(\cdot)$ (for $\alpha > 0$) does not have lower square loss than $u_t(\cdot)$. In other words, if Equation 6 did not hold, then note that this would imply that for a sufficiently small $\alpha > 0$, the function $u(\cdot) + \alpha\mathrm{I}(\cdot)$ would be a better monotonic function for the data $((w_t\cdot x_1, y_1), \ldots, (w_t\cdot x_m, y_m))$, which violates the optimality of PAV.

For the second term in (4), Jensen's inequality implies

$$\left\|\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})x_i\right\|^2 \;\leq\; \frac{1}{m}\sum_{i=1}^{m}\|(y_i-\hat{y}_{t,i})x_i\|^2 \;=\; \frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_{t,i})^2\,\|x_i\|^2 \;\leq\; \hat{L}(u_t, w_t) \tag{7}$$

which completes the proof. □

For the proof of the theorem, we have that (for all T):

$$\sum_{t}^{T}\hat{L}(u_t, w_t) \leq \left\|w^1 - w\right\|^2 - \|w_{T+1} - w\|^2 \leq \|w\|^2$$

which completes the proof.