Statistical learning (for regression/classification)

Data (x, y) drawn iid from unknown distribution D
features ↑ ↑ labels

Loss function $\ell(\hat{y}, y) \geq 0$ e.g. $(\hat{y}-y)^2$

Predictor $h : \mathbb{X} \to \mathbb{Y}$

Statistical risk $\ell_D(h) = \mathbb{E}\left[\ell(h(X), Y)\right]$

Training set S : pairs (X, Y) drawn iid from D

Learning algorithm A maps S to h

$\mathcal{H}_A$: ~~set~~ set of all predictors output by A (e.g., linear predictors)

In stat learning we care about variance error

$$\underbrace{\ell_D(A(S))}_{\text{random var.}} - \inf_{h \in \mathcal{H}_A} \ell_D(h)$$

Online learning asks: can we study machine learning without using statistical assumptions?

Online learning

- Data points arrive one by one in a stream
- No stat assumptions on the stream

Motivations:

- Sensor data, user interaction data, financial data
- large datasets can only be accessed sequentially

Online learning is incremental

Start from initial predictor $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$ (stream)

- observe next data point (x_t, y_t) in stream
- test current predictor h_t and measure loss $\ell(h_t(x_t), y_t)$
- update $h_t \to h_{t+1} \in \mathcal{H}$

Note: update is local as it depends on $h_t, (x_t, y_t)$

Notation: $\ell(h_t(x_t), y_t) \to \ell_t(h_t) \in [0,1]$

$h_1, h_2, \dots$ stream of predictors generated

Sequential risk of $h_1, h_2, \dots$ $\sum_{t=1}^{T} \ell_t(h_t)$

Regret (cfr. variance error):

$$R_T = \sum_{t=1}^{T} \ell_t(h_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(h)$$

Since $\ell_t \in [0,1]$ $R_T = O(T)$

Learning takes place when $R_T = o(T)$

Learning w/ expert advice: a notable special case

K actions (e.g. stocks), $\ell_t(i) \in [0,1]$ is loss of action i at time t

We consider randomized predictors

Model at time t is distribution p_t over $\underbrace{\{1, \dots K\}}_{\text{actions}}$

Regret $R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots K} \sum_{t=1}^{T} \ell_t(i)$

For $t = 1, 2, \dots$

- Draw $I_t \sim p_t$ and pay $\ell_t(I_t)$
- Observe ℓ_t and update $p_t \to p_{t+1}$

$L_{t-1}(i) = \sum_{s=1}^{t-1} \ell_s(i)$ total loss of i up to time $t-1$

Hedge algorithm: $p_t(i) \propto \exp\left(-\eta L_{t-1}(i)\right)$

$\eta > 0$

Regret analysis

$W_t(i) = e^{-\eta L_{t-1}(i)}$, $W_1(i) = 1$ as $L_0(i) = 0$

$\overline{W}_t = \sum_j W_t(j) \qquad p_t(i) = W_t(i)/\overline{W}_t$

<u>Note</u>: $W_{t+1}(i) = W_t(i)\, e^{-\eta \ell_t(i)}$ update rule

$$\ln \frac{\overline{W}_{T+1}}{\overline{W}_1} = \sum_{t=1}^{T} \ln \frac{\overline{W}_{t+1}}{\overline{W}_t}$$

$$\ln \frac{\overline{W}_{t+1}}{\overline{W}_t} = \ln \sum_i \frac{W_{t+1}(i)}{\overline{W}_t} = \ln \sum_i \overbrace{\boxed{\frac{W_t(i)}{\overline{W}_t}}}^{p_t(i)} e^{-\eta \ell_t(i)}$$

$$\leq \ln \sum_i p_t(i)\left(1 - \eta \ell_t(i) + \frac{\eta^2}{2} \ell_t(i)^2\right) \qquad e^{-x} \leq 1 - x + \frac{x^2}{2} \ (x>0)$$

$$= \ln\left(1 - \eta \sum_i p_t(i)\ell_t(i) + \frac{\eta^2}{2} \sum_i p_t(i)\, \ell_t(i)^2\right)$$

$$\leq -\eta \underbrace{\sum_i p_t(i)\ell_t(i)}_{\mathbb{E}[\ell_t(I_t)]} + \frac{\eta^2}{2} \sum_i p_t(i)\, \ell_t(i)^2 \qquad \ln(1-x) \leq -x$$

Summing over $t = 1, \dots T$:

$$① \quad \ln \frac{W_{T+1}}{W_1} \leq -\eta\, \mathbb{E}\Big[\sum_t \ell_t(I_t)\Big] + \frac{\eta^2}{2} \sum_t \sum_i \ell_t(i)^2 p_t(i)$$

$$K_T^* = \underset{i=1,\dots T}{\mathrm{argmin}} \sum_{t=1}^{T} \ell_t(i) \quad \text{best action in first } T \text{ steps}$$

$$② \quad \ln \frac{W_{T+1}}{W_1} = \ln \sum_i e^{-\eta L_T(i)} - \ln K \geq \ln e^{-\eta L_T(K_T^*)} - \ln K$$

$$= -\eta L_T(K_T^*) - \ln K$$

Divide ① and ② by $\eta > 0$ and rearrange:

$$R_T \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_t \sum_i \underbrace{\ell_t(i)^2}_{\leq 1} \underbrace{p_t(i)}_{\sum_i p_t(i) = 1}$$

Hence $R_T \leq \dfrac{\ln K}{\eta} + \dfrac{\eta}{2} T$

choosing $\eta \simeq \sqrt{\dfrac{\ln K}{T}}$ gives $R_T \leq \sqrt{T \ln K}$

* Not improvable! (except for constants)

Dynamic tuning $\eta_t \simeq \sqrt{\dfrac{\ln K}{t}}$ gives

$R_T \leq \sqrt{T \ln K}$ simultaneously for all T