Summer school
Online Learning
28.06.2015-02.07.2015
Copenhagen University

Participants

Participants & Lecturer

- Around 60 participants
- Lecturer
  - Shai Shalev-Shwartz, Hebrew University
  - Peter Auer, Leoben University
  - Nicola Cesa-Bianchi, Miland University
  - Csaba Szepesvari, Alberta University
  - Yevgeny Seldin, Copenhagen University





Topics

- Basics of online learning
- Online convex optimization
- Bandits (stochastic, adversarial….)
- Online reinforcement learning
- Space of online learning problems

- Theory only, proof of bounds

General online learning problem

- for $t = 1,2,\ldots T$
  - receive question $x_t \in X$
  - predict $p_t \in D$
  - receive true answer $y_t \in Y$
  - suffer loss $l(p_t, y_t)$
  - (update model)

- Given a fixed hypothesis class H(e.g. halfspaces $h_t = \text{sign}(\langle w_t | x_t \rangle)$, $\forall t$ find $h_t \in H, h_t: X \to Y$ s.t. $\sum_{t=1}^{T} l(p_t, y_t)$ is minimized $(p_t = h_t(x_t))$.

- Loss functions
  - $l(p_t, y_t) = |p_t - y_t|$ "0 − 1 loss" or "absolut loss"
  - $l(p_t, y_t) = (p_t - y_t)^2$ "quadratic loss"

2 alternative restrictions/goals

- No assumption about sequence (determenistic, stochastic, adversarial)

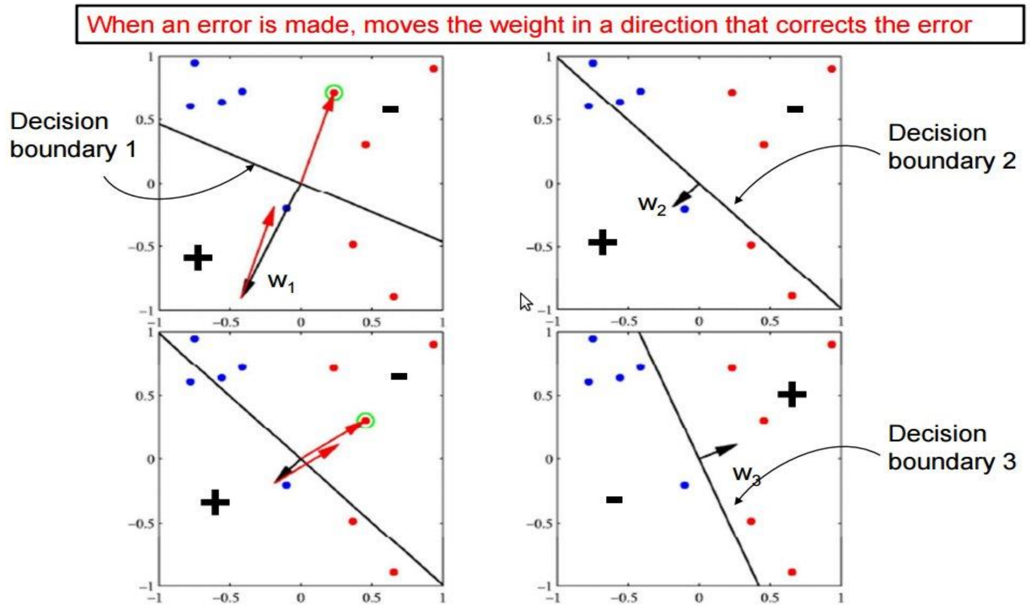
- Realizability
  - $\exists\, \hat{h}: X \rightarrow Y \text{ s.t. } y_t = \hat{h}(x_t) \forall t, \hat{h} \in H$
  - Goal : Find algorithm with minimal mistake bound (sublinear with T)

- No realizability
  - Goal: Find algorithm with minimal regret $R(\hat{h})$ compared to the best fixed predictor $\hat{h} \in H$
  - $R(\hat{h})_T = \sum_{t=1}^{T} l(p_t, y_t) - \sum_{t=1}^{T} l(\hat{h}(p_t), y_t)$

- Given an algorithm-> find & proof the corresponding mistake-bound/regret-bound

Perceptron for realizability case

- Seperate data points for binary classification
- $\mathrm{Y} = \{-1,1\}$, hypothesis class H = all halfspaces in $\mathbb{R}^n$
- $\mathrm{l}(\mathrm{w_t}) = \max(1 - \mathrm{y_t}{*}\langle \mathrm{w_t} | \mathrm{x_t} \rangle, 0)$, similar to "hinge − loss"



Red points belong to the positive class, blue points belong to the negative class

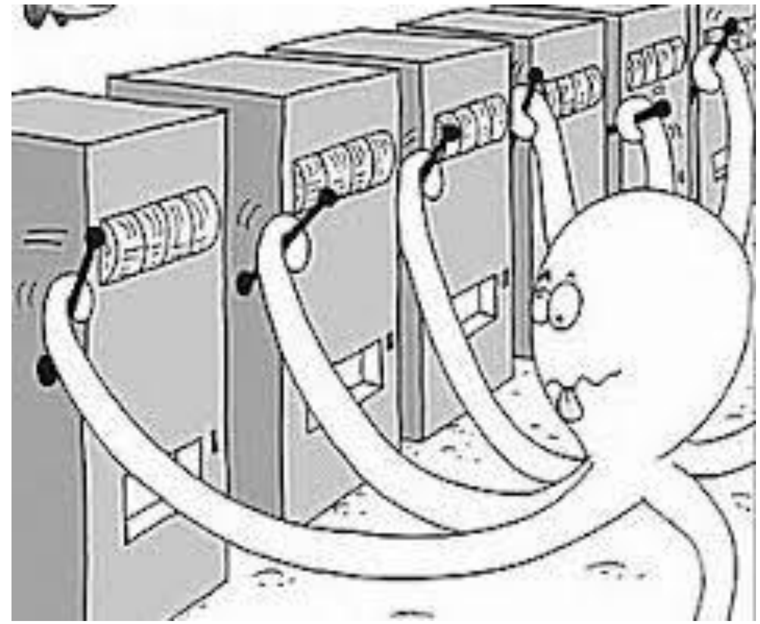
- initialize: $\mathrm{w_1} = 0$
- for $\mathrm{t} = 1,2,\ldots,\mathrm{T}$
  - receive $\mathrm{x_t}$
  - predict $\mathrm{p_t} = \mathrm{sign}(\langle \mathrm{w_t} | \mathrm{x_t} \rangle)$
  - if $y_t\langle \mathrm{w_t} | \mathrm{x_t} \rangle \le 0$
    - $\mathrm{w_{t+1}} = w_t + \mathrm{y_t x_t}$
  - else $\mathrm{w_{t+1}} = w_t + \mathrm{y_t x_t}$

- Suppose that $||x_t|| \le D \; and \; \exists \; \widehat{w}(\big||\widehat{w}|\big|_2 = 1) \; s.t. \, y_t\langle \widehat{w} | \mathrm{x_t} \rangle \ge \gamma \; \forall t$

$$\rightarrow M \le \left(\frac{D}{\gamma}\right)^2$$

Multi-armed Bandits

- for t = 1,2,...T
  - play action $y_t \in Y = \{1, .. K\}$
  - receive reward $r_t(y_t) = [0,1]$



- Only reward of played action is seen
- Goal: maximize reward!
  - minimize regret $R = T\hat{\mu} - \sum_{t=1}^{T} r_t$
- Exploitation vs. Exploration
- Old, but very popular nowadays....Why?!

  → Google uses it

Multi-armed Bandits

- Various versions
  - Stochastic stationary/non-stationary
  - Adversarial, Contextual
  - Graph-based

- Applications
  - Web-searches (contextual, max adv. income of Google, Bing etc.)
  - Clinical trials (minimize patient losses)
  - Adaptive routing (minimize delays)

Stochastic, stationary Bandits

- Rewards for actions are generated by stationary distributions
- $R(T) = T\hat{\mu} - \sum_{j=1}^{K} \mu_j * T_j, T_j$ is the number action j was played
- Estimate μ_k accurately to minimize the regret

Upper Confidence Bound (UCB) algorithm

- Calculate confidence bounds for each action
- Chernoff-Hoeffding bound:
  - Let $X_1, X_2, \ldots X_K$ independent random variables in the range [0,1] with $\mu = \mathrm{Exp}(X) = \frac{1}{K}\sum_{j=1}^{K} \mu_j \rightarrow P\left(\frac{1}{T}\sum_{t=1}^{T} X_i \geq \mu + a\right) \leq e^{-2a^2 T}$
- $a = \sqrt{\frac{2\log(T)}{T_j}} \rightarrow P\left(\frac{1}{T}\sum_{t=1}^{T} X_i \geq \mu + a\right) \leq T^{-4}$, converges quickly to 0
- Choose the action with highest upper confidence bound

$$\max \mu_j + \sqrt{\frac{2\log(T)}{T_j}}$$

- Balances exploitation vs. exploration

- $R_{UCB} \leq \sum_{j=1}^{K} \frac{2\log(T)}{\Delta_j} + (1 + \frac{\pi^2}{3})\,\Delta_j, \qquad \Delta_j = \hat{\mu} - \mu_j$

Further readings

- Got interested ?
  - Shai Shalev-Shwartz “Online Learning and Online Convex Optimization”
  - Sebastien Bubeck, Nicolò Cesa-Bianchi „Regret Analysis of Stochastic andNon-stochastic Multi-armedBandit Problems“
  - Nicolò Cesa-Bianchi “Prediction, learning, and games”