COS324: Introduction to Machine Learning

Lecture 3: online learning part II

Prof. Elad Hazan & Prof. Yoram Singer

Recap + today

- last lecture:
  1. online decision making
  2. our first (serious) learning algorithm: weighted majority
- today: the power of randomness in learning
  1. randomization in decision making
  2. the Kelly criterion

Reminder: online learning

- Initialize $\mathbf{w}^1$; $\mathcal{L}^1 = 0$
- For $t = 1, 2, \dots, T, \dots$
  1. Predict $\hat{y}^t$ using $\mathbf{w}^t$
  2. Observe true outcome y^t
  3. Endure loss: $\ell^t = \ell(y^t, \hat{y}^t)$; $\mathcal{L}^{t+1} = \mathcal{L}^t + \ell^t$
  4. Update $\mathbf{w}^{t+1} := F(\mathbf{w}^t, \mathbf{x}^t, y^t)$

Reminder: Weighted Majority Algorithm

- Initialize $\mathbf{w}^1 = \mathbf{1}$; $\mathcal{L}^1 = 0$
- For $t = 1, 2, \ldots, T, \ldots$

  1. Observe predictions $\mathbf{x}^t \in \{-1, +1\}^n$
  2. Predict $\hat{y}^t := \text{sign}(\mathbf{w}^t \cdot \mathbf{x}^t)$
  3. Observe true outcome y^t
  4. Endure loss: $\ell^t = \mathbf{1}[y^t \neq \hat{y}^t]$; $\mathcal{L}^{t+1} = \mathcal{L}^t + \ell^t$
  5. Update:

$$w_j^{t+1} = \begin{cases} w_j^t & x_j^t = y^t \\ (1-\eta) w_j^t & x_j^t \neq y^t \end{cases}$$

Bag Of Words (BOW) model

- Pre-defined dictionary of n tokens (words, html, arch-codes)

kale	1
plate	2
kohlrabi	3
ate	4
fork	5

- Represent a document as a vector $\mathbf{x} \in \{-1, +1\}^n$ s.t. $x_j = +1$ iff token j appears in document
- Tokens not in the dictionary are ignored
- Examples:

$$\text{"The kohlrabi ate kale on a plate"} \mapsto (+1, +1, +1, +1, -1)$$

$$\text{"A monkey ate a banana with a fork"} \mapsto (-1, -1, -1, +1, +1)$$

BOW + WM $\Rightarrow$ Text Classifier

- Each dictionary word is an expert
- Initialize weight of experts $\mathbf{w}^1 = \mathbf{1}$
- For $t = 1, \ldots, m$: // m is #document
    - Convert document t to a vector $\mathbf{x}^t \in \{-1, +1\}^n$
    - Update weights using WM with provided tagging y^t: $\mathbf{w}^t \rightsquigarrow \mathbf{w}^{t+1}$
- Output $\mathbf{w}^{m+1}$

Wait, but what if $\nexists$ *single* accurate expert ?
Do we obtain a good classifier? **Yes!**

(Future) Refinement

- In many applications the vocabulary size n is much larger than length of each individual document
- Therefore $\mathbf{x}^i$ consists mostly of -1's and few $+1$'s
- Most of the contribution to the weighted majority is due to words that do not appear in the document
- We can represent a document as a vector in $\{0, 1\}^n$
  - If word j appears in document then $x_j = 1$ o.w. $x_j = 0$
  - Algorithmic advantage – represent $\mathbf{x}$ as a list of indices
- However, $\mathbf{w} \cdot \mathbf{x} > 0$ since all weights and inputs are non-negative
- Introduce an *bias term* (indexed 0) which is always -1:
$$\mathbf{x} \mapsto (-1, \mathbf{x})$$

To be continued...

Reminder: guarantee

$\mathcal{L}_i^T$ number of mistakes made by expert i during $t = 1, ..., T$
$\mathcal{L}^T$ number of mistakes WM made during during $t = 1, ..., T$

Theorem: For every sequence $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^T, y^T)$ the number of mistakes of WM is at most,

$$\forall i \in [n] \ : \quad \mathcal{L}^T \ \leq \ 2(1+\eta)\mathcal{L}_i^T \ + \ \frac{2\log(\mathrm{n})}{\eta}$$

Theorem 2: **any** deterministic decision making algorithm has

$$\mathcal{L}^T \ \geq \ 2\min_i 2\mathcal{L}_i^T$$

But can we still do better??

Randomized Weighted Majority

- Little and Warmuth derived randomized version of WM (RWM)
- RWM replaces the deterministic weighted majority rule with a randomized prediction:
  1. Define a distribution over experts

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^{n} w_j^t}$$

  2. Pick an expert i^t at random according to $\mathbf{p}^t$
- How is this random choice implemented on a computer?

Randomized Weighted Majority

- Initialize $\mathbf{w}^1 = \mathbf{1}$; $\mathcal{L}^1 = 0$
- For $t = 1, 2, \dots, T, \dots$

1. Observe predictions $\mathbf{x}^t \in \{-1, +1\}^n$
2. Form distribution $p_i^t = \frac{w_i^t}{\sum_{j=1}^{n} w_j^t}$
3. Pick an index e with probability p_e^t and predict $\hat{y}^t := x_e^t$
4. Observe true outcome y^t
5. Endure loss: $\ell^t = \mathbf{1}\left[y^t \neq \hat{y}^t\right]$; $\mathcal{L}^{t+1} = \mathcal{L}^t + \ell^t$
6. Update:

$$w_j^{t+1} = \begin{cases} w_j^t & x_j^t = y^t \\ (1-\eta) w_j^t & x_j^t \neq y^t \end{cases}$$

Randomized Weighted Majority

- The expected number of mistakes of RWM is bounded above,

$$\mathbb{E}[\mathcal{L}^T] \leq (1+\eta)\mathcal{L}_{i^*}^T + \frac{\log(n)}{\eta}$$

- This bound is tight – any randomized prediction algorithm in the experts setting makes at least,

$$(1+\eta)\mathcal{L}_{i^*}^T + \frac{\log(n)}{\eta}$$

mistakes for some $\eta \in (0, \frac{1}{2})$

Proof

- Let i^* be the best expert in hindsight (the one who made the least number of mistakes)
- Let $\Phi^t = \sum_{i=1}^{n} w_i^t$
- Let m_i^t be 1 if expert i made a mistake on round t and 0 o.w.
- Notice that $\mathcal{L}_i^T = \sum_{t=1}^{T} m_i^t$
- Expected number of mistakes by RWM at time t is

$$p^t \cdot m^t = \sum_{i=1}^{n} p_i^t m_i^t$$

and overall expected #mistakes from 1 thru T is $\sum_{t=1}^{T} p^t \cdot m^t$

Observation I

$$
\begin{aligned}
\Phi^T &= \sum_{i=1}^{n} w_i^T \\
&\geq w_{i^*}^T \\
&= w_{i^*}^0 \times (1-\eta)^{\mathcal{L}_{i^*}^T} \\
&= (1-\eta)^{\mathcal{L}_{i^*}^T}
\end{aligned}
$$

Observation II

$$\Phi^T \leq \Phi^0 e^{-\eta \sum_{t=1}^{T} p^t \cdot m^t}$$

Proof outline:

- Expand Φ^{t+1}

$$\Phi^{t+1} = \sum_{i=1}^{n} w_i^{t+1} \; = \; \sum_{i=1}^{n} w_i^t (1 - \eta m_i^t)$$

- Since $p_i^t = \frac{w_i^t}{\Phi^t} \Rightarrow w_i^t = \Phi^t p_i^t$

$$\Phi^{t+1} = \Phi^t - \eta \sum_i \Phi^t p_i^t m_i^t \; = \; \Phi^t (1 - \eta p^t \cdot m^t)$$

- Use $1 - a \leq e^{-a}$

$$\Phi^{t+1} \leq \Phi^t e^{-\eta p^t \cdot m^t}$$

- Use induction on t to get observation

Proof (cont.)

- Combining both observations:

$$(1-\eta)^{\mathcal{L}_{i^*}^T} \leq \Phi^T \leq \Phi^0 e^{-\eta \mathbb{E}[\mathcal{L}^T]}$$

- Taking the logarithm:

$$-\eta \mathbb{E}[\mathcal{L}^T] + \log(n) \geq \mathcal{L}_{i^*}^T \log(1-\eta)$$

- From the Taylor approximation, for $\eta < \frac{1}{2}$:

$$-\eta - \eta^2 \leq \log(1-\eta) \leq -\eta$$

- Plugging that back in:

$$-\eta \mathbb{E}[\mathcal{L}^T] + \log(n) \geq \mathcal{L}_{i^*}^T(-\eta - \eta^2)$$

- Shifting sides and multiplying by $\frac{1}{\eta}$:

$$\mathbb{E}[\mathcal{L}^T] \leq \frac{\log(n)}{\eta} + (1+\eta)\mathcal{L}_{i^*}^T$$

Randomized Weighted Majority

- The expected number of mistakes of RWM is bounded above:

$$\mathbb{E}[\mathcal{L}^T] \leq (1+\eta)\mathcal{L}_{i^*}^T + \frac{\log(n)}{\eta}$$

- How good is this bound?

Kelly criterion

Kelly criterion

- Horse race - how to bet on a favorable horse? (prior information tilt the odds in your favor)
- Two possible outcomes, both happen w.p. $\frac{1}{2}$:
  - Loose everything
  - Make $3\times$ on your bet
- Bet of \$1. Outcome after race:

$$\text{reward} = \begin{cases} \$0, & w.p. \ \frac{1}{2} \\ \$3, & w.p. \ \frac{1}{2} \end{cases}$$

- Given \$100, how much would you bet?

Kelly criterion

- Repeated investing: wealth increases by factor of b with probability p such that $pb > 1$
- Given that we have 100 rounds of investing, what fraction of wealth to iteratively invest?
- μ^t = wealth at time t ; $\rho^t = \frac{\mu^t}{\mu^{t-1}}$
- $f \in [0, 1]$ fraction of wealth to bet on
- Expectation (one round):

$$
\begin{aligned}
\mathbb{E}[\rho^t] &= (1-p)(1-f) + p\,[(1-f) + f\,b] \\
&= 1 + f(pb - 1) > 1
\end{aligned}
$$

- Maximized at $f = 1$, why?

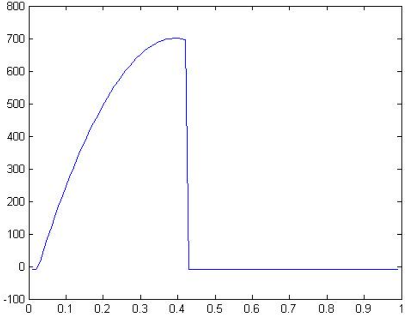
Kelly criterion

- After 100 rounds of investing...
- Expectation:

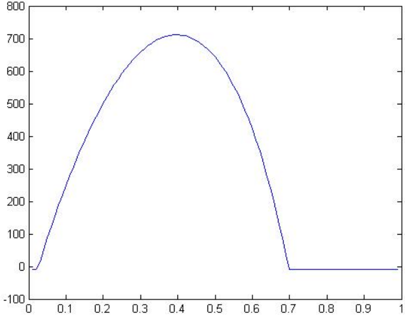
$$\begin{aligned}
\mathbb{E}[\mu^{100}] &= \mu^1 \, \mathbb{E}[\prod_{t=1}^{T} \rho^t] \\
&= \mu^1 \prod_{t=1}^{100} \mathbb{E}[\rho^t] && \text{independence} \\
&= \mu^1 \, (1 + f(bp - 1))^{100}
\end{aligned}$$

- So, how much would you bet?

Kelly criterion - simulation



Kelly criterion - simulation



Kelly criterion

- The Kelly Criterion – Maximize

$$\mathbb{E}[\log(\rho^t)]$$

- Results in:

$$f^* = \frac{pb-1}{b-1}$$

- Theorem: betting f^* results in more wealth than any other fractional-betting method with probability one, as number of rounds $\mapsto \infty$!
- To be continued later in the course...

Summary

- The power of randomization in learning
- Randomized weighted majority
- Use in text classification
- Expectation vs. high probability, Kelly criterion
- Next week: statistical and computational learning theory.