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 $w^1$; $\mathcal{L}^1 = 0$

- For $t = 1, 2, \ldots, T, \ldots$
  1. Predict $\hat{y}^t$ using $w^t$
  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 $w^{t+1} := F(w^t, x^t, y^t)$
Reminder: Weighted Majority Algorithm

- Initialize $w^1 = 1$ ; $L^1 = 0$

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>kale</td>
<td>1</td>
</tr>
<tr>
<td>plate</td>
<td>2</td>
</tr>
<tr>
<td>kohlrabi</td>
<td>3</td>
</tr>
<tr>
<td>ate</td>
<td>4</td>
</tr>
<tr>
<td>fork</td>
<td>5</td>
</tr>
</tbody>
</table>

- 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:

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

  "A monkey ate a banana with a fork" $\mapsto (-1, -1, -1, +1, +1)$
BOW + WM ⇒ Text Classifier

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

Wait, but what if $\not\exists$ single accurate expert?
Do we obtain a good classifier? Yes!
In many applications the vocabulary size $n$ is much larger than length of each individual document.

Therefore $x^i$ 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 $x$ as a list of indices

However, $w \cdot x > 0$ since all weights and inputs are non-negative.

Introduce an *bias term* (indexed 0) which is always $-1$: $x \mapsto (-1, x)$

To be continued...
Reminder: guarantee

\[ \mathcal{L}_i^T \] number of mistakes made by expert \( i \) during \( t = 1, \ldots, T \)
\[ \mathcal{L}_w^T \] number of mistakes WM made during during \( t = 1, \ldots, T \)

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

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

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

\[
\mathcal{L}_w^T \geq 2 \min_i 2\mathcal{L}_i^T
\]

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 \( p^t \)
- How is this random choice implemented on a computer?
Randomized Weighted Majority

- Initialize $w^1 = 1$ ; $L^1 = 0$
- For $t = 1, 2, \ldots, T, \ldots$
  1. Observe predictions $x^t \in \{-1, +1\}^n$
  2. Form distribution $p^t_i = \frac{w^t_i}{\sum_{j=1}^{n} w^t_j}$
  3. Pick an index $e$ with probability $p^t_e$ and predict $\hat{y}^t := x^t_e$
  4. Observe true outcome $y^t$
  5. Endure loss: $\ell^t = 1[y^t \neq \hat{y}^t]$ ; $L^{t+1} = L^t + \ell^t$
  6. Update:
    $$w^{t+1}_j = \begin{cases} w^t_j & x^t_j = y^t \\ (1 - \eta)w^t_j & x^t_j \neq y^t \end{cases}$$
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 1

\[
\Phi^T = \sum_{i=1}^{n} w_i^T \geq w_{i^*}^T = w_{i^*}^0 \times (1 - \eta) L_{i^*}^T = (1 - \eta) L_{i^*}^T
\]
Observation II

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

Proof outline:

- Expand \( \Phi^{t+1} \)

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

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

- Taking the logarithm:
  \[-\eta \mathbb{E}[\mathcal{L}^T] + \log(n) \geq L_{i*}^T \log(1 - \eta)\]

- From the Taylor approximation, for \(\eta < \frac{1}{2}\):
  \[-\eta - \eta^2 \leq \log(1 - \eta) \leq -\eta\]

- Plugging that back in:
  \[-\eta \mathbb{E}[\mathcal{L}^T] + \log(n) \geq L_{i*}^T (-\eta - \eta^2)\]

- Shifting sides and multiplying by \(\frac{1}{\eta}\):
  \[\mathbb{E}[\mathcal{L}^T] \leq \frac{\log(n)}{\eta} + (1 + \eta)L_{i*}^T\]
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:

\[
\text{reward} = \begin{cases} 
0, & \text{w.p. } \frac{1}{2} \\
3, & \text{w.p. } \frac{1}{2}
\end{cases}
\]
- 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?
- $\mu^t = \text{wealth at time } t$ ; $\rho^t = \frac{\mu^t}{\mu^{t=1}}$
- $f \in [0, 1]$ fraction of wealth to bet on
- Expectation (one round):

$$\mathbb{E}[\rho^t] = (1 - p)(1 - f) + p [(1 - f) + fb]$$
$$= 1 + f(pb - 1) > 1$$

- Maximized at $f = 1$, why?
Kelly criterion

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

\[ \mathbb{E}[\mu^{100}] = \mu^1 \mathbb{E}\left[\prod_{t=1}^{T} \rho^t\right] \]

\[ = \mu^1 \prod_{t=1}^{100} \mathbb{E}[\rho^t] \quad \text{independence} \]

\[ = \mu^1 (1 + f(bp - 1))^{100} \]

- 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 \( \to \infty \)!

• 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.