Generalization Bounds

UNIVERSITY of WASHINGTON

Realizable case

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. Assume there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) \leq \frac{\log(|\mathcal{H}|/\delta)}{n}$$

where $(X, Y) \sim \nu$.

Realizable case - Proof

Realizable case - Proof

Realizable case

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \dots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. Assume there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) \leq \frac{\log(|\mathcal{H}|/\delta)}{n}$$

where $(X, Y) \sim \nu$.

Corollary Under the conditions of the theorem (i.e., there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$, $(x_i, y_i) \overset{iid}{\sim} \nu$, and $\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$) we have $\mathbb{E}[R(\widehat{h})] \leq \int_{\epsilon=0}^{d} \mathbb{P}(R(\widehat{h}) \geq \epsilon) \leq \frac{2\log(|\mathcal{H}|)}{n}$

Agnostic (Non-realizable) case

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) - R(h_*) \leq \sqrt{\frac{2\log(|\mathcal{H}|/\delta)}{n}}.$$

$$h_* = \arg\min_{h \in \mathcal{H}} R(h)$$

$$R(\hat{h}) - R(h_*) = R(\hat{h}) - \hat{R}_n(\hat{h}) + \underbrace{\hat{R}_n(\hat{h}) - \hat{R}_n(h_*)}_{\leq 0} + \hat{R}_n(h_*) - R(h_*)$$

$$\leq \frac{1}{n}\sum_{i=1}^n \Big(\underbrace{\mathbb{P}(\hat{h}(X) \neq Y)}_{\mu} - \underbrace{\mathbb{1}\{\hat{h}(x_i) \neq y_i\}}_{Z_i}\Big) + \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{1}\{h_*(x_i) \neq y_i\} - \mathbb{P}(h_*(X) \neq Y)\Big)$$

$$\mathbb{P}\left(\bigcup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{P}(h(X) \neq Y) - \mathbb{1}\{h(x_i) \neq y_i\} \right) > \varepsilon \right\} \right)$$

$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{P}(h(X) \neq Y) - \mathbb{1}\{h(x_i) \neq y_i\} \right) > \varepsilon \right)$$

$$\leq \delta |\mathcal{H}| \leq \delta'$$

$$\varepsilon = \sqrt{\frac{\log(1/\delta)}{2n}} \quad \Rightarrow \quad \varepsilon' = \sqrt{\frac{\log(|\mathcal{H}|/\delta)}{2n}}$$

Agnostic (Non-realizable) case - Proof

~~**Lemma**~~ Corollary **(Hoeffding's inequality)**: Let $Z_1, \dots, Z_n \overset{iid}{\sim} \nu$ where $\mathbb{E}[Z_i] = \mu$ and $Z_i \in [a, b]$ almost surely. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \geq \mu + \epsilon\right) \leq \exp\left(\frac{2n\epsilon^2}{|b-a|^2}\right).$$

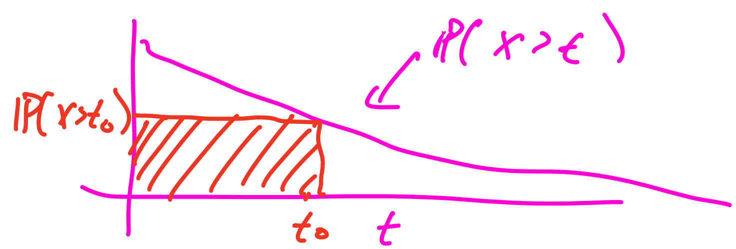
$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt$$

Under above conditions

$$\mathbb{E}[\exp(\lambda(Z-\mu))] \leq \exp(\lambda^2 (b-a)^2/8)$$

For any positive R.V. X (i.e. $X \geq 0$ a.s.)

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}$$



$$\mathbb{P}\left(\frac{1}{n}\sum_i Z_i > \mu + \varepsilon\right) \overset{\lambda > 0}{=} \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^{n}(Z_i - \mu)\right) > \exp(\lambda \varepsilon n)\right)$$

$$\leq e^{-\lambda \varepsilon n}\, \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n}(Z_i - \mu)\right)\right]$$

$$= e^{-\lambda \varepsilon n}\, \mathbb{E}\left[\prod_{i=1}^{n} \exp(\lambda(Z_i - \mu))\right]$$

$$= e^{-\lambda \varepsilon n} \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda(Z_i - \mu))\right]$$

$$= e^{-\lambda \varepsilon n}\, \mathbb{E}\left[\exp(\lambda(Z_1 - \mu))\right]^n$$

$$= e^{-\lambda \varepsilon n + \lambda^2 (b-a)/8}$$

Optimize λ

$$= e^{-2n\varepsilon^2/(b-a)^2}$$

Agnostic (Non-realizable) case - Proof

Agnostic (Non-realizable) case

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) - R(h_*) \leq \sqrt{\frac{2\log(|\mathcal{H}|/\delta)}{n}}.$$

Corollary Under the conditions of the theorem (i.e., $(x_i, y_i) \overset{iid}{\sim} \nu$, and $\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$) and $|\mathcal{H}| \geq n$, we have $\mathbb{E}[R(\widehat{h})] - R(h_*) \leq \sqrt{\frac{8\log(|\mathcal{H}|)}{n}}$

Agnostic (Non-realizable) case - Interpolation

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. If $\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) - R(h_*) \leq \sqrt{\frac{2R(h_*)\log(2|\mathcal{H}|/\delta)}{n}} + \frac{\log(2|\mathcal{H}|/\delta)}{n}.$$

Proof: Use Bernstein's inequality instead of Hoeffding. ■

Infinite classes

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \stackrel{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) - R(h_*) \leq \sqrt{\frac{2R(h_*) \log(2|\mathcal{H}|/\delta)}{n}} + \frac{\log(2|\mathcal{H}|/\delta)}{n}.$$

What if $|\mathcal{H}|$ is *infinite* such as the space of all hyperplane classifers?

Infinite classes

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) - R(h_*) \leq \sqrt{\frac{2R(h_*) \log(2|\mathcal{H}|/\delta)}{n}} + \frac{\log(2|\mathcal{H}|/\delta)}{n}.$$

What if $|\mathcal{H}|$ is *infinite* such as the space of all hyperplane classifers?

Lots of tools to address this:
- minimum description length
- VC-dimension and Rademacher complexity
- Covering number / log-entropy bounds

Online Learning

UNIVERSITY of WASHINGTON

Realizable case

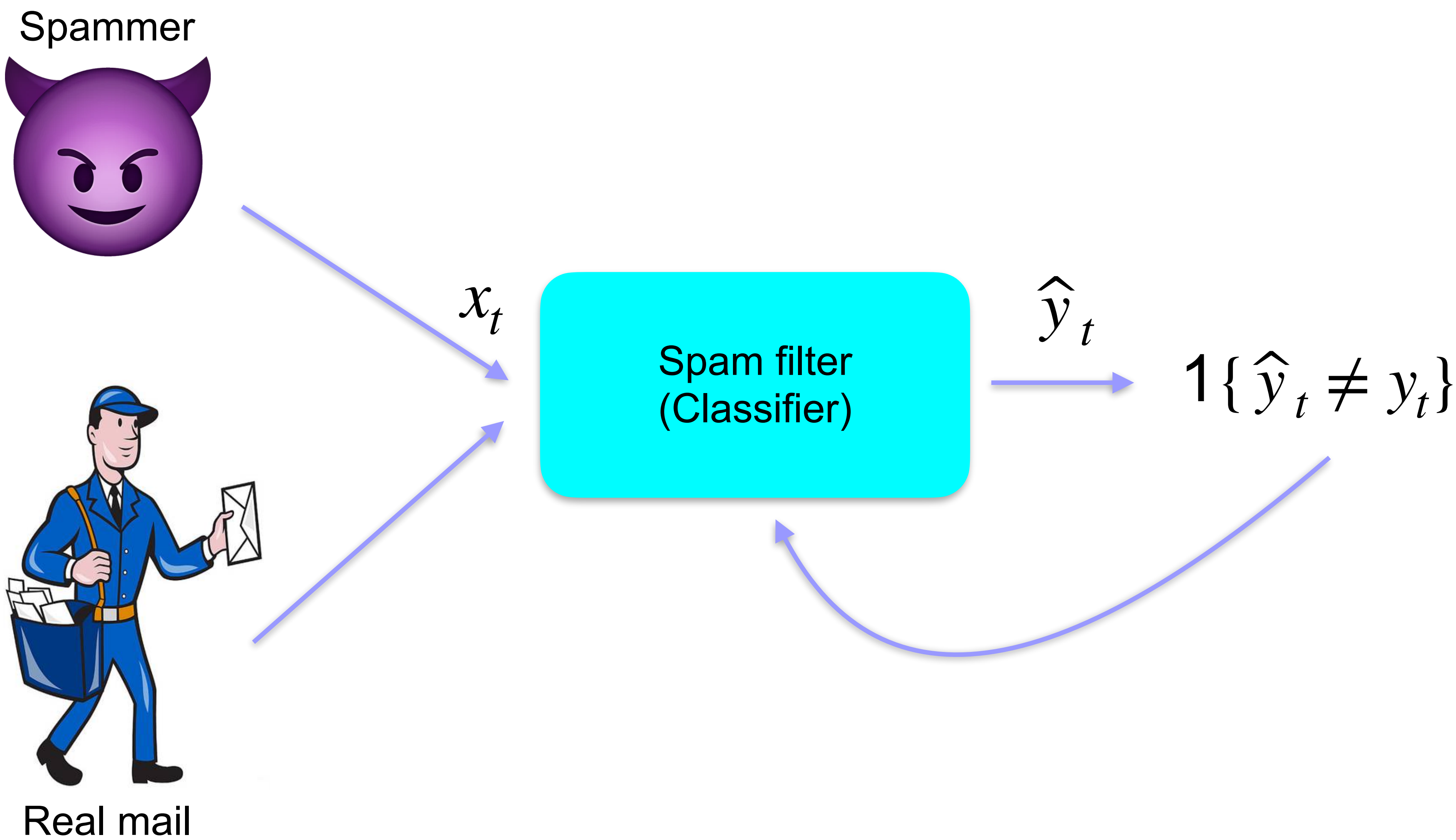
Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n) \overset{iid}{\sim} \nu$ where $y_i \in \{-1, 1\}$. For any $h \in \mathcal{H}$ define $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\}$ and $R(h) = \mathbb{P}(h(X) \neq Y)$ where $(X, Y) \sim \nu$. Assume there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$. If $\widehat{h} = \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)$ then with probability at least $1 - \delta$ we have

$$R(\widehat{h}) \leq \frac{\log(|\mathcal{H}|/\delta)}{n}$$

where $(X, Y) \sim \nu$.

All the guarantees of the previous section (and the entirety of this class so far) has relied critically on *(x,y)* being drawn **IID**. Can we say anything if *(x,y)* are chosen **adversarially**?

Online learning



Online learning

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Settings of interest:

IID $(x_t, y_t) \sim \nu$

Adversarial (x_t, y_t) arbitrary

Online learning - IID

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
  x_t arrives
  Player picks $h_t \in \mathcal{H}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

IID $(x_t, y_t) \sim \nu$

We know learning theory! Choose $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$

Online learning - IID

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

IID $\quad (x_t, y_t) \sim \nu$

Corollary Under the conditions of the theorem (i.e., there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$, $(x_i, y_i) \overset{iid}{\sim} \nu$, and $\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$) we have $\mathbb{E}[R(\widehat{h})] \leq \int_{\epsilon=0}^{d} \mathbb{P}(R(\widehat{h}) \geq \epsilon) \leq \frac{2\log(|\mathcal{H}|)}{n}$

Online learning - IID

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

IID $\quad (x_t, y_t) \sim \nu$

Corollary Under the conditions of the theorem (i.e., there exists an $h_* \in \mathcal{H}$ such that $R(h_*) = 0$, $(x_i, y_i) \overset{iid}{\sim} \nu$, and $\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$) we have $\mathbb{E}[R(\widehat{h})] \leq \int_{\epsilon=0}^{d} \mathbb{P}(R(\widehat{h}) \geq \epsilon) \leq \frac{2\log(|\mathcal{H}|)}{n}$

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}\right] \leq 1 + \sum_{t=2}^{T} \mathbb{E}[\mathbb{P}(h_t(x_t) \neq y_t)]$$
$$\leq 1 + \sum_{t=2}^{T} \mathbb{E}[R(h_t)] \leq 1 + \sum_{t=2}^{T} \frac{2\log(|\mathcal{H}|)}{t-1} \leq 2 + 2\log(|\mathcal{H}|)\log(T)$$

# of mistakes grows only logarithmically!

Online learning - Adversarial

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$

for $t = 1, 2, \ldots$

x_t arrives

Player picks $h_t \in \mathcal{H}$

y_t is revealed

Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:

Minimize mistakes

$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary

Online learning - Adversarial

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

`Input:` $\mathcal{H}$ with $|\mathcal{H}| < \infty$
`for` $t = 1, 2, \ldots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary $y_t = h_*(x_t)$ for $h_* \in \mathcal{H}$

We know learning theory! Choose $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$?

Online learning - Adversarial

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
Simultaneously { x_t arrives
Player picks $h_t \in \mathcal{H}$ }
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary $y_t = h_*(x_t)$

We know learning theory! Choose $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$?

Claim There exists a sequence $\{(x_t, y_t)\}_{t=1}^{T}$ and $\widehat{h}_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$ such that the strategy makes $\min\{|\mathcal{H}|, T\}$ mistakes.

Hint: many classifiers achieve minimum, assume adversary knows your tie-breaking strategy

Online learning - Adversarial

Goal:
Minimize mistakes $\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
    x_t arrives
    Player picks $h_t \in \mathcal{H}$
    y_t is revealed
    Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary $y_t = h_*(x_t)$

Halving Algorithm

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
Initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
    x_t arrives
    Player picks a $h_t \in V_t : \sum_{h \in V_t} \mathbf{1}\{h(x_t) = h_t(x_t)\} > \sum_{h \in V_t} \mathbf{1}\{h(x_t) = -h_t(x_t)\}$
    y_t is revealed
    Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$
    Update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$

Online learning - Adversarial

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
  x_t arrives
  Player picks $h_t \in \mathcal{H}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary

Halving Algorithm

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
Initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
  x_t arrives
  Player picks a $h_t \in V_t$: $\sum_{h \in V_t} \mathbf{1}\{h(x_t) = h_t(x_t)\} > \sum_{h \in V_t} \mathbf{1}\{h(x_t) = -h_t(x_t)\}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$
  Update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$

Either the algorithm doesn't make mistake, or *at least half* of hypotheses are discarded

Online learning - Adversarial

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Goal:
Minimize mistakes
$\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\}$

Adversarial (x_t, y_t) arbitrary

Theorem: Fix a finite hypothesis class $\mathcal{H}$ so that $|\mathcal{H}| < \infty$ and for all $h \in \mathcal{H}$ we have $h(x) \in \{-1, 1\}$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ where x_t is arbitrary and $y_t = h_*(x_t)$ for some $h_* \in \mathcal{H}$. Then if h_t is recommended by the Halving algorithm, we have that $\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} \leq \log_2(|\mathcal{H}|)$

Online learning

Assuming that your data is IID is a **very** strong assumption that is almost never true in practice. Online learning is a different paradigm that makes no assumptions but still yields meaningful guarantees.

Assuming there exists a perfect classifier h_*:

- When x_t is drawn IID, empirical risk minimization results in only a number of mistakes that grows like $\log(T)\log(H)$

- When x_t is chosen adversarially empirical risk minimization can do arbitrarily badly. But there exist smarter approaches (like Halving algorithm) that make only $\log(H)$ mistakes

Questions?

Exponential weights

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	…
Expert 1	.7					
Expert 2	.4					
Expert 3	.6					

Truth z_t: .5

$\ell_t(i) = |z_t - b_t(i)|$

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

t=1 *t=2* *t=3* *t=4* *t=5* …

Expert 1

Expert 2

Expert 3

$$z_t(i) = |b_t(i) - y_t|$$

ith expert's prediction

True temperature

$\texttt{Input:}$ d experts
$\texttt{for}$ $t = 1, 2, \ldots$
    Player picks $p_t \in \triangle_d$ and plays $I_t \sim p_t$
    Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
    Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

$z_t(I_t)$

Expert prediction

Suppose $b_t \in [0,1]^d$ is a vector of **d** experts predictions of tomorrow's temperature.

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	...
Expert 1						
Expert 2						
Expert 3						

$$z_t(i) = |b_t(i) - y_t|$$

*i*th expert's prediction

True temperature

$\texttt{Input:}$ d experts
$\texttt{for}$ $t = 1, 2, \dots$
Player picks $p_t \in \triangle_d$ and plays $I_t \sim p_t$
Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

Goal: Minimize regret wrt best $\max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle$

$= \max_i \mathbb{E}\left[\sum_{t=1}^{T} z_t(I_t) - z_t(i) \right]$

Expert prediction

Goal: Minimize regret wrt best $\max_{i \in [d]} \sum_{t=1}^{T} \langle p_t, z_t \rangle - \langle \mathbf{e}_i, z_t \rangle$

$\texttt{Input:}$ d experts
$\texttt{for}$ $t = 1, 2, \ldots$
 Player picks $p_t \in \triangle_d$ and plays $I_t \sim p_t$
 Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
 Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$

Exponential weights algorithm

$\texttt{Input:}$ d experts, $\eta > 0$
$\texttt{Initialize:}$ $w_1 \in [1, \ldots, 1]^\top \in \mathbb{R}^d$
$\texttt{for}$ $t = 1, 2, \ldots$
 Player plays $I_t \sim p_t$ where $p_t(i) = w_t(i) / \sum_{j=1}^{d} w_t(j)$
 Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$
 Player pays loss $\langle p_t, z_t \rangle = \mathbb{E}[z_t(I_t)]$
 Player updates weights $w_{t+1}(i) = w_t(i) \exp(-\eta z_t(i))$

Expert prediction

Goal: Minimize regret wrt best $\max_{i\in[d]} \sum_{t=1}^{T} \langle p_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle$

Exponential weights algorithm

`Input:` d experts, $\eta > 0$

`Initialize:` $w_1 \in [1,\ldots,1]^\top \in \mathbb{R}^d$

`for` $t = 1, 2, \ldots$

Player plays $I_t \sim p_t$ where $p_t(i) = w_t(i)/\sum_{j=1}^{d} w_t(j)$

Adversary simultaneously reveals expert losses $z_t \in [0,1]^d$

Player pays loss $\langle p_t, z_t\rangle = \mathbb{E}[z_t(I_t)]$

Player updates weights $w_{t+1}(i) = w_t(i)\exp(-\eta z_t(i))$

Theorem: If $z_t \in [0,1]^d$ $\forall t$, and I_t, p_t are chosen by exponential weights then
$\max_{i\in[d]} \mathbb{E}\left[\sum_{t=1}^{T} \langle I_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle\right] = \max_{i\in[d]} \sum_{t=1}^{T} \langle p_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle \le \frac{\log(d)}{\eta} + \frac{T\eta}{8}$

Choosing $\eta = \sqrt{\frac{8\log(d)}{T}}$ gives regret bound of $\sqrt{T\log(d)/2}$

Online learning in non-separable case

Online learning

Goal: Minimize regret wrt best

$$\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

$\texttt{Input:}$ $\mathcal{H}$ with $|\mathcal{H}| < \infty$
$\texttt{for}$ $t = 1, 2, \dots$
x_t arrives
Player picks $h_t \in \mathcal{H}$
y_t is revealed
Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Settings of interest:

IID $(x_t, y_t) \sim \nu$

Adversarial (x_t, y_t) arbitrary

Online learning

Goal: Minimize regret wrt best

$$\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

Input: $\mathcal{H}$ with $|\mathcal{H}| < \infty$
for $t = 1, 2, \ldots$
  x_t arrives
  Player picks $h_t \in \mathcal{H}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Settings of interest:

IID $(x_t, y_t) \sim \nu$

Choose $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$

Corollary Under the conditions of the theorem (i.e., $(x_i, y_i) \overset{iid}{\sim} \nu$, and $\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$) and $|\mathcal{H}| \geq n$, we have $\mathbb{E}[R(\widehat{h})] - R(h_*) \leq \sqrt{\frac{8 \log(|\mathcal{H}|)}{n}}$

$$\implies \max_{h \in \mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}\right] \leq \sqrt{8T \log(|\mathcal{H}|)}$$

Online learning

Goal: Minimize regret wrt best

$$\max_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

$\texttt{Input:}$ $\mathcal{H}$ with $|\mathcal{H}| < \infty$
$\texttt{for}$ $t = 1, 2, \ldots$
  x_t arrives
  Player picks $h_t \in \mathcal{H}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Settings of interest:

IID $(x_t, y_t) \sim \nu$

Adversarial (x_t, y_t) arbitrary

Theorem: If $z_t \in [0,1]^d$ $\forall t$, and I_t, p_t are chosen by exponential weights then
$\max_{i\in[d]} \mathbb{E}\left[\sum_{t=1}^T \langle I_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle\right] = \max_{i\in[d]} \sum_{t=1}^T \langle p_t, z_t\rangle - \langle \mathbf{e}_i, z_t\rangle \leq \sqrt{T\log(d)/2}$

$$\implies \max_{h\in\mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}\right] \leq \sqrt{T\log(|\mathcal{H}|)/2}$$

Online learning

Goal: Minimize regret wrt best

$$\max_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

`Input:` $\mathcal{H}$ with $|\mathcal{H}| < \infty$
`for` $t = 1, 2, \ldots$
  x_t arrives
  Player picks $h_t \in \mathcal{H}$
  y_t is revealed
  Player receives loss $\ell(h_t, (x_t, y_t)) = \mathbf{1}\{h_t(x_t) \neq y_t\}$

Settings of interest:

IID $\quad (x_t, y_t) \sim \nu$

$$\implies \max_{h\in\mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}\right] \leq \sqrt{8T\log(|\mathcal{H}|)}$$

Adversarial $\quad (x_t, y_t)$ arbitrary

$$\implies \max_{h\in\mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}\right] \leq \sqrt{T\log(|\mathcal{H}|)/2}$$

Online learning

Assuming that your data is IID is a **very** strong assumption that is almost never true in practice. Online learning is a different paradigm that makes no assumptions but still yields meaningful guarantees.

This section does not assume there exists a perfect classifier h_* but still has strong guarantees on the regret even under adversarially chosen data!

$$\implies \max_{h \in \mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}\right] \leq \sqrt{T \log(|\mathcal{H}|)/2}$$

But requires enumerating hypotheses… not computationally efficient.
What about infinite hypotheses?

Questions?

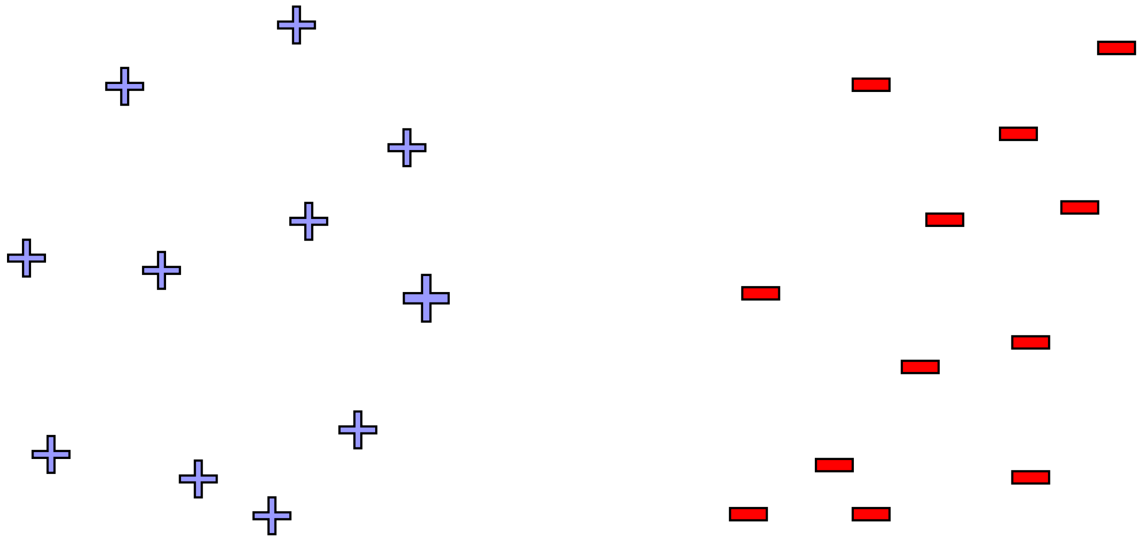
Perceptron

Online learning

- Halving algorithm is efficient, but what about infinite hypothesis classes and computational efficiency?
- Click prediction for ads is a streaming data task:
  - User enters query, predict if a particular ad will be clicked on or not
    - Observe $x_t \in \mathbb{R}^d$, and must predict $y_t \in \{-1, 1\}$
  - User either clicks or doesn't click on ad
    - Label y_t is revealed afterwards
      - Google gets a reward if user clicks on ad
  - Update model for next time

Binary Classification

Assume data is linearly separable:



The Perceptron Algorithm [Rosenblatt '58, '62]

- Classification setting: $y_t \in \{-1, 1\}$
- Linear model
  - Prediction:
- Training:
  - Initialize weight vector:
  - At each time step:
    - Observe features:
    - Make prediction:
    - Observe true class:
    - Update model:
      - If prediction is not equal to truth

The Perceptron Algorithm [Rosenblatt ‘58, ‘62]

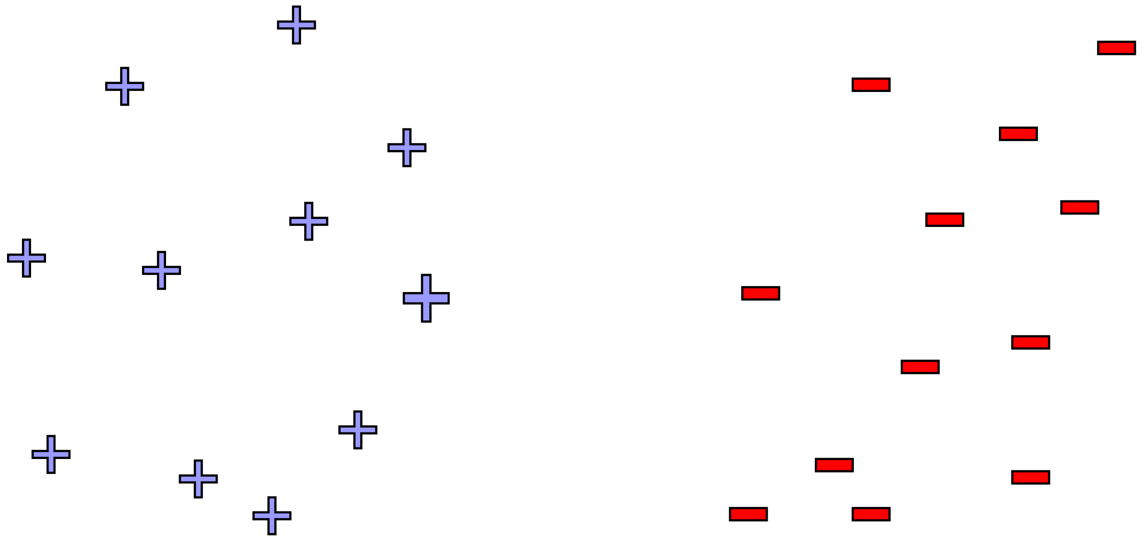
- Classification setting: $y_t \in \{-1,1\}$
- Linear model
  - Prediction: $\text{sign}(w^\top x_t)$

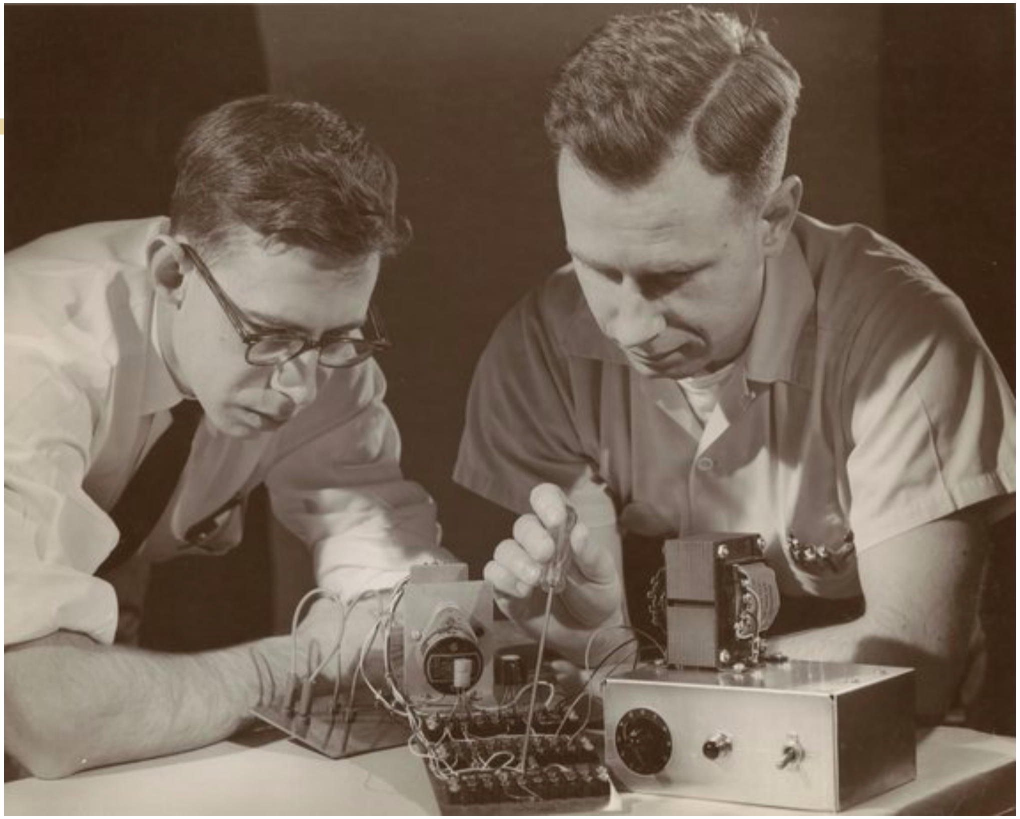
- Training:
  - Initialize weight vector: $w_1 = 0 \in \mathbb{R}^d$
  - At each time step:
    - Observe features: $x_t \in \mathbb{R}^d$
    - Make prediction: $\text{sign}(w_t^\top x_t)$
    - Observe true class: $y_t \in \{-1,1\}$

    - Update model:
      - If prediction is not equal to truth $w_{t+1} = w_t + x_t y_t$

Binary Classification

Assume data is linearly separable:





Rosenblatt 1957

"the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

The New York Times, 1958

Perceptron Analysis: Linearly Separable Case

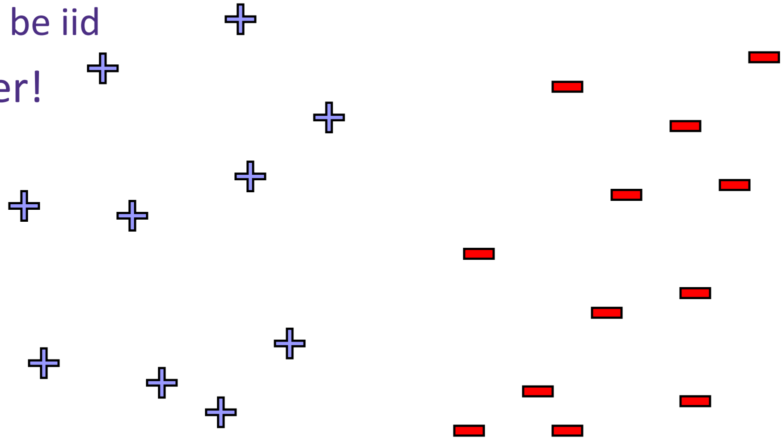
- **Theorem** [Block, Novikoff]:
  - Given a sequence of labeled examples: $(x_1, y_1), (x_2, y_2), \ldots$
  - Each feature vector has bounded norm: $\|x\|_2^2 \leq R^2$
  - If dataset is linearly separable with a margin:

    Exists $w_* \in \mathbb{R}^d$ such that $w_*^\top x_t y_t \geq \gamma$

then for w_t from perceptron we have $\sum_{t=1}^{T} \mathbf{1}\{\text{sign}(w_t^\top x_t) \neq y_t\} \leq \frac{R^2}{\gamma^2}$

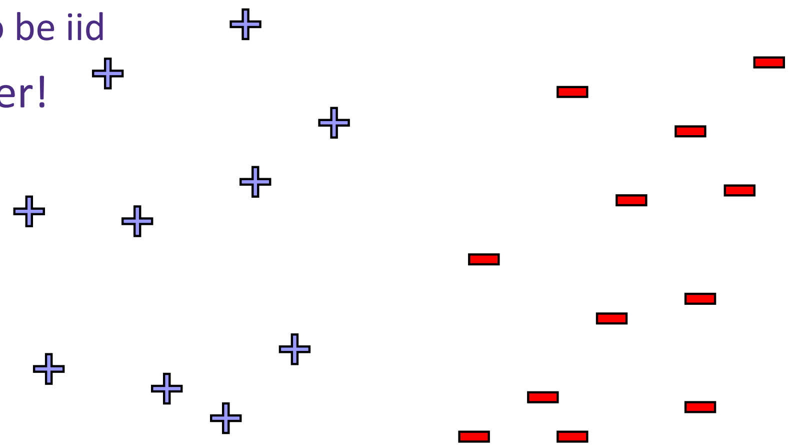
Beyond Linearly Separable Case

- Perceptron algorithm is super cool!
  - No assumption about data distribution!
    - Could be generated by an oblivious adversary, no need to be iid
  - Makes a fixed number of mistakes, and it's done for ever!
    - Even if you see infinite data



Beyond Linearly Separable Case

- Perceptron algorithm is super cool!
  - No assumption about data distribution!
    - Could be generated by an oblivious adversary, no need to be iid
  - Makes a fixed number of mistakes, and it's done for ever!
    - Even if you see infinite data



- Perceptron is useless in practice!
  - Real world not linearly separable
  - If data not separable, cycles forever and hard to detect
  - Even if separable may not give good generalization accuracy (small margin)

What is the Perceptron Doing???

- When we discussed logistic regression:
  - Started from maximizing conditional log-likelihood

- When we discussed the Perceptron:
  - Started from description of an algorithm

- What is the Perceptron optimizing???? (Wait a few slides)

Online Convex Optimization

UNIVERSITY *of* WASHINGTON

Convex surrogate loss functions

Previous section for the adversarial case suggested using multiplicative weights over the |H| hypotheses, which is completely intractable in practice.

And in the stochastic case we used $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$ which is also intractable to compute!

So it seems we have no practical algorithm! Solution: relax the objective.

Convex surrogate loss functions

Previous section for the adversarial case suggested using multiplicative weights over the |H| hypotheses, which is completely intractable in practice.

And in the stochastic case we used $h_t \in \arg\min_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbf{1}\{h(x_s) \neq y_s\}$ which is also intractable to compute!

So it seems we have no practical algorithm! Solution: relax the objective.

Instead of $$\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbf{1}\{h_t(x_t) \neq y_t\} - \mathbf{1}\{h(x_t) \neq y_t\}$$

We use $$\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h_t, (x_t, y_t)) - \ell(h, (x_t, y_t))$$ with $\mathcal{H}$ convex

Example: Linear classification takes $\mathcal{H} \subset \mathbb{R}^d$ and $\ell(h, (x_t, y_t)) = \log(1 + \exp(-y_t h^\top x_t))$

Convex surrogate loss functions

Goal: $\max_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h_t, (x_t, y_t)) - \ell(h, (x_t, y_t))$ with $\mathcal{H}$ convex

Online gradient descent

$\texttt{Input:}$ $\mathcal{H} \subset \mathbb{R}^d$, convex loss function ℓ, step size $\eta > 0$

$\texttt{Initialize:}$ Choose any $h_1 \in \mathcal{H}$

$\texttt{for}$ $t = 1, 2, \ldots$

Player plays $h_t \in \mathcal{H}$

Adversary simultaneously reveals (x_t, y_t)

Player pays loss $\ell_t(h_t) := \ell(h_t, (x_t, y_t))$

Player updates $w_{t+1} = \Pi_{\mathcal{H}}(w_t - \eta \nabla_h \ell_t(h_t))$

Theorem Online gradient descent satisfies for any $h_* \in \mathcal{H}$

$$\sum_{t=1}^{T} \ell(h_t, (x_t, y_t)) - \ell(h_*, (x_t, y_t)) \leq \frac{\|h_*\|_2^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_h \ell_t(h_t)\|_2^2$$

if $\max_{h \in \mathcal{H}} \|h_*\|_2 \leq R$ and $\ell(\cdot)$ is G-Lipschitz then regret $\leq RB\sqrt{T}$

Proof

Theorem Online gradient descent satisfies for any $h_* \in \mathcal{H}$

$$\sum_{t=1}^T \ell(h_t,(x_t,y_t)) - \ell(h_*,(x_t,y_t)) \leq \frac{\|h_*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\nabla_h \ell_t(h_t)\|_2^2$$

Questions?