ANNOUNCEMENTS

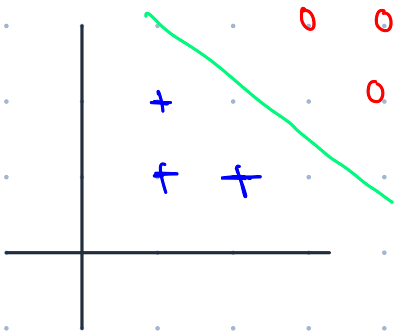
NOT 11:59 pm

1. HW5 due immediately after spring break (04/07, 5pm)
   (04/09, 5pm - slip days)
2. Tushaar's Th OH cancelled [due to travel], OH over break are cancelled also!
3. Academic Integrity - use of GenAI tools must be disclosed w/ prompts!
4. Other deadlines (P4, P5, etc.), check Ed!

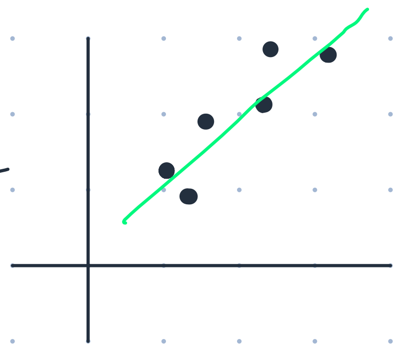
As usual, turn your non-note-taking devices off!

__So far__ : " much time has passed, many algorithms learned "

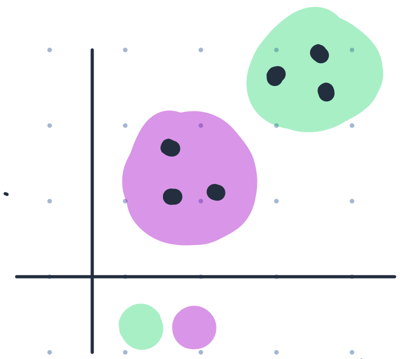
SUPERVISED UNSUPERVISED



OR

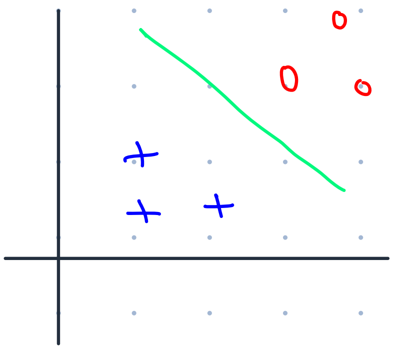


vs.

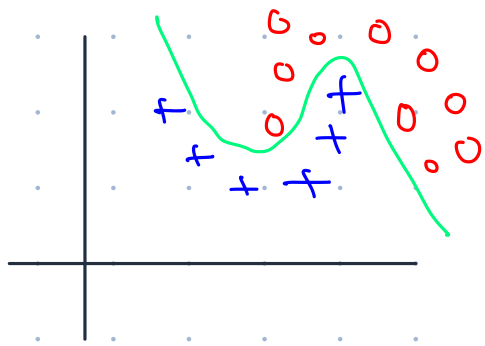


(classification) (Regression) Unsupervised

linearly-separable



non-linearly-separable



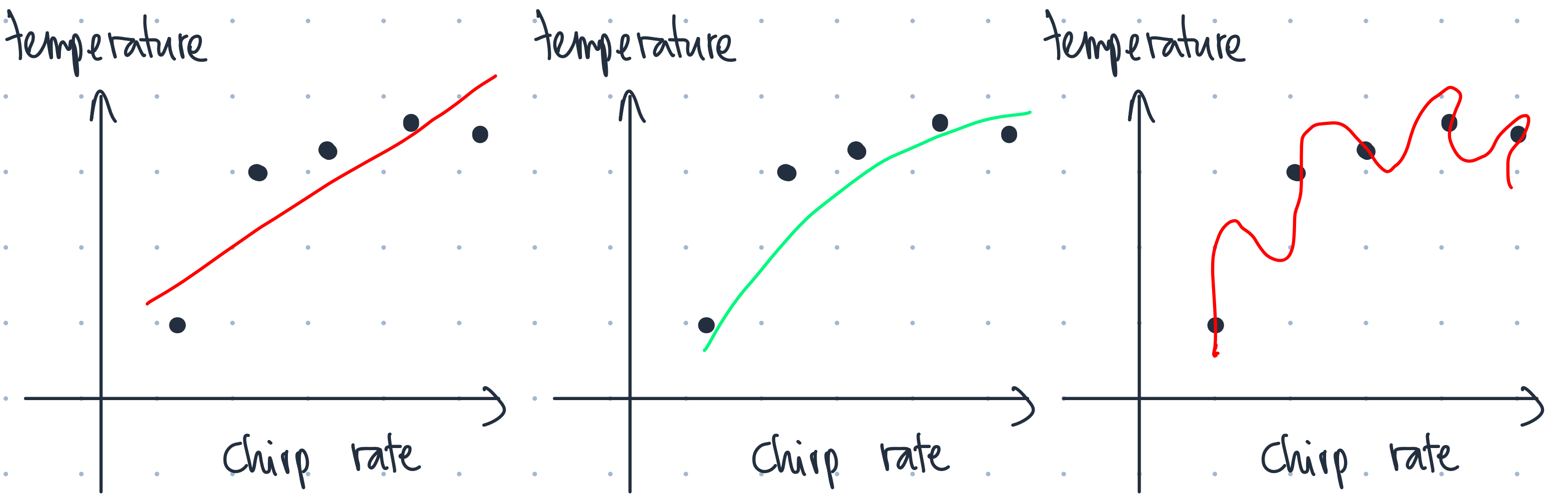
Today — view the fundamental "learning problem" from the lens of risk min

NOT a specific algorithm

Find θ that minimizes some cost fn. $J(\theta)$

Q. When does an algorithm / class of algorithm succeed?

Crickets chirp and temperature raises! — "generalization"

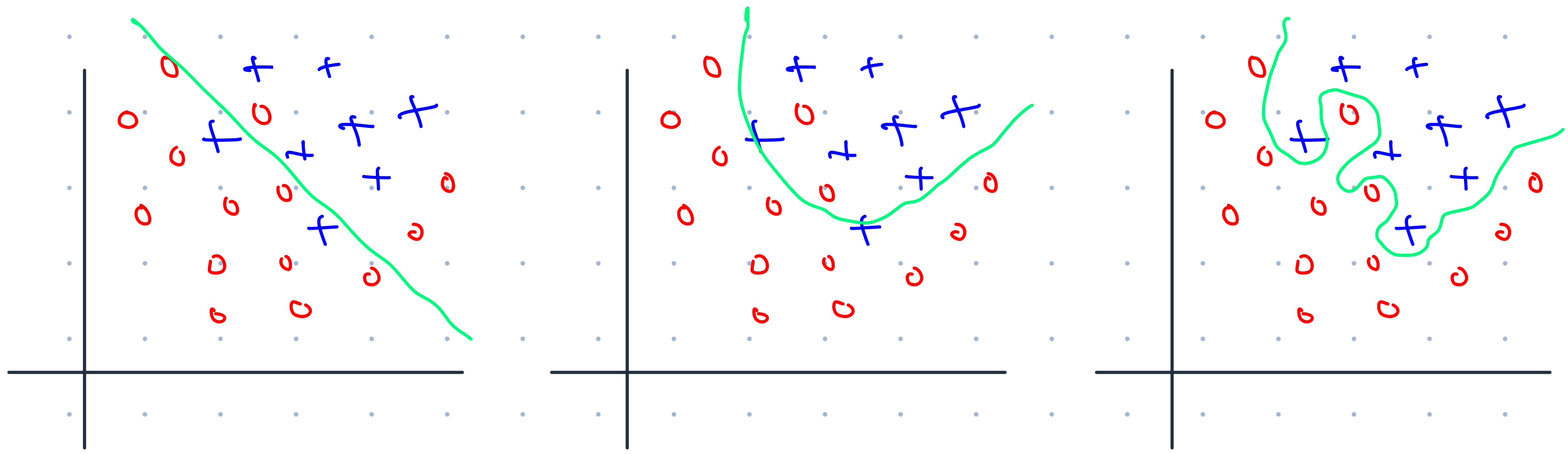


temp^r $\propto$ chirp rate

temp^r = $\theta_0 + \theta_1$ rate + θ_2 rate^2

Griffin says too specific to data

Same game, different name — Classification w/ logistic regression



A case of binary classification

Given $D = \{(x^{(j)}, y^{(j)})\}_{j=1}^{n}$ of n samples.

$$(x^{(j)}, y^{(j)}) \overset{\text{iid}}{\sim} P$$

Given a fn, hypothesis "h", the training error

$$\hat{\varepsilon}(h) = \frac{1}{n} \sum_{j=1}^{n} \boxed{1\{\underbrace{h(x^{(j)})}_{\text{prediction}} \neq y^{(j)}\}}$$

misclassification

train error OR

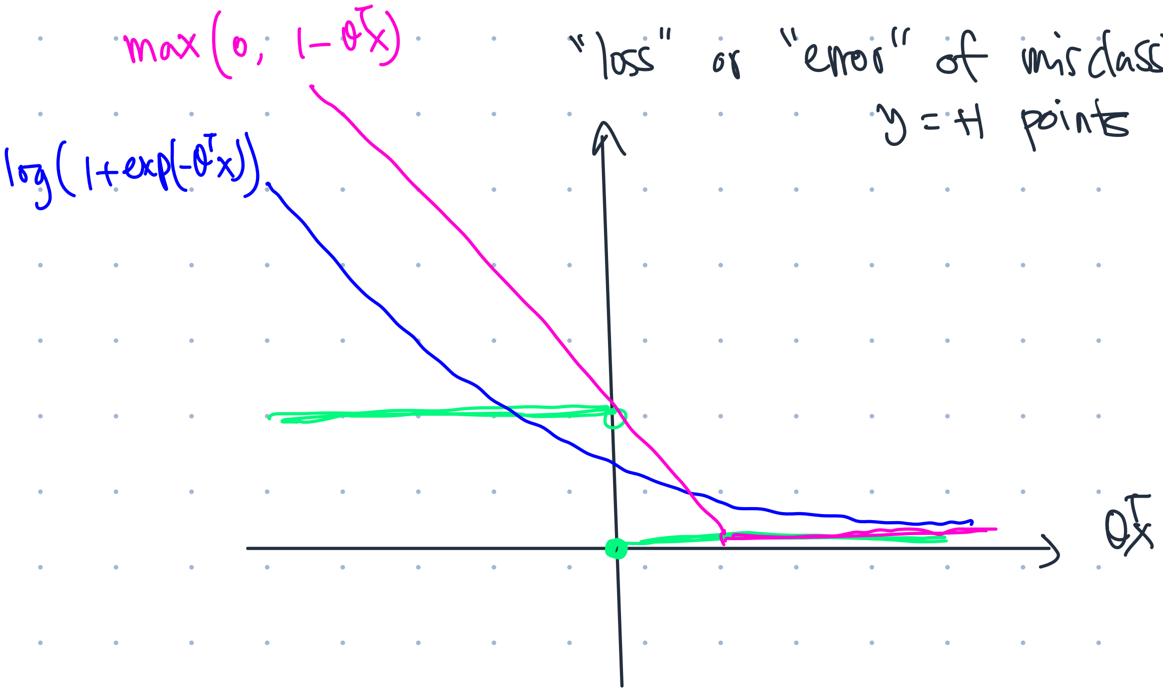
$\hat{\varepsilon}_D(h)$ Empirical risk

We want to choose "$\hat{\theta}$" s.t.

$$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h)$$

Empirical risk minimization!

logistic regression, SVMs, etc. are convex approximates to 0/1 loss



From "find the optimal θ to minimize $\hat{\varepsilon}(h_\theta)$" to
<u>Empirical</u> risk minimization

before : $\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta)$

given $\mathcal{H}$, a class of all hypotheses,

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

linear classifiers : $\mathcal{H} = \{ x \to \sigma(\theta^T x) \mid \theta \in \mathbb{R}^{d+1} \}$

linear regression : $\mathcal{H} = \{ x \to \theta^T x \mid \theta \in \mathbb{R}^{d+1} \}$.

<u>What do we care about?</u>

<u>NOT</u> $\hat{\varepsilon}(h)$ = train error

Generalization error

$$\varepsilon(h) = P_{(x,y)\sim \mathcal{D}} \, 1\{ h(x) \neq y \}$$

probability of misclassifying some "x"

What do we want?

1) given the training error $\hat{\varepsilon}(h)$, can we guarantee anything about generalization? $\varepsilon(h)$

2) Can we estimate this "generalization error" — $\varepsilon(h)$

if the traing dataset is at least ___ large,

$\mathcal{H}$ at most ___ complex,

then ___ w.p., we have training error within γ of test error.

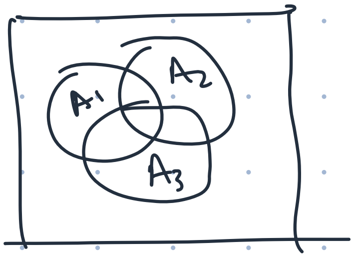
Some preliminaries

Lemma

(The union bound). $P(A_1 \cup A_2 \ldots \cup A_k) \leq P(A_1) + P(A_2) + \ldots + P(A_k)$

"or"

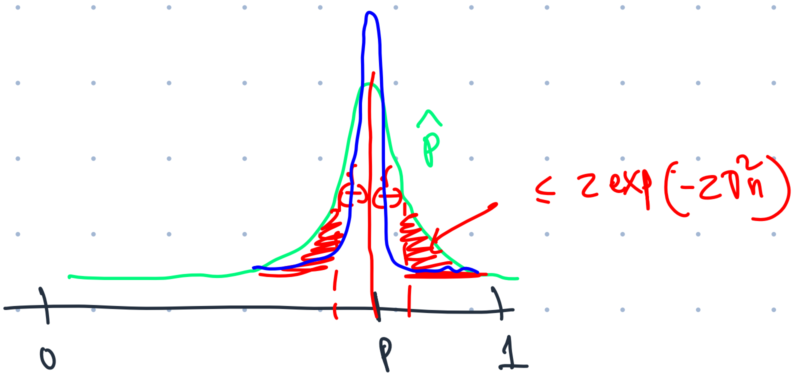
$A_1, \ldots, A_k$ events, not necessarily independent



Lemma

(The Chernoff bound). If we had a coin w/ $P(H) = p$, flipped "n" times, then $\hat{p}$ = fraction of times we see H

$$P(|p - \hat{p}| > r) \leq 2\exp(-2r^2 n)$$



Takeaway: As $n \uparrow$, we get better estimates for $\hat{p}$

'exp' better

Relate train error to generalization error

"hat" vs. "no hat"

$1\{h(x) \neq y\} = z$ — (x, y) misclassified?

$P(z=1) = \varepsilon(h)$ → p in coin toss example

$\varepsilon(h)$ = generalization

$\hat{\varepsilon}(h)$ = train

$1\{h(x^{(j)}) \neq y^{(j)}\} = z^{(j)}$ — $(x^{(j)}, y^{(j)})$ misclassified

now, $\frac{1}{n}\sum_{j=1}^{n} z^{(j)} = \hat{\varepsilon}(h)$ → $\hat{p}$ in coin toss example

$$P\left(|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma\right) \leq 2\exp(-2\gamma^2 n)$$

for <u>ONE</u> $h \in \mathcal{H}$

$$P\left(\left| \varepsilon(h) - \hat{\varepsilon}(h) \right| > \gamma \right) \leq 2\exp(-2\gamma^2 n)$$

for $\underline{\text{ONE}}$ $h \in \mathcal{H}$

Wish to extend to any $h \in \mathcal{H}$:

Q. what is the probability that one or more h_js result in $|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma$?

let A_j to be the event that $|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma$

$$P(A_1 \cup A_2 \cup \dots \cup A_k) \leq P(A_1) + \dots + P(A_k)$$

$$\leq \sum_{j=1}^{k} 2\exp(-2\gamma^2 n)$$

$$= 2k\exp(-2\gamma^2 n)$$

error margin that we chose

$|\mathcal{H}|$ = num of hypothesis n = dataset size

$$P(\neg\, h_j \in \mathcal{H} \text{ s.t. } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma) \geq 1 - 2k\exp(-2\gamma^2 n)$$

probability that $\underline{\text{none}}$ of $h_1, \dots, h_k$ result in $|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma$

Alternatively, ...

Given some $\gamma > 0$, $0 < \delta < 1$, how large a dataset is needed to guarantee w.p. $1-\delta$, the training error is within "γ" of generalization error?

previous result –

$$\underbrace{P(\neg\, h_j \in \mathcal{H} \text{ s.t. } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma)}_{\text{probability that none of } h_1, \dots, h_k \text{ result in } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma} \geq 1 - 2k\exp(-2\gamma^2 n)$$

$$1 - \delta \geq 1 - 2k\exp(-2\gamma^2 n)$$

$$\delta \leq 2k\exp(-2\gamma^2 n)$$

$$\log\frac{\delta}{2k} \leq -2\gamma^2 n \quad \Rightarrow -2\gamma^2 n \geq \log\frac{\delta}{2k}$$

$$n \geq -\frac{1}{2\gamma^2}\log\frac{\delta}{2k}$$

$$\text{(or)} \quad n \geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta}.$$

$$n = O_{\gamma,\delta}(\log k)$$

Error bound

Fix n and δ, solve for margin -

$$|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| \leq \sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}$$

So, what can we say about $\underline{\underline{\hat{h}}}$?

ERM - identifies $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h) \longrightarrow$ best ERM hypothesis

$h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h) \longrightarrow$ best-in-$\mathcal{H}$ hypothesis

$$\gamma = 2\sqrt{\ldots}$$

$$\begin{aligned}
\varepsilon(\hat{h}) &\le \hat{\varepsilon}(\hat{h}) + \gamma \\
&\le \hat{\varepsilon}(h^*) + \gamma \\
&= \boxed{\hat{\varepsilon}(h^*) - \varepsilon(h^*)} + \varepsilon(h^*) + \gamma \\
&\le \gamma + \varepsilon(h^*) + \gamma
\end{aligned}$$

since $\hat{h}$ is chosen to minimize train err $\hat{\varepsilon}$, no h^* can achieve lower tr error

$$|\hat{\varepsilon}(h^*) - \varepsilon(h^*)| \le \gamma$$

OR $\varepsilon(\hat{h}) \le \varepsilon(h^*) + 2\gamma$

ORACLE INEQUALITY!

Finite $\mathcal{H}$?

DeepSeek-671B model, $2^{32 \times 671B}$ hypotheses

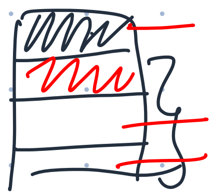
We need 190T examples, to say with 50% proba that generalization error of my model is 20% worse that the best model

Model selection

- Can we estimate $\varepsilon(h)$ somehow?

  what if $\hat{\varepsilon}(h)$ chosen to be $\varepsilon(h)$?

1. 70/30 split, estimate $\varepsilon(h)$ using 30% split
2. k-fold CV –



3. leave-one-out – every data point is used to test!