Discussion 5: (9/19/2018)

1- iid vs id

2- VC dimension

3- handle missing

- iid (independent & identical distribution)

  $X_1, X_2 .., X_n$ they are generated from an iid Source, the it should satisfy the following

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i)$$

$$P(X_i) \text{ is the same } \forall i \in \{1, \dots, n\}$$

- id (indepent distribution)

$$Y_i = X_i + \varepsilon \xrightarrow{} \varepsilon \perp X_i, \ \varepsilon \sim N(0,1)$$

if X_i's are iid, what is the distribution of Y_i given X_i are observed.

$$P(Y_1, \dots, Y_n | X_1 \dots X_n) = \prod_{i=1}^{n} P(y_i | X_i)$$

$$P(Y_i | X_i) \sim N(x_i, 1) \rightsquigarrow = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}(y_i - x_i)^2\right)$$

(1)

$p(Y_i|X_i)$ is a function x_i ⇒ it is Not identical for all i, but $p(Y_i|X_i)$ are indep. from each other.

VC dimension :

we have seen how complexity a model can affect the performance of the model

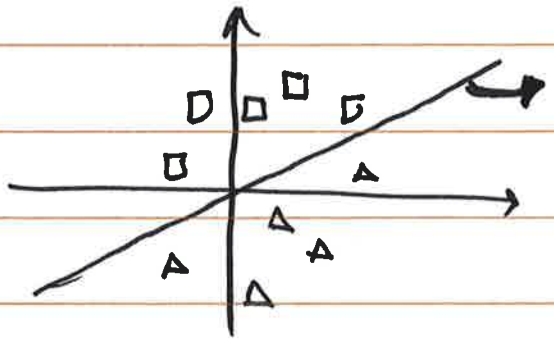
need to define complexity of model

VC dimension can be used to define complexity.

representation power is the ability of model to lear a wide variety of input-output relationship / or ability to memorize to fit data

Example: perceptron for 2D

$$\hat{c}(\underline{x}) = Sgn(\theta_1 x_1 + \theta_2 x_2 + \theta_0)$$



choose a different form of classifier

⇒ results in different model & performance

$$\hat{c}(\underline{x}) = \text{sgn}(\theta_1 x_1 + \theta_2 x_2)$$

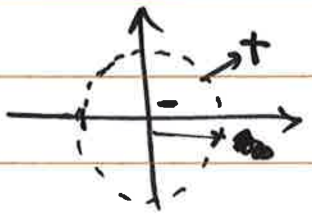
↓

classifier (learner) should pass through origin while classify data points.

Example: $\hat{c}(\underline{x}) = \text{sgn}(\underbrace{x^T x}_{\|x\|^2} - \theta_0)$

if value of $\|x\|$ is large → it is classified as +1

otherwise → " " " " −1



$\|x\|^2 > \theta_0 \to +1$

$\|x\|^2 < \theta_0 \to -1$

Tradeoff:

the more representation power we can have for classifier, the more accurate model on training data we have, but this is at the cost of overfitting

the less rep. power ↓ ⇒ poor performance on training

How can we $\underline{\underline{\text{quantify}}}$ represtation power?
(VC dim.)

Let's have some preliminary on VC:

- let's assume our training data are iid from dist. $p(x)$
- let's define "risk" & "emperical risk"

These are just "long term" test & "observed" training errors

$$R(\theta) = \text{Test Error} = E\big(\delta(c \neq c(x;\theta))\big)$$

$$R^{emp}(\theta) = \text{Train Error} = \frac{1}{m}\sum_i \delta\big(c^{(i)} \neq c(x^{(i)};\theta)\big)$$

Q: How are these errors related?

- underfitting domain:

  Test error is similar to train error

- over fitting domain:

  Test error is be way worse!

VC dim. tells us about risk:

- Given some classifier, let H be its VC dim.
- with " high prob" $1-\eta$, it can be shown that

$$\text{Test error} \le \text{Train Error} + \sqrt{\frac{H \log\left(\frac{2m}{H}\right) + H - \log\frac{\eta}{4}}{m}}$$

(m is # of training data)

$H \uparrow \Longrightarrow$ overfitting

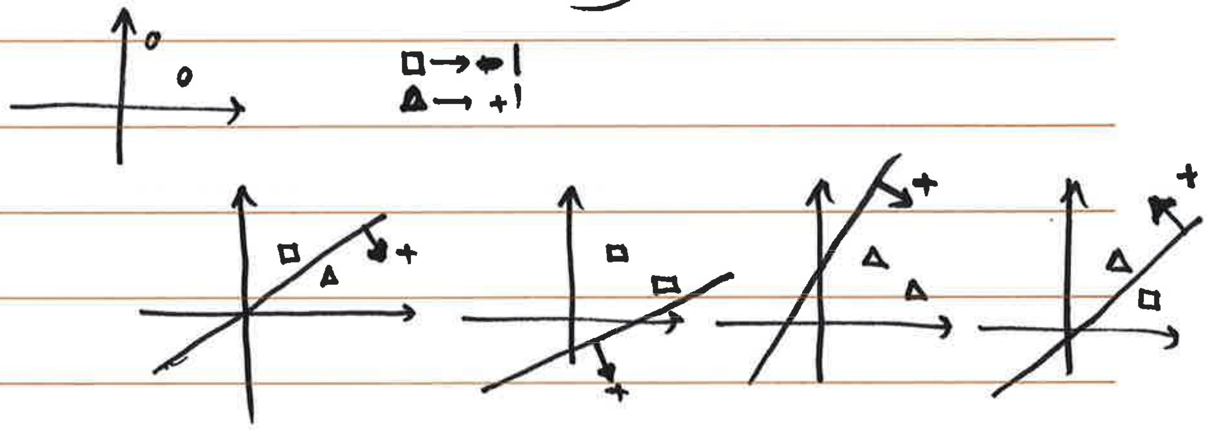
$H \downarrow \Longrightarrow$ Test & Trains are similar

Shattering:

We say a classifier $f(x)$ can shatter points $x^{(1)} \dots x^{(n)}$ iff <u>for all</u> $y^{(1)} \dots y^{(n)}$ $f(x)$ can achieve zero error on training data.

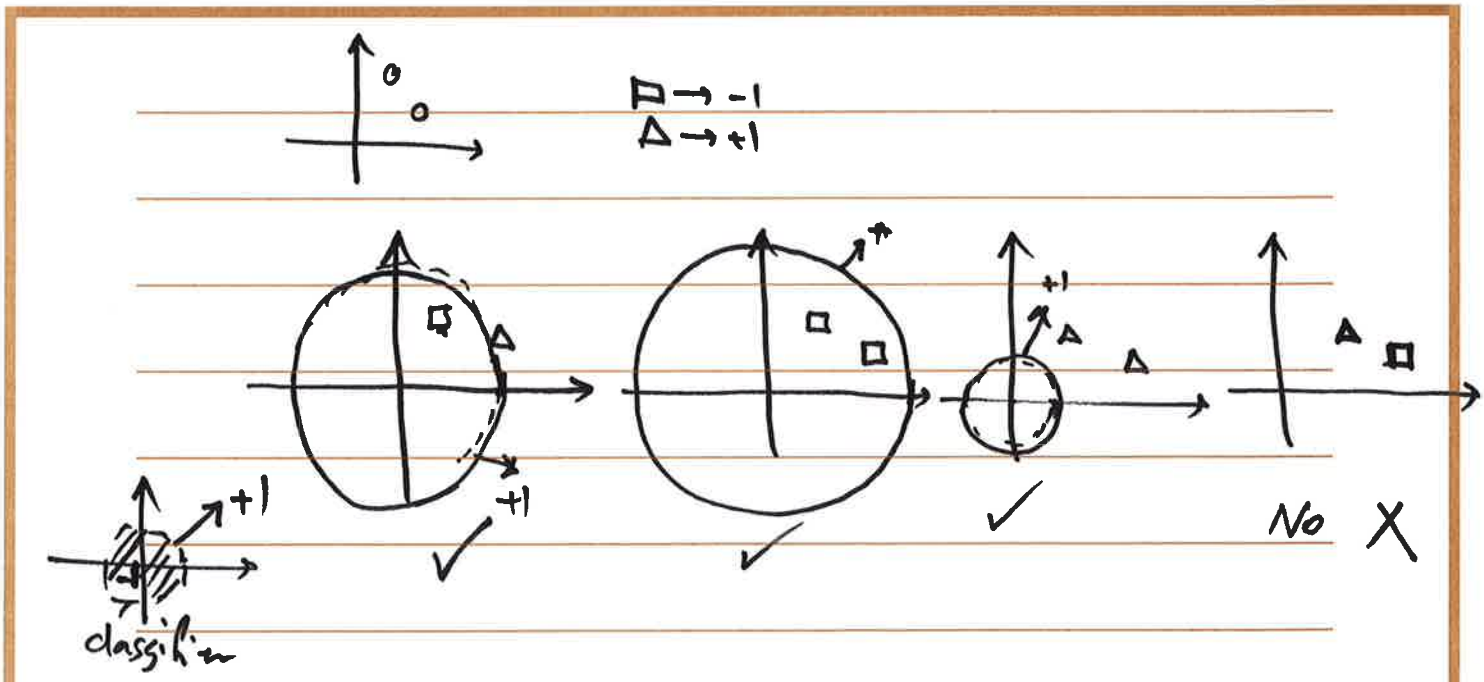
Example: (2D) Can $f(x;\theta) = \text{Sgn}(\theta_1 x_1 + \theta_2 x_2 + \theta_3$

shatter the following points? Yes



Example : (2D) Can $f(x;\theta) = \text{Sgn}(\bullet x^T x - \theta)$

shatter the following points? NO



VC dim :

The VC dim H is defined as the max. # of points h that can be arranged so that f(x) can shatter all of them.

A game :

- fix the del $f(x;\theta)$
- Player 1 : choose location of $x^{(1)} \dots x^{(h)}$
- " 2 : " target labels of $x^{(1)} \dots x^{(h)}$ $(y^{(1)} \dots y^{(h)})$
- Player 1 : choose value θ

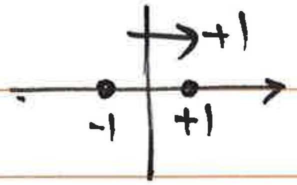
if $f(x;\theta)$ can reproduce the target labels, Player 1 wins

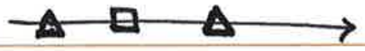
$$\exists \{x^{(i)}\}_{i=1}^{h} \text{ s.t. } \underline{\forall} \{y^{(i)}\}_{i=1}^{h} \ \underline{\exists} \theta \text{ s.t.}$$

$$f(x^{(i)};\theta) = y^{(i)} \quad \underline{\forall} i \in \{1, \dots, h\}$$

Example: what is the VC dimension of 1D linear perception

$$\hat{c}(x) = \text{Sgn}(\theta_1 x + \theta_0)$$

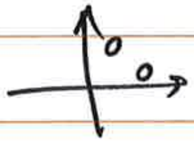


3 points  ✗

$\Longrightarrow$ VC of 1D linear perceptron is 2.

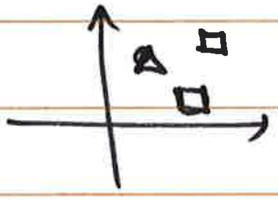
Example: VC dim. of 2D

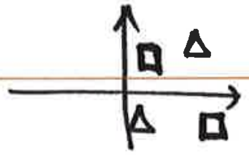
$\partial(x) = \text{Sgn}(\theta_1 x_1 + \theta_2 x_2 + \theta_0)$

2 points ✓ 

3 points ✓

(8 possible labling)



4 points 

VC dimensin of this problem is 3.

* VC dimension of d dimensional linear perceptron class is equal to d+1.

VC dimension measures the power of the learner (NOT necessarily equal to # of parameters)

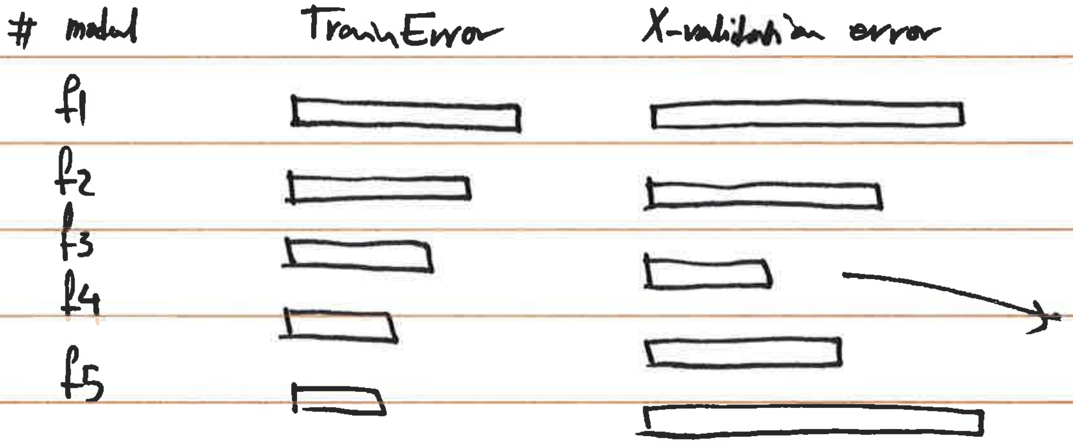
↓

so # of parameters does NOT necessarly equals to the complexity of the model

$(VC(1NN) = \infty)$

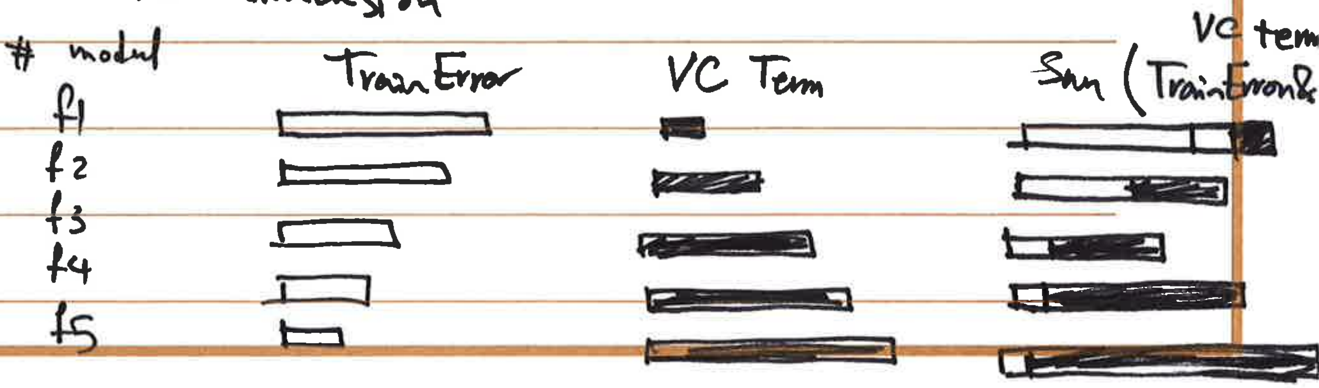
~~Trade off Analysis with VC:~~

Cross-Validation



best model is f3

VC dimension


# model
TrainError
VC Term
VC term
Sum (TrainError&
f1
f2
f3
f4
f5