PRML Summer School 2019

[Session 2 (Learning Theory): ML Basics]

Prof. Sungroh Yoon

ECE | Seoul National University

(last compiled at 23:37:00 on 2019/07/23)

Outline

References

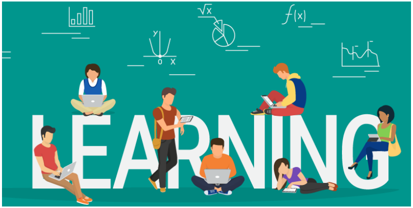
- books
  - *Learning from Data* by Abu-Mostafa et al.
  - *Pattern Recognition & Machine Learning* by Bishop
  - *Deep Learning* by Goodfellow, Bengio and Courville ▸ Link
- online resources:
  - *Deep Learning Specialization (coursera)* ▸ Link
  - *Stanford CS231n: CNN for Visual Recognition* ▸ Link
  - *Machine Learning Yearning* ▸ Link

Outline

Machine learning

- learning from ____
- what do we mean by learning?
  - Mitchell (1997):

    "A computer program is said to learn from experience E
    with respect to some class of tasks T and performance measure P,
    if its performance at tasks in T,
    as measured by P,
    improves with experience E."
- common types:
  - supervised
  - unsupervised
  - reinforcement
  - many more



Tasks in ML

- described in terms of how to process an **example**
- an "example":
  - a collection of **features** quantitatively measured from object/event
  - represented as a vector $\boldsymbol{x} \in \mathbb{R}^n$ (each entry x_i : a feature)

  e.g. features of an image: pixels values
- common ML tasks:

  T1. classification

  T2. classification with missing inputs

  T3. regression

  T4. transcription

  T5. machine translation

  T6. structured output

  T7. anomaly detection

  T8. synthesis and sampling

  T9. imputation of missing values

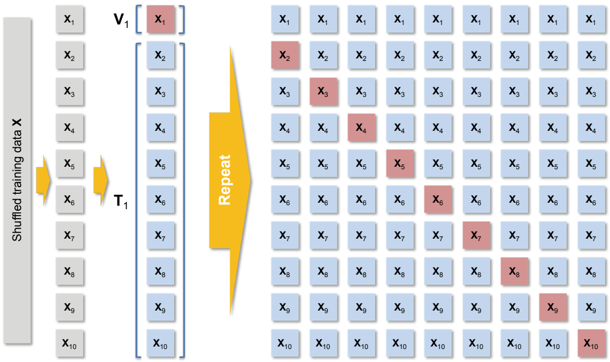
  T10. denoising

  T11. density/pmf estimation

Data set

- a collection of examples
  - training set: for fitting
  - validation set ("dev set"): for model selection
  - test set: for ___________

10-fold
cross-validation:



Performance measure

- specific to task T

  e.g. classification: accuracy, error rate E $\leftarrow$ we focus on this for a while

  density estimation: average log-probability the model assigns to examples

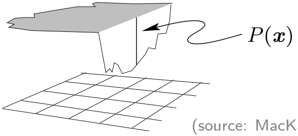
- evaluated using data sets
  - training/dev/test sets $\Rightarrow \mathrm{E}_{\text{train}}, \mathrm{E}_{\text{dev}}, \mathrm{E}_{\text{test}}$

- often challenging to choose
  1. difficult to decide what to measure

     e.g. penalize frequent mid-sized mistakes or rare large mistakes?
  2. know ideal measure but measurement is ________

     e.g. density estimation

a lake whose depth at $\boldsymbol{x} = (x, y)$ is $P(\boldsymbol{x})$



(source: MacKay)

Central challenge in ML

- $\boxed{}$
  - ability to perform well on previously unobserved examples
- generalization error $\mathrm{E}_{\mathrm{gen}}$
  - expected error on a new example $\Rightarrow$ implausible to calculate
- training error $\mathrm{E}_{\mathrm{train}}$
  - measured on a training set $\Rightarrow$ bad proxy for $\mathrm{E}_{\mathrm{gen}}$
- test error $\mathrm{E}_{\mathrm{test}}$
  - measured on a test set (not used in training) $\Rightarrow$ better proxy for $\mathrm{E}_{\mathrm{gen}}$

Two specific objectives

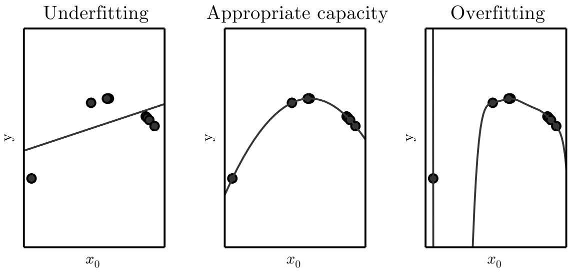
- objective: $\boxed{\mathrm{E}_{\mathrm{gen}} = 0}$ in theory or $\boxed{\mathrm{E}_{\mathrm{test}} \simeq 0}$ in practice
- split into two objectives:
  1. $\mathrm{E}_{\mathrm{test}} \simeq \mathrm{E}_{\mathrm{train}}$
  2. $\mathrm{E}_{\mathrm{train}} \simeq 0$
- objective 1: make $\mathrm{E}_{\mathrm{test}} \simeq \mathrm{E}_{\mathrm{train}}$
  - failure: ________ $\rightarrow$ high variance
  - cure: regularization, more data
- objective 2: make $\mathrm{E}_{\mathrm{train}} \simeq 0$
  - failure: underfitting $\rightarrow$ high bias
  - cure: optimization, more complex model

Capacity of a model

- the ability of the $\underbrace{\text{model}}_{\uparrow}$ to fit various functions

  representation (+ learning algorithm)

- altering capacity controls over/underfitting
  - example (truth: quadratic; fit: linear, quadratic, degree-9)



Choosing a model (conventional advice)

- Occam's razor (a principle of parsimony)
  - among competing hypotheses, choose the "______" one
- why? **VC generalization bound**: for any $\epsilon > 0$ and $N > 0$

$$\mathbb{P}[\ \underbrace{\overbrace{|\mathrm{E}_{\text{train}}(f) - \mathrm{E}_{\text{test}}(f)|}^{\text{generalization gap}} > \epsilon}_{\text{bad event}}\ ] \le \underbrace{4 \cdot (2N)^{\overbrace{\color{red}{d_{\mathrm{VC}}}}^{\text{capacity}}} \cdot e^{-\frac{1}{8}\epsilon^2 N}}_{\text{VC bound}}$$

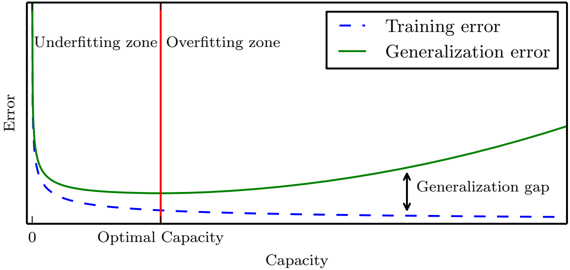
  - N : # of training examples
  - f : a model (d_{VC} : its *VC dimension*, a measure of model capacity)
- in words: discrepancy between $\mathrm{E}_{\text{train}}$ and E_{test}
  - grows as model capacity grows (but $\underbrace{\text{shrinks as } N \text{ increases}}_{\uparrow}$)

power of big data

A tradeoff: the main challenge in ML

- approximation-generalization tradeoff or bias-variance tradeoff

$$\underbrace{\mathrm{E_{test}} \simeq \overbrace{\mathrm{E_{train}}}^{\text{complex model is better}}}_{\text{simple model is better}} \simeq 0$$



- in theory: choose simpler functions
  - better generalization (smaller gap between training/test error)
- in practice: must still choose a sufficiently complex hypothesis
  - to achieve low training error

Two major weapons to fight the tradeoff

- **optimization**: ___ reduction (better approximation)
  - finds model parameters that minimize error

  e.g. stochastic gradient descent

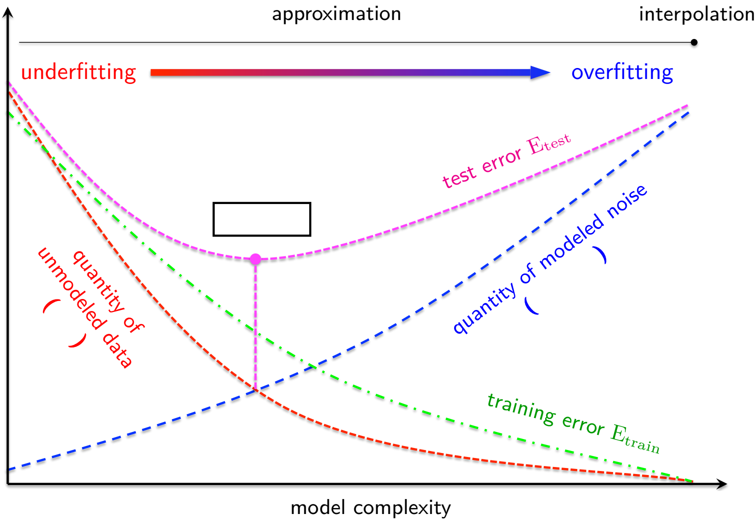
- **regularization**: _______ reduction (better generalization)
  - constrains model capacity by reflecting prior knowledge

  e.g. dropout, weight decay

Choosing a model (modern advice)

- $\boxed{\text{complex model + effective} \qquad\qquad\qquad \text{+ big data}}$
- complex model
  - higher chance of fitting data $\rightarrow \mathrm{E}_{\mathrm{train}} \simeq 0$
- regularization + big data
  - reduces generalization gap $\rightarrow \mathrm{E}_{\mathrm{test}} \simeq \mathrm{E}_{\mathrm{train}}$

Big picture



approximation
interpolation
underfitting
overfitting
test error E_{test}
quantity of unmodeled data
quantity of modeled noise
training error E_{train}
model complexity

Outline

Summary

- machine learning: learn from data to achieve generalization
  - objectives: making $\mathrm{E}_{\mathrm{test}} \simeq \mathrm{E}_{\mathrm{train}}$ + making $\mathrm{E}_{\mathrm{train}} \simeq 0$
  - challenge: approximation-generalization or bias-variance tradeoff
  - weapons: big data, optimization, regularization
  - example: linear models for classification/regression/prob estimation

PRML Summer School 2019

[Session 2 (Learning Theory): VC Analysis]

Prof. Sungroh Yoon

ECE | Seoul National University

Outline

Readings

- *Learning from Data* by Abu-Mostafa, Magdon-Ismail, and Lin
  - Chapter 2: Training versus Testing (Sections 2.1 & 2.2)

Recap

- questions on *why* and *how* machines can learn:

> 1. can we make sure that $\mathrm{E_{out}}(g) \approx \mathrm{E_{in}}(g)$?
> 2. can we make $\mathrm{E_{in}}(g)$ small enough?

- how the complexity of finite $\mathcal{H}$ affects learning:

	complex $\mathcal{H}$	simple $\mathcal{H}$	why?
Q1	☹	☺	$\mathbb{P}[\text{bad}] \leq 2M \cdots$
Q2	☺	☹	to fit training data $\mathcal{D}$

- choosing the right $\mathcal{H}$ is therefore critical
  - what if $M = \|\mathcal{H}\| = \infty$?

Today's plan

- we know machines can learn (for finite M) with enough data:

$$\mathbb{P}[\underbrace{|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon}_{\text{bad event}}] \leq \underbrace{2 \cdot M \cdot e^{-2\epsilon^2 N}}_{\text{small with large } N} \tag{1}$$

- can machines learn even when M is infinite?
  - yes, we will derive a new bound

$$\mathbb{P}[|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon] \leq 4 \cdot m_{\mathcal{H}}(2N) \cdot e^{-\frac{1}{8}\epsilon^2 N}$$

  where

$$m_{\mathcal{H}}(2N) \leq (2N)^{d_{\mathrm{VC}}}$$

- that is, we will find a ____ quantity that can replace ______ M
  - the growth function polynomially bounded by VC dimension

⇒ gives VC generalization bound

Outline

Key observation

- let $\mathcal{B}_m$ be the ($\mathcal{B}$ad) event

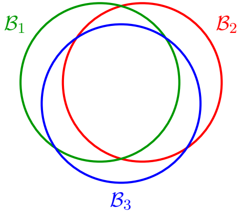
$$|\mathrm{E}_{\mathrm{in}}(h_m) - \mathrm{E}_{\mathrm{out}}(h_m)| > \epsilon$$

  - then

$$\mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \cdots \text{ or } \mathcal{B}_M] \leq \underbrace{\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]}_{\text{no overlaps: } M \text{ terms}}$$

  - this is how we got M in generalization bound

- the union bound becomes loose
  - if $\mathcal{B}_1, \ldots, \mathcal{B}_M$ strongly ______

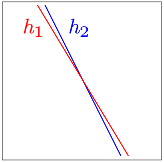


- typical learning model
  - many hypotheses: very ______
  - if $h_1 \approx h_2$, $\mathcal{B}_1$ and $\mathcal{B}_2$ are likely to coincide for most data

  $\Rightarrow$ $\mathcal{B}_m$'s do often strongly overlap

- ex) perceptron
  - if you slowly vary weight $\mathbf{w}$

  $\Rightarrow$ you will get infinitely many hypotheses that differ only infinitesimally



Overlap engineering

- theory of generalization hinges on the observation:
  - many hypotheses are indeed very similar

- idea:

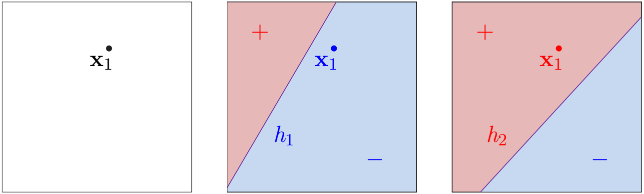
1. categorize similar hypotheses into m groups/types 2. regard m as the '_______' number of hypotheses 3. replace M with m in the bound, if m is finite

- how to group similar/overlapping hypotheses?

How many line types (seen by 1 point)?

hypothesis set $\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$

- how many lines in $\mathcal{H}$? $\boxed{}$
- how many types of lines does input point $\mathbf{x}_1$ see? $\boxed{}$

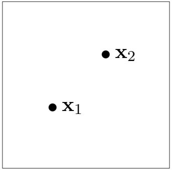


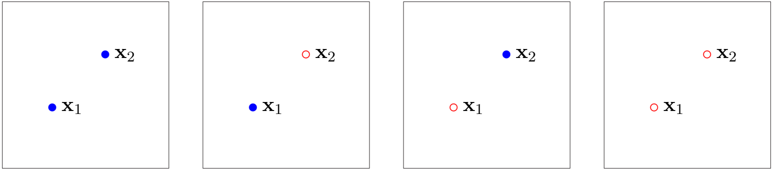
  - type 1: h_1-like lines that classify $\mathbf{x}_1$ as -1
  - type 2: h_2-like lines that classify $\mathbf{x}_1$ as $+1$

How many line types (seen by 2 points)?

$\boxed{\text{hypothesis set } \mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}}$

- for two input points $\mathbf{x}_1$ and $\mathbf{x}_2$? $\boxed{}$

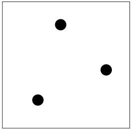




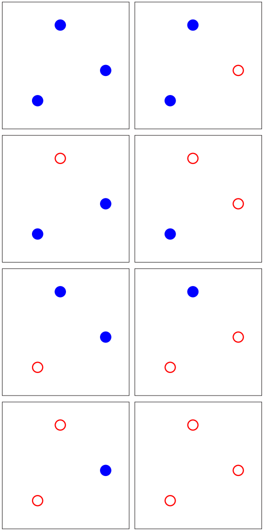
How many line types (seen by 3 points)?

$$\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$$

- for three input points?



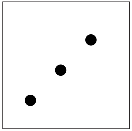
- $\boxed{}$ for this specific configuration
  - for **any** three inputs?



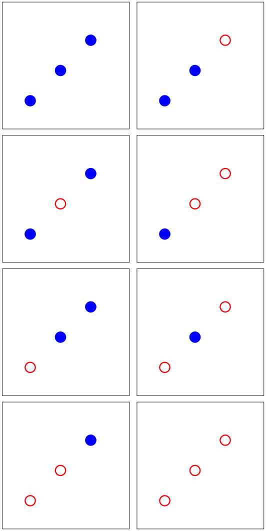
How many line types (seen by another 3 points)?

$$\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}$$

- how about these three?



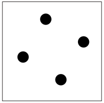
- $\boxed{}$ for this specific configuration
  - fewer than $8 = 2^3$
- at most $\boxed{}$ for any three inputs



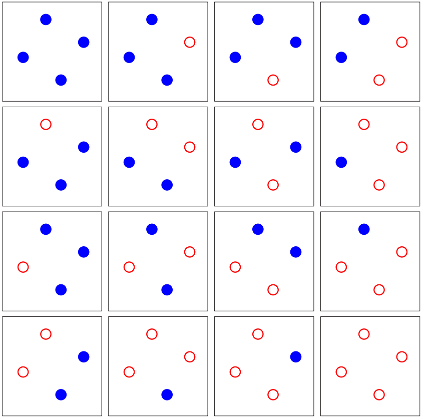
How many line types (seen by 4 points)?

$$\boxed{\mathcal{H} = \{\text{all lines in } \mathbb{R}^2\}}$$

- for four input points?



- at most $\boxed{\quad}$ for any four inputs
  - fewer than $16 = 2^4$



How many lines types (in general)?

- how many line 'groups' do N points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^2$ see?
  - according to previous examples $\longrightarrow$
  - this can be considered as the effective number of lines in $\mathcal{H}$
  - this number must be $\leq 2^N$ in any case (why?)

N	# line types
1	2
2	4
3	6, 8
4	14
5	22

- if this number is $\ll 2^N$ for sufficiently large N

  $\Rightarrow$ we can plug it into the bound (1) to replace M

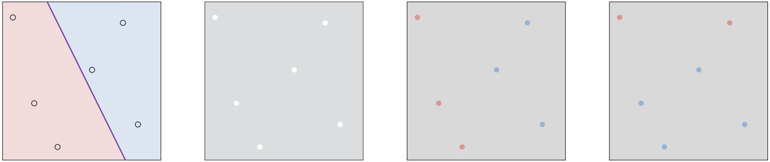
  $\Rightarrow$ ______ is feasible with infinite lines!

- let's formulate this idea

Outline

Concept

- assume a binary target function
  - each $h \in \mathcal{H}$ maps $\mathcal{X}$ to $\{-1, +1\}$
- instead of the whole input space $\mathcal{X}$
  - consider a _____ set of input points, and
  - count the number of *dichotomies* (________________)
- example ($N = 6$, perceptron): how many different dichotomies?



Dichotomy

- definition: given $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$

$$\underbrace{\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)}_{\substack{\uparrow \\ \boxed{\text{dichotomies}}\ \text{generated by } \mathcal{H} \text{ on these points}}} = \{\ (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N)) \mid h \in \mathcal{H}\} \qquad (2)$$

- dichotomies $\approx$ 'mini-hypotheses'
  - a set of hypotheses (just like $\mathcal{H}$)
  - these mini-hypotheses: seen through the ____ of N points only

Comparison

- domain
  - hypothesis $h : \mathcal{X} \to \{-1, +1\}$
  - dichotomy $h : \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \to \{-1, +1\}$

- diversity
  - the number of hypotheses $M = |\mathcal{H}|$: can be ______
  - the number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)|$: at most ___

- key point
  - $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)| \leq 2^N$ even for infinite $|\mathcal{H}|$

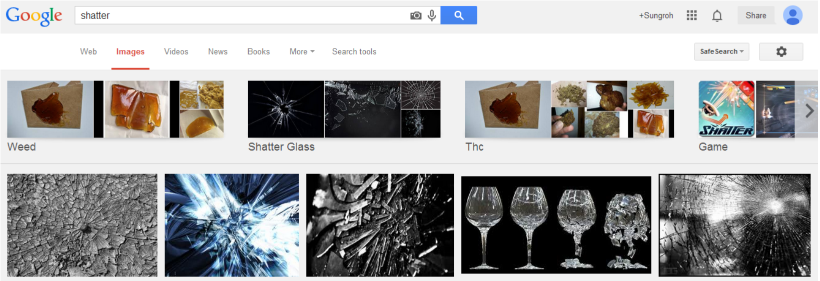
  $\Rightarrow$ candidate for replacing M

Shatter?

`v. shatter`

- to (make something) suddenly break into small pieces
- to destroy something completely (esp. feelings, hopes)

[from Oxford Advanced American Dictionary]



Definition

hypothesis set $\mathcal{H}$ can *shatter* $\mathbf{x}_1, \ldots, \mathbf{x}_N$

$\Leftrightarrow$ $\mathcal{H}$ can generate __________ dichotomies on $\mathbf{x}_1, \ldots, \mathbf{x}_N$

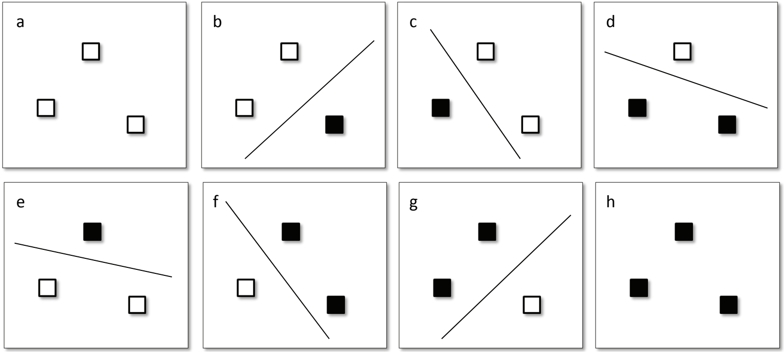
$\Leftrightarrow$ $\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \{-1, +1\}^N$

$\Leftrightarrow$ $|\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)| =$

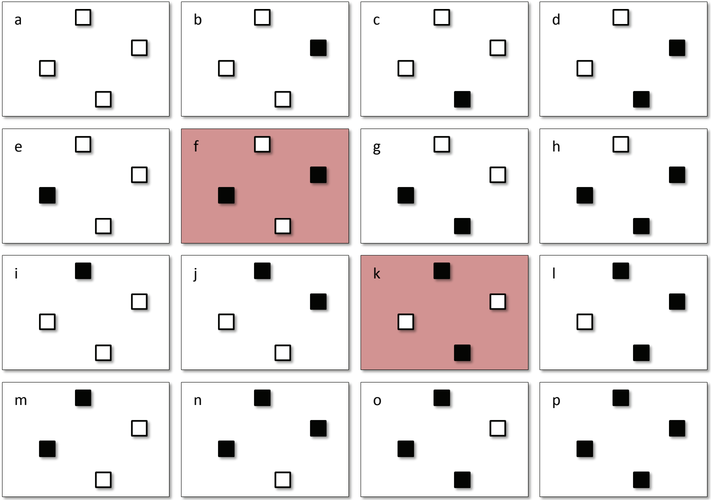
- this signifies that
  - $\mathcal{H}$ is as diverse as can be on the particular example
  - any learning problem definable by N examples can be learned with no training ____ by a hypothesis drawn from $\mathcal{H}$

Example

- $\mathcal{H}_1 = \{\text{lines in } \mathbb{R}^2\}$
  - can shatter ____ points in $\mathbb{R}^2$



- can $\mathcal{H}_1$ shatter four points in $\mathbb{R}^2$?



Outline

Generalization bound

- bounds $\mathrm{E}_{\mathrm{out}}$ in terms of E_{in}

  e.g. Hoeffding inequality

$$\mathbb{P}[|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} \tag{3}$$

  - equivalently

$$\mathrm{E}_{\mathrm{out}}(g) \leq \mathrm{E}_{\mathrm{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \tag{4}$$

    with probability $\geq 1 - \delta$ for a tolerance level δ (*e.g.* 0.05)

  - meaningless if M is ______

- key observation: infinitely many h's differ only infinitesimally
  - we can find something _____ that can replace infinite M

Growth function

- notation: $m_{\mathcal{H}}(N)$
  - the growth function of $\mathcal{H}$ on N points

- $m_{\mathcal{H}}(N)$ captures how different h's in $\mathcal{H}$ are

  $\Rightarrow$ gives effective # of h's

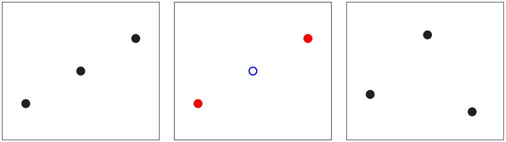
  $\Rightarrow$ can replace M in the bound (4)

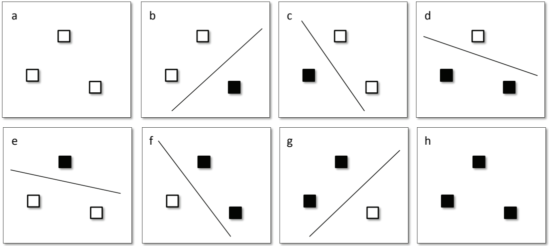
- definition
  - the max number of dichotomies $\mathcal{H}$ generates on N points

  $\Rightarrow$ $m_{\mathcal{H}}(N) \leq 2^N$

Example

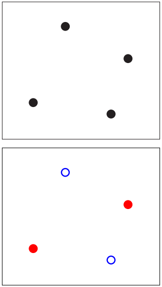
- $\mathcal{X}$: Euclidean plane $\mathbb{R}^2$, and $\mathcal{H}$: 2D perceptrons
- what is $m_{\mathcal{H}}(3)$? ans: $m_{\mathcal{H}}(3) = \_\_$

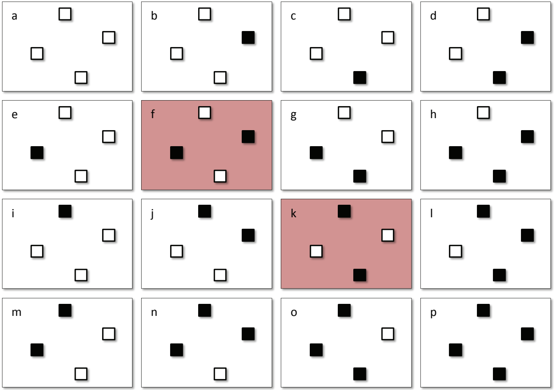




Example

- $\mathcal{X}$: Euclidean plane $\mathbb{R}^2$, and $\mathcal{H}$: 2D perceptrons
- how about $m_{\mathcal{H}}(4)$? ans: $m_{\mathcal{H}}(4) =$ ____





Summary so far

- we have tried to replace

$$\mathrm{E}_{\text{out}}(g) \leq \mathrm{E}_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

  - with

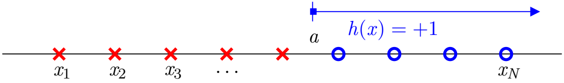
$$\mathrm{E}_{\text{out}}(g) \overset{?}{\leq} \mathrm{E}_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}$$

- key for learning: having $\mathcal{H}$ with $\boxed{\text{polynomial } m_{\mathcal{H}}(N)}$
  - why?

Example: positive rays

- $\mathcal{H} = \{h \,|\, h(x) = \text{sign}(x - a), x \in \mathbb{R}\}$

  i.e. -1 to the left of some a and $+1$ to the right of a



- given N points
  - line: split into $N + 1$ regions
  - dichotomy on N points: decided by which region has a

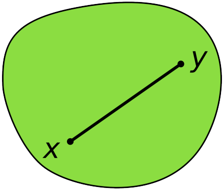
location of a	x_1	x_2	x_3	x_4
$-\infty < a < x_1$	○	○	○	○
$x_1 < a < x_2$	×	○	○	○
$x_2 < a < x_3$	×	×	○	○
$x_3 < a < x_4$				
$x_4 < a < \infty$	×	×	×	×

- thus, $m_{\mathcal{H}}(N) =$ _____ $\ll 2^N$ for sufficiently large N

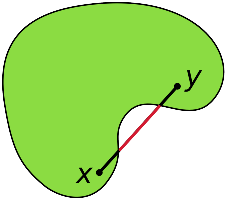
Example: convex sets

- $\mathcal{H}$ consists of all hypotheses in 2D $h : \mathbb{R}^2 \to \{-1, +1\}$
  - that are positive inside a convex set and negative elsewhere

- a set is convex
  - if the line segment connecting any two points in the set lies entirely _____ the set



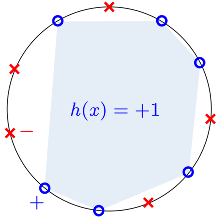
(a) convex set



(b) non-convex set

- let's choose N points on the perimeter of a circle
    - consider any dichotomy on these points by assigning an arbitrary pattern of ± 1's to them
- observe:
    - the polygon formed by connecting $+1$'s: always a _________
    - no matter how you assign ± 1's, you can always separate $+$'s and $-$'s perfectly

$\Rightarrow$ $\mathcal{H}$ manages to shatter these points



- therefore: $m_{\mathcal{H}}(N) =$ ___

Checkpoint

- example growth functions
  - positive rays $m_{\mathcal{H}}(N) = N + 1$
  - convex sets $m_{\mathcal{H}}(N) = 2^N$
  - 2D perceptrons $m_{\mathcal{H}}(N) < 2^N$ for $N > 2$
- what if $m_{\mathcal{H}}(N)$ replace M in the generalization bound?

$$\mathbb{P}[|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon] \leq 2 \cdot m_{\mathcal{H}}(N) \cdot e^{-2\epsilon^2 N}$$

  - GOOD if $m_{\mathcal{H}}(N)$ is __________ in N
  - BAD if $m_{\mathcal{H}}(N)$ is __________ in N
- computing $m_{\mathcal{H}}(N)$ is not trivial $\Rightarrow$ any alternative?

Challenge & solution

- it is not practical to compute $m_{\mathcal{H}}(N)$ for every $\mathcal{H}$ we use
  - fortunately, we don't have to
- our approach: find a $\boxed{\text{polynomial bound on } m_{\mathcal{H}}(N)}$

  to show $m_{\mathcal{H}}(N)$ is polynomial

  we show $m_{\mathcal{H}}(N) \leq \cdots \leq \cdots \leq$ a __________
- getting a good bound on $m_{\mathcal{H}}(N)$
  - will be much easier than computing $m_{\mathcal{H}}(N)$ itself
  - thanks to the notion of a *break point*

Outline

Concept

- if the condition $m_{\mathcal{H}}(N) = 2^N$ breaks at any point k
  i.e. $m_{\mathcal{H}}(k) < 2^k$ and $\mathcal{H}$ cannot shatter k examples

- then we can bound $m_{\mathcal{H}}(N)$ by a simple polynomial of N
  - this bound is based on break point k
  - spoiler: $m_{\mathcal{H}}(N) = O(\qquad)$

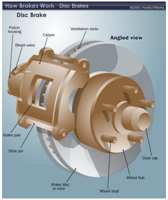
Definition

> if no data set of size k can be shattered by $\mathcal{H}$
> $\Rightarrow$ k is said to be a *break point* for $\mathcal{H}$

- for any break point k, $m_{\mathcal{H}}(k) < 2^k$

  i.e. 'brake' for shattering

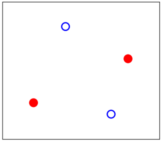
  - $k+1, k+2, \ldots$ are also break points
  - we focus on the ________ break point



- in general, a break point k is easier to find than $m_{\mathcal{H}}(N)$

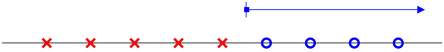
  e.g. _____ for 2D perceptron

  - a bigger data set cannot be shattered either



Examples

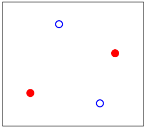
- positive rays: $m_{\mathcal{H}}(N) = N + 1$



  - break point
  - $m_{\mathcal{H}}(2) = 3 < 2^2$

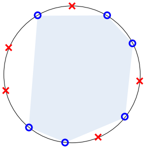
  o ×

- 2D perceptron: $m_{\mathcal{H}}(N) < 2^N$



  - break point
  - $m_{\mathcal{H}}(4) = 14 < 2^4$

- convex sets: $m_{\mathcal{H}}(N) = 2^N$



  - break point
    (*i.e.* no break point)

Key fact

- theorem (see textbook for proof):
  - if $\underbrace{m_{\mathcal{H}}(k) < 2^k}_{k\text{: break point}}$ for some $k \Longrightarrow m_{\mathcal{H}}(N) \le \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\substack{\text{polynomial in } N \\ \text{of degree } k-1}}, \forall N$

- in words
  - if $\mathcal{H}$ has a __________ $\Rightarrow$ polynomial bound on $m_{\mathcal{H}}(N)$ exists

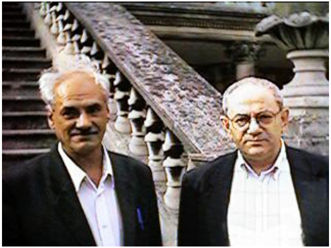
  $\Rightarrow$ we have what we want to ensure good generalization

- the degree of the polynomial bound on $m_{\mathcal{H}}(N)$
  - $k-1 \Rightarrow$ called ____________

Outline

Vapnik-Chervonenkis (VC) dimension

- formal name of the __________________ point
  - a single parameter that characterizes the growth function
  - measures the capacity of a learning algorithm



Alexey Chervonenkis and Vladimir Vapnik

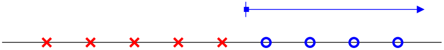
- $d_{\mathrm{VC}}(\mathcal{H})$: the VC dimension of $\mathcal{H}$
  - the largest N that $\mathcal{H}$ can ______

  i.e. the largest N for which $m_{\mathcal{H}}(N) = 2^N$
  - if $m_{\mathcal{H}}(N) = 2^N$ for all $N \Rightarrow d_{\mathrm{VC}}(\mathcal{H}) \triangleq \infty$

- property:
  - $k = d_{\mathrm{VC}} + 1$: the minimum break point for $m_{\mathcal{H}}$
  - d_{VC}: the order of the polynomial bound on $m_{\mathcal{H}}(N)$
  - the polynomial bound on $m_{\mathcal{H}}(N)$: $m_{\mathcal{H}}(N) \leq$ _____

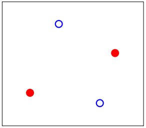
Examples

- positive rays: $m_{\mathcal{H}}(N) = N + 1$



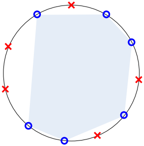
  - break point $k = 2$
  - $d_{\mathrm{VC}} = 1$

- 2D perceptron: $m_{\mathcal{H}}(N) \leq N^3$



  - break point $k = 4$
  - $\boxed{d_{\mathrm{VC}} = 3}$

- convex sets: $m_{\mathcal{H}}(N) = 2^N$



  - break point $k = \infty$
  - $\boxed{d_{\mathrm{VC}} = \infty}$

General idea: good vs bad models

$\boxed{\mathrm{E_{out}} \overset{?}{\leq} \mathrm{E_{in}} + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}}$

- good models: $\boxed{}$

  $\Rightarrow$ $m_{\mathcal{H}}(N)$ is bounded by a polynomial in N

  $\Rightarrow$ the term $\ln m_{\mathcal{H}}(N)$ grows logarithmically in N

  $\Rightarrow$ so it will be crushed by the $\frac{1}{N}$ factor

  $\Rightarrow$ for any fixed tolerance δ, the bound on $\mathrm{E_{out}}$ will be arbitrarily close to $\mathrm{E_{in}}$ for sufficiently large N

  $\Rightarrow$ $\mathrm{E_{out}} \approx \mathrm{E_{in}}$ for sufficiently large N ($\mathrm{E_{in}}$ "generalizes" to $\mathrm{E_{out}}$)

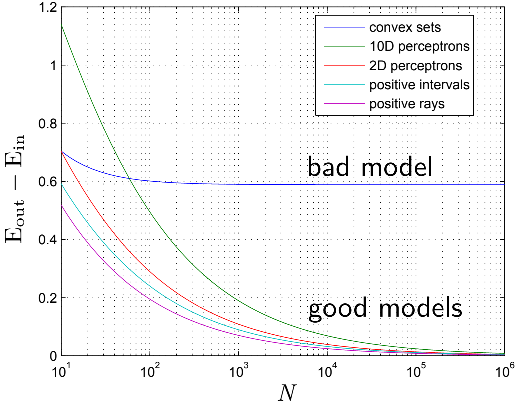
- bad models: $\boxed{}$

  $\Rightarrow$ the above arguments will all fail

  $\Rightarrow$ no matter how large data set is, cannot make generalization conclusion from $\mathrm{E_{in}}$ to $\mathrm{E_{out}}$ based on VC analysis

Example: good versus bad models

- generalization performance ($\delta = 0.1$)



$\mathcal{H}$	d_{VC}
convex sets	∞
10D perceptrons	11
2D perceptrons	3
positive rays	1

error bar used: $\sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}$; for the perceptrons, additional bound used: $m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$

The VC generalization bound

for any tolerance $\delta > 0$

$$\mathrm{E}_{\mathrm{out}} \leq \mathrm{E}_{\mathrm{in}} + \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}} \tag{5}$$

with probability $\geq 1 - \delta$

- the most important mathematical result in theory of learning
- holds for any binary target function f, any hypothesis set $\mathcal{H}$, any learning algorithm $\mathcal{A}$, and any input prob. distribution P

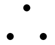
- meaning: if $d_{\mathrm{VC}}(\mathcal{H}) \neq \infty$ (*i.e.* $\mathcal{H}$ has a ____ VC dimension)
  $\Rightarrow$ with enough data $(N \to \infty)$, *each and every* hypothesis h (even in an infinite $\mathcal{H}$) will ________ well from E_{in} to $\mathrm{E}_{\mathrm{out}}$

Outline

VC dimension versus # parameters

- 1-dim perceptron
  - $d_{\mathrm{VC}} = 2$

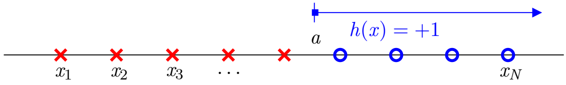
- 2-dim perceptron
  - $d_{\mathrm{VC}} = 3$



- d-dim perceptron
  - $d_{\mathrm{VC}} = d + 1$

- what is # of parameters of d-dim perceptron?

- positive rays ($d_{\mathrm{VC}} = 1$):



- however, this is not always the case in general
  - is there any physical intuition behind d_{VC}?

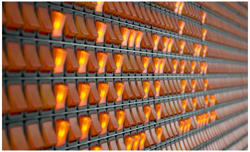
Interpreting d_{VC}

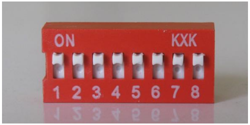
- parameters create '________________' (DOF)
  - the more parameter a model has, the more diverse its $\mathcal{H}$ is
- perceptron: $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$
  - parameters: $w_0, w_1, \ldots, w_d \Rightarrow d+1$ in total
- in other models (*e.g.* multi-layer perceptrons)
  - some parameters may not directly contribute to DOF

  $\Rightarrow$ effective parameters may be less obvious or implicit
- d_{VC} measures these $\boxed{\text{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of parameters}}$ or DOF

Degrees of freedom (DOF)

- hypothesis parameters $\mathbf{w} = (w_0, \ldots, w_d)$
  - creates degrees of freedom
- # hypotheses $M = |\mathcal{H}|$
  - 'continuous' degrees of freedom
- # dichotomies reflected in $m_{\mathcal{H}}(N)$
  - 'binary' degrees of freedom
- VC dimension d_{VC}
  - ______________ degrees of freedom





Outline

Decomposing VC generalization bound

- two parts make up the bound (5):

$$\mathrm{E}_{\text{out}} \leq \underbrace{\mathrm{E}_{\text{in}}}_{\text{1st part}} + \underbrace{\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}}_{\text{2nd part}} \tag{5}$$

- second part: increases as ___ increases

$$\mathrm{E}_{\text{out}}(g) \leq \mathrm{E}_{\text{in}}(g) + \underbrace{\Omega(N, \mathcal{H}, \delta)}_{\uparrow} \tag{6}$$

$$\Omega(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N}\ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)} \leq \sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{VC}}}+1)}{\delta}\right)}$$

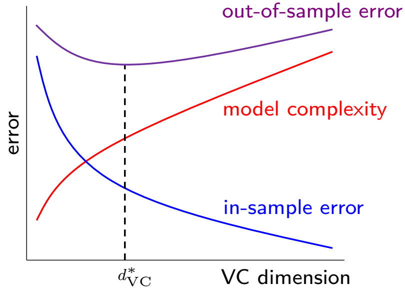
Interpreting $\Omega(N, \mathcal{H}, \delta)$ as penalty for model complexity

with probability $\geq 1 - \delta$, $\quad \mathrm{E_{out}} \leq \mathrm{E_{in}} + \underbrace{\Omega(N, \mathcal{H}, \delta)}_{=\sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}}$

- penalty $\Omega(N, \mathcal{H}, \delta)$ gets worse ($\Rightarrow$ worse bound on $\mathrm{E_{out}}$) if
  - we have a smaller training set
  - we use a more complex $\mathcal{H}$ (_____ d_{VC})
  - we insist on higher confidence (_____ δ)
- penalty $\Omega(N, \mathcal{H}, \delta)$ gets better if
  - we have more training examples
  - we use a simpler model
  - we want lower confidence (higher δ)

Tradeoff

model complexity ↑	d_{VC} ↑	$\Rightarrow$	E_{in} ↓	but	Ω ↑	and	$\mathrm{E}_{\mathrm{out}} - \mathrm{E}_{\mathrm{in}}$ ↑
model complexity ↓	d_{VC} ↓	$\Rightarrow$	Ω ↓	but	E_{in} ↑		



using powerful $\mathcal{H}$ is not always good!

- regularization: instead of using E_{in} as proxy for $\mathrm{E}_{\mathrm{out}}$
  - use ___ and _ together (*i.e.* augmented error $\mathrm{E}_{\mathrm{aug}}$)

Outline

Alternative #1: test-set based approach

- VC analysis
  - we do not know g (the best $h \in \mathcal{H}$) in advance

  $\Rightarrow$ should consider all cases by using the union bound

$$\mathbb{P}[|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon] \leq 2M e^{-2\epsilon^2 N}$$

  $\Rightarrow$ infinite M issue $\Rightarrow$ (all the hassles) $\Rightarrow$ VC bound

- $\mathrm{E}_{\mathrm{test}}$ approach
  - g is fixed by training before we compute $\mathrm{E}_{\mathrm{test}}$ (*i.e.* ______)

  $\Rightarrow$ can use the single inequality

$$\mathbb{P}[|\mathrm{E}_{\mathrm{in}}(g) - \mathrm{E}_{\mathrm{out}}(g)| > \epsilon] \leq 2 e^{-2\epsilon^2 N}$$

  $\Rightarrow$ much tighter than VC bound

Test set versus training set

- common: both are finite samples
  - normally have some variance due to sample size

- a training set has an __________ bias in its estimate of $\mathrm{E}_{\mathrm{out}}$
  - $\because$ it was used to choose a hypothesis that looked good *on it*
  - VC bound implicitly considers that bias $\Rightarrow$ huge error bar

- a test set has no optimistic/pessimistic bias
  - $\Rightarrow$ when you report $\mathrm{E}_{\mathrm{test}}$ to customers and they try on new data
  - mostl likely: not surprised ($\because$ generalization of $\mathrm{E}_{\mathrm{test}}$)

$$\boxed{\mathrm{E}_{\mathrm{out}}(g) \underbrace{\approx}_{} \mathrm{E}_{\mathrm{in}}(g) \underbrace{\approx}_{} 0}$$

Alternative #2: bias-variance analysis

- VC analysis: based on binary target functions, but
  - can be extended to real-valued functions
  - as well as to other types of functions

- proofs in those cases: quite technical

  $\Rightarrow$ no addition to insight VC analysis of binary functions provides

- an alternative approach for real-valued functions
  - ____________ analysis: $\boxed{\mathrm{E}_{\mathrm{out}} = \text{bias} + \text{variance}}$
  - provides new insights into generalization

Outline

Summary

$$\boxed{E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \quad (4)}$$

if $\mathcal{H}$ has	no break point	any break point
$m_{\mathcal{H}}(N)$	2^N	polynomial in N
if $m_{\mathcal{H}}(N)$ replaced M in inequality (4)	$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \nrightarrow 0$ regardless of N	$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \rightarrow 0$ as $N \rightarrow \infty$
generalize well?	no	yes
example	convex set	perceptron

- $d_{\mathrm{VC}}(\mathcal{H})$, VC dimension of $\mathcal{H}$: the most points $\mathcal{H}$ can shatter
  - definition: the largest non-break point (minimum $k-1$)
  - example: $d_{\mathrm{VC}} = d+1$ for d-dimensional perceptron
  - physical intuition: $d_{\mathrm{VC}} \approx \#$ (effective) parameters
  - utility: estimating model complexity & sample complexity
  - rule of thumb: $N \geq 10 \times d_{\mathrm{VC}}$ for decent generalization
  - generalization bound: $\mathrm{E}_{\mathrm{out}} \leq \mathrm{E}_{\mathrm{in}} + \Omega(N, \mathcal{H}, \delta)$
  - bottom line: models with lower d_{VC} tend to generalize better

- alternatives to VC analysis
  - test set: use $\mathrm{E}_{\mathrm{test}}$ as proxy for $\mathrm{E}_{\mathrm{out}}$ (tighter than VC bound)
  - bias-variance analysis: for real-valued targets