Learning and VC Dimension

Chen Huang

2015/05/21

OutLine

Learning

Learning and Loss

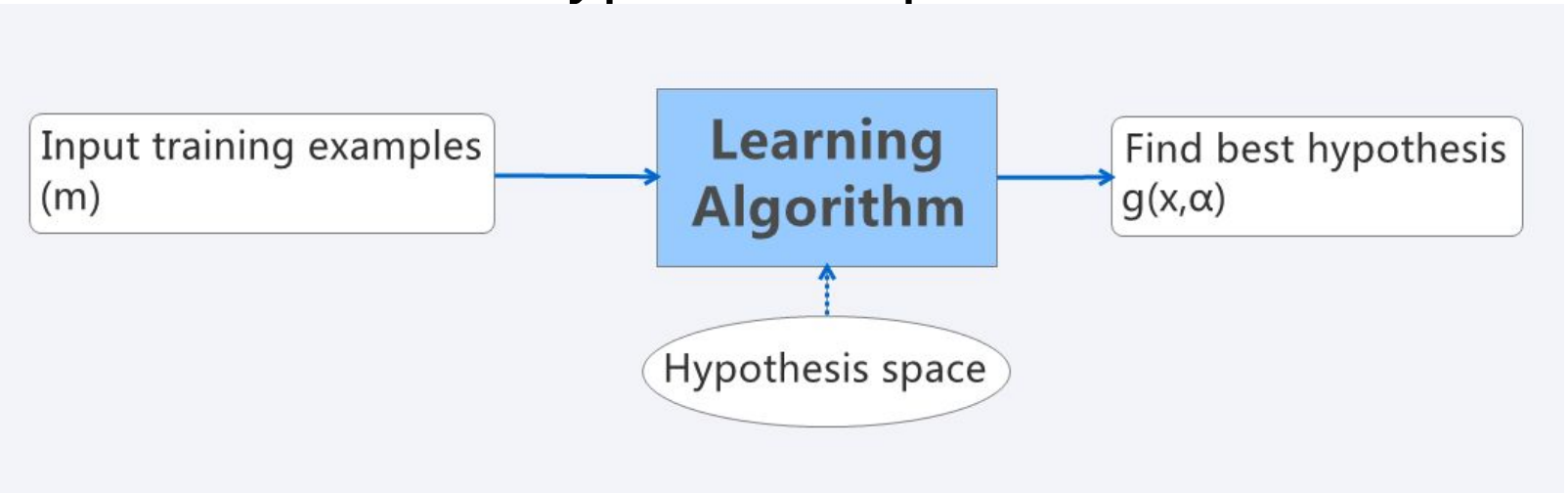
VC Dimension

VC

Model Selection

What's Learning

Learning is about applying learning algorithm to training examples *m*, which are from the whole data *D=f(x)*, in order to find a **ideal** hypothesis *g* from the hypothesis space *H*.



How to define this ideal

- We want this learned g∈H to work well on future data.
  Define *Error(h) = l(h,x,y)* to be the loss function, and expirical loss $E_{train}(h)$,expected loss $E_{true}(h)$, we want :

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

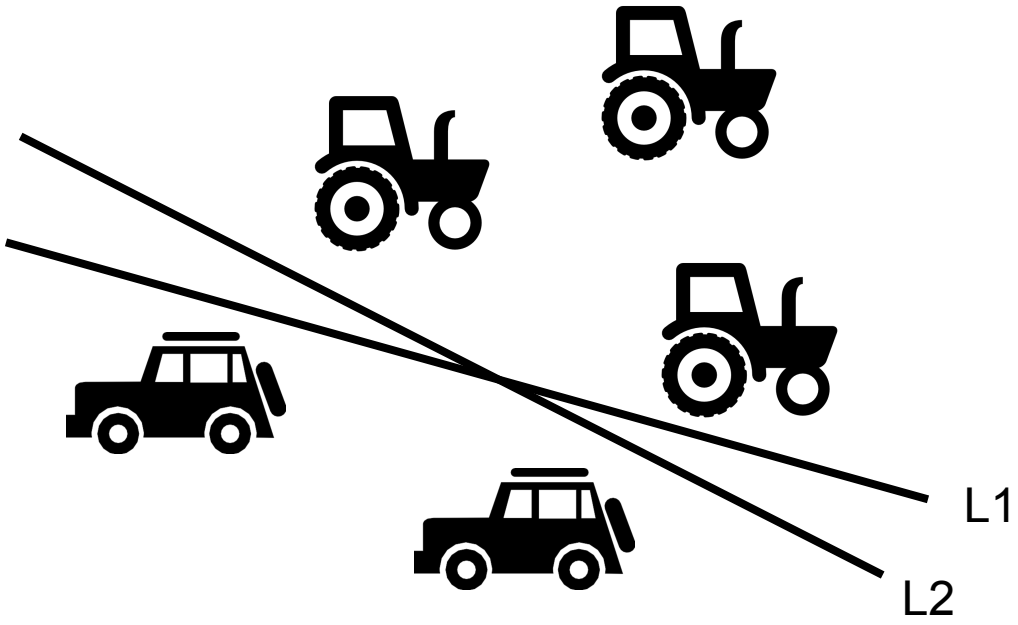
$$E_{train}(h) = \mathrm{E}_{x\in \mathrm{m},h\in \mathrm{H}}[l(h,x,y)] = \frac{1}{|m|}\sum_{x\in m,h\in H}\delta(h(x)\neq y)$$

$$E_{true}(h) = \mathrm{E}[l(h,X,Y)]$$

Note : For a fixed h, train error is likely to be an underestimate to true error.

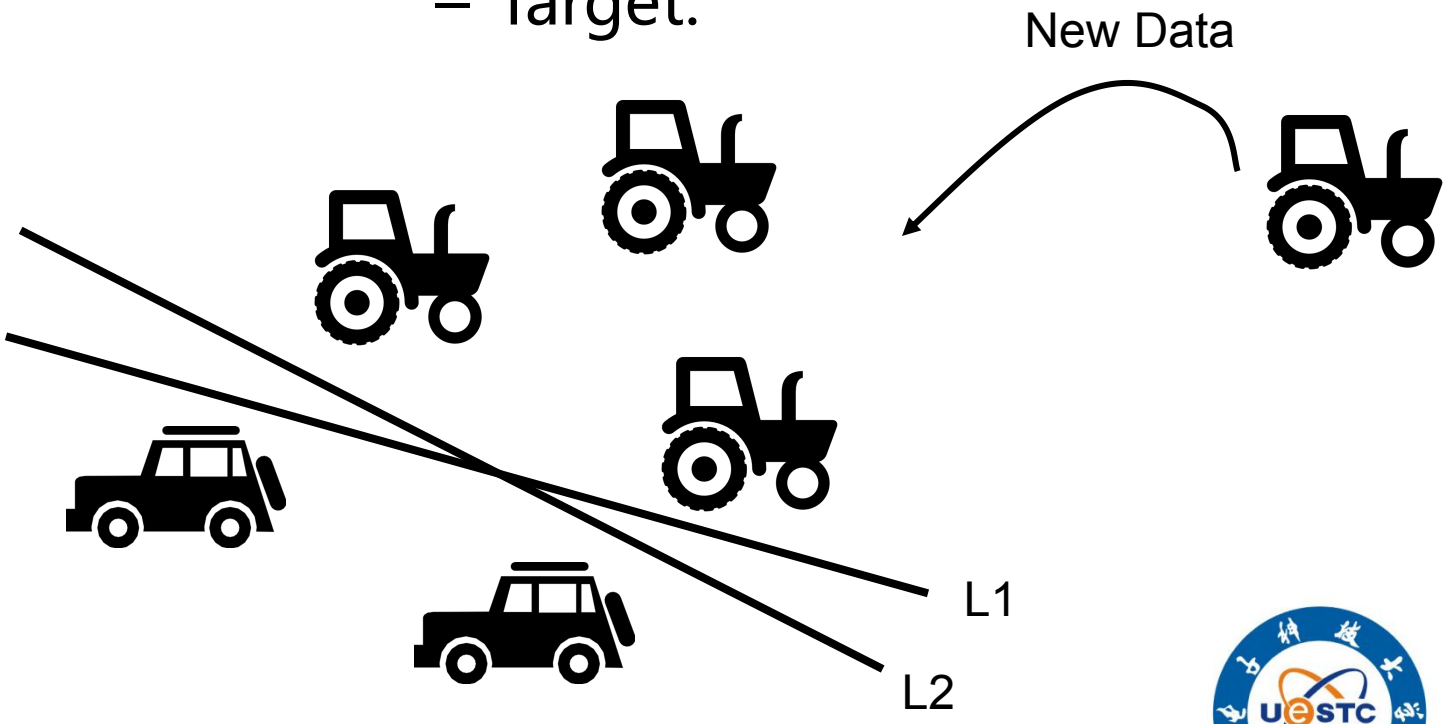
Learning Example

Suppose we want an algorithm to distinguish among different types of motor vehicles such as cars and tractors. And the objective is to design a "prediction" algorithm that given a vector will correctly predict the corresponding type of vehicle



What we got

- The number of training sample : m
  - Hypothesis space H : lots of lines
    - Target:



What we got

- The number of training sample : m
- Hypothesis space H : lots of lines
- Target:

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

So what is the probability of $E_{train}(g)$ close enough to $E_{true}(g)$?

Hoeffding Inequality

Let mutually independent random variables ξ1,...ξN , N is a large number, and define:

$$\bar{\xi} = \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

So for any ε > 0 , we have

$$P\left(\bar{\xi} - E[\bar{\xi}] > \varepsilon\right) \le \exp\left(-2N\varepsilon^2\right)$$

Hoeffding shows the difference between the true probability of an event and the observation of its independent trials

Learning Example

According to Hoeffding, for a fixed h, we have

$$\mathrm{P}[E_{true}(h) - E_{train}(h) > \varepsilon] \leq \exp\left(-2\mathrm{m}\varepsilon^2\right)$$

where m is the size of training example and it should be a large number

How large?

How many training example will suffice

$$\mathrm{P}[E_{true}(h) - E_{train}(h) > \varepsilon] \le \exp\left(-2\mathrm{m}\varepsilon^2\right)$$

Let's say we want $E_{true}(g) \le E_{train}(g) + \varepsilon$ holds with the confidence of at least $1-\delta$

Then according to other formulas, we have the answer:

$$\mathrm{m} \ge \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

$$\mathrm{E}_{true}(h) \le \mathrm{E}_{train}(h) + \varepsilon = \mathrm{E}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

the Upper Bound

Now we have, for a fixed h,

$$\mathrm{P}[E_{true}(h) - E_{train}(h) > \varepsilon] \le \exp\left(-2\mathrm{m}\varepsilon^2\right)$$

As for the whole hypothesis space H

$$\mathrm{P}[E_{true}(h_1) - E_{train}(h_1) > \varepsilon \cup .... \cup E_{true}(h_{|H|}) - E_{train}(h_{|H|}) > \varepsilon]$$

$$\le \mathrm{P}[E_{true}(h_1) - E_{train}(h_1) > \varepsilon] + ... + \mathrm{P}[E_{true}(h_{|H|}) - E_{train}(h_{|H|}) > \varepsilon]$$

$$\le |H| \exp\left(-2\mathrm{m}\varepsilon^2\right)$$

Learning Example

- Abviously, the upper bounds for hypothesis space H is

$$|H|\exp(-2m\varepsilon^2)$$

- So the answer to our target:

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

- $E_{train}(g)$ can close enough to $E_{true}(g)$ if m is large and H is finite

- $E_{train}(g)$ can close enough to zero if H is reasonable

Infinite Hypothesis Space

- However, the line to distinguish among different types of motor vehicles is infinite

- What measure of complexity should we use in place of |H| ?

Vapnik-Chervonenkis Dimension

Shatter and VC dimension

- 对于一个系统(U, S), U是一个集合，S是U的子集的集合。

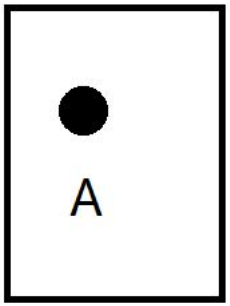
- 如果样本A $\subseteq$ U，A中的每一个子集都可以表示为S中的一个元素与A的交集，则称A可以被S打散（Shatter）

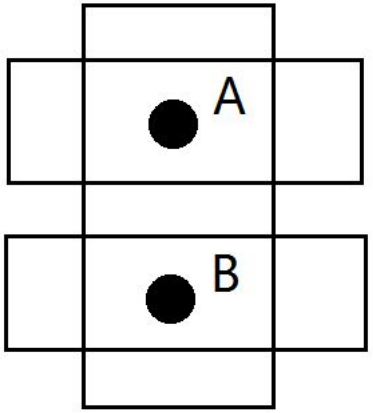
- 假设空间H的VC维，VC(H)，定义为：
  VC(H) = max(|A|), A是可以被H打散的最大样本集合

太抽象！

Example

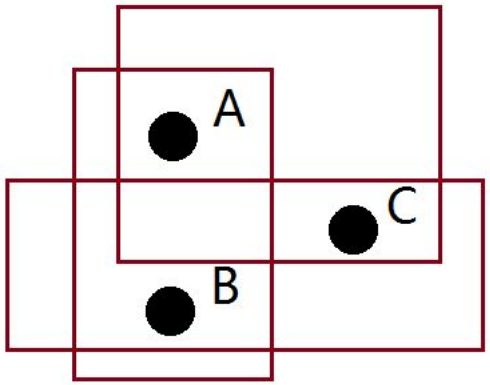
- U = R^2 of points in the plane,
- S = the collection of all axis-parallel rectangles.
- When m = 1 and m=2



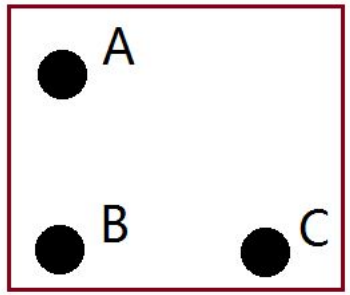


Example

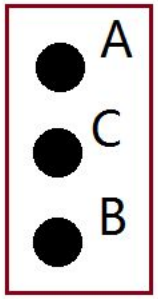
When m = 3, There are many ways to place 3 points



Shattered



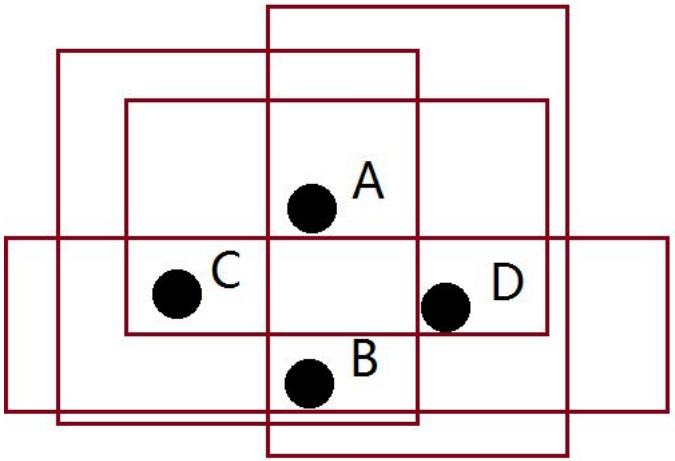
Choose AC failed



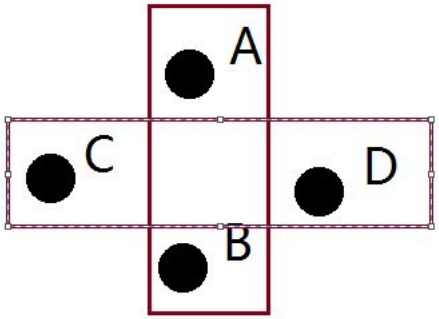
Choose AB failed

Example

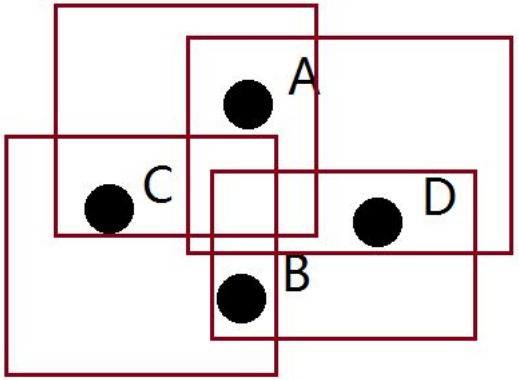
When m = 4, shattered



Choose three

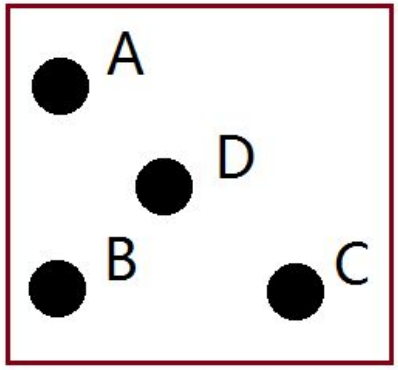


Choose two

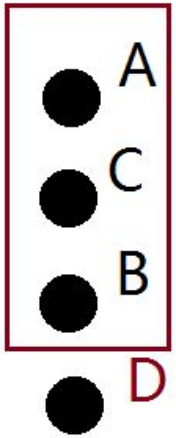


Example

When m = 4, not shattered



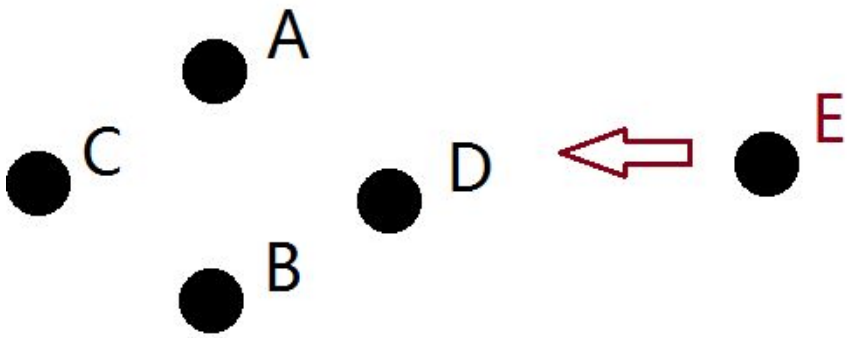
Choose ABC failed



Choose AB failed

Example

When any m > 4, not shattered



- 5 points in line : NO
- Put E inside ABCD : NO
- Convex Polygon(凸边形) : NO
- .....

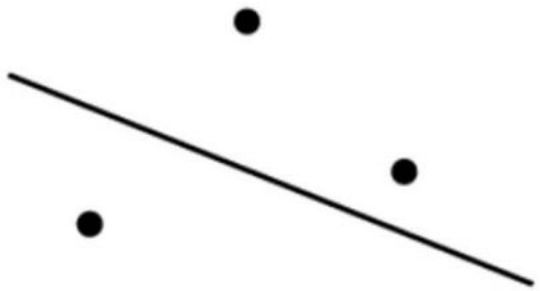
Example

- U = R^2 of points in the plane,
- S = the collection of all axis-parallel rectangles.

- VC = 4

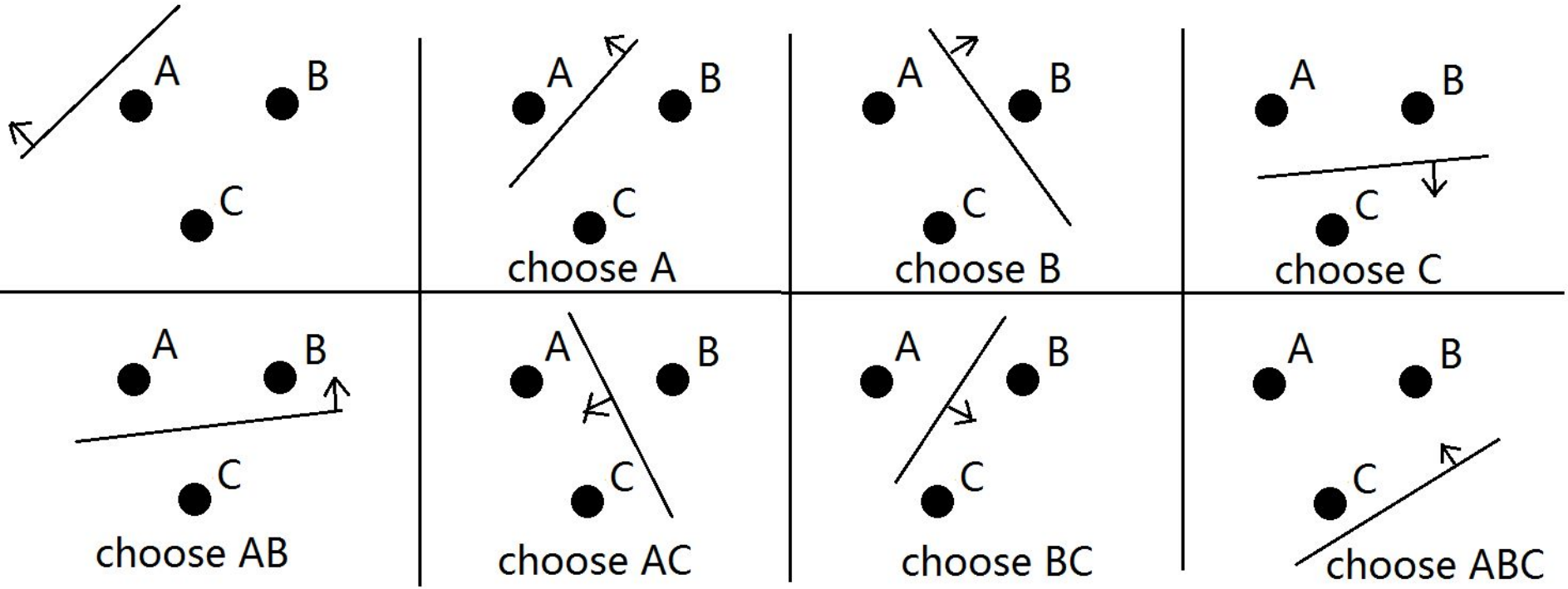
Example 2

- A set system is the set U of points in the plane, with S linear separating hyperplanes in n dimension.
- VC(H) = n + 1



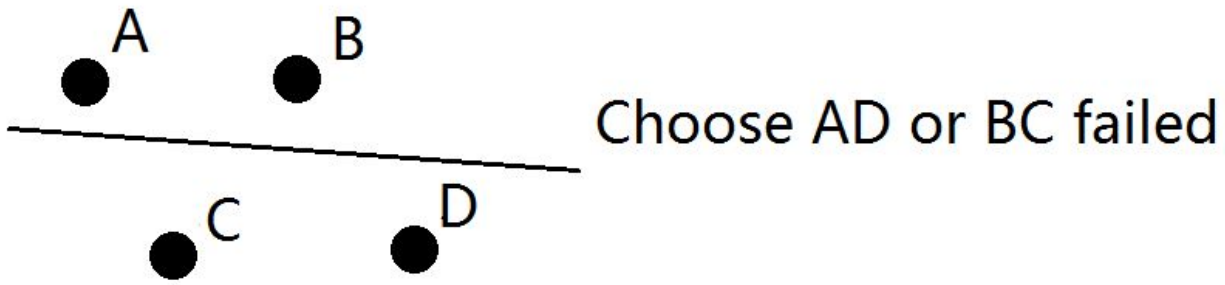
Example 2

- U =R^2 of points in the plane,
- S = linear separating hyperplanes in 2 dimension.
- m = 3



Example 2

When m = 4:



Any m >4 is not shattered, VC(H) = 3

Example 2

- Known that VC(Linear) = 3 and VC(Axis-parallel Rectangle) = 4

- To separate training example of size 4:

- E_{train}(Linear) > E_{train}(Rectangle)

- More likely E_{train}(Rectangle) ≈ 0

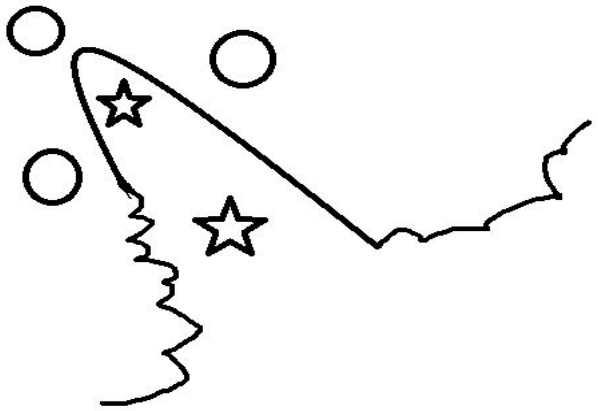
So the smaller VC(H), the harder to find a hypothesis h∈H , E_{train}(h) ≈ 0

the Meaning of VC Dimension

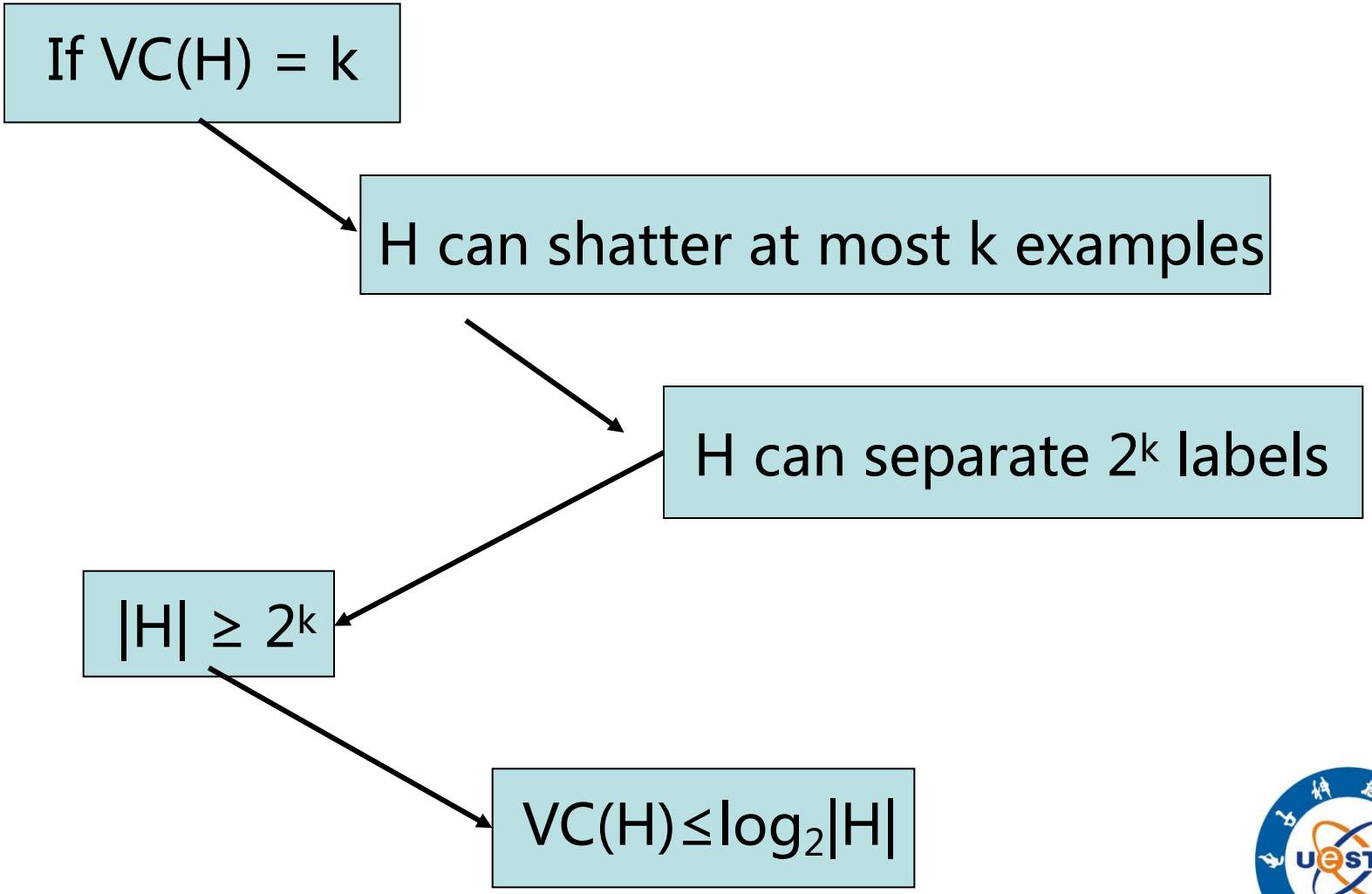
- VC(H) is a measure of **complexity** and measures the expressive **power or flexibility** of a set of functions(or hypothesis space) by assessing how wiggly(扭动的,起伏的) its members can be.

- The bigger VC(H) shows H can shatter more point.

- VC(H) is infinite, if H can shatter any n examples, then, model is very complicated



VC(H) and |H|


If VC(H) = k
H can shatter at most k examples
H can separate 2^k labels
$|H| \geq 2^k$
$VC(H) \leq \log_2|H|$

VC Bound

- Let's review the problem left before and replace |H| with VC(H):
  - Bound on m using other complex quantities

$$\mathrm{m} \geq \frac{1}{\varepsilon}\left(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\varepsilon}\right)$$

  - Bound on E using other complex quantities

$$\mathrm{E}_{true}(h) < \mathrm{E}_{train}(h) + \varepsilon = \mathrm{E}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln\frac{2m}{VC(H)} + 1\right) + \ln\frac{4}{\delta}}{m}}$$

VC Bound

$$\mathrm{E}_{true}(h) < \mathrm{E}_{train}(h) + \varepsilon = \mathrm{E}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln\frac{2m}{VC(H)}+1\right)+\ln\frac{4}{\delta}}{m}}$$

- So back to our target:

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

- Bound can be found on finite and infinite hypothesis space H, E_{train} can be close enought to E_{true}.

Model Selection

- Known that $E_{train}(g) \approx E_{true}(g)$, But how to select a model g most fitting data and having a ideal performance?

- What about $E_{train}(g) \approx 0$ in our target ?

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

Empirical Risk Minimization

- To simply minimize $E_{train}(g) \approx 0$ in our target ?

$$E_{train}(g) \approx E_{true}(g) \approx 0$$

- So we need
- To minimize $E_{train}(g)$
- To fit more train examples as possible
- A powerful H to shatter more examples
- A big VC(H), but a bad performance on true data
- Thus, lead overfitting

What is overfitting

- Hypothesis h overfits training data, if there is a h' that:

$$Error_{train}(h) < Error_{train}(h') \text{ and}$$

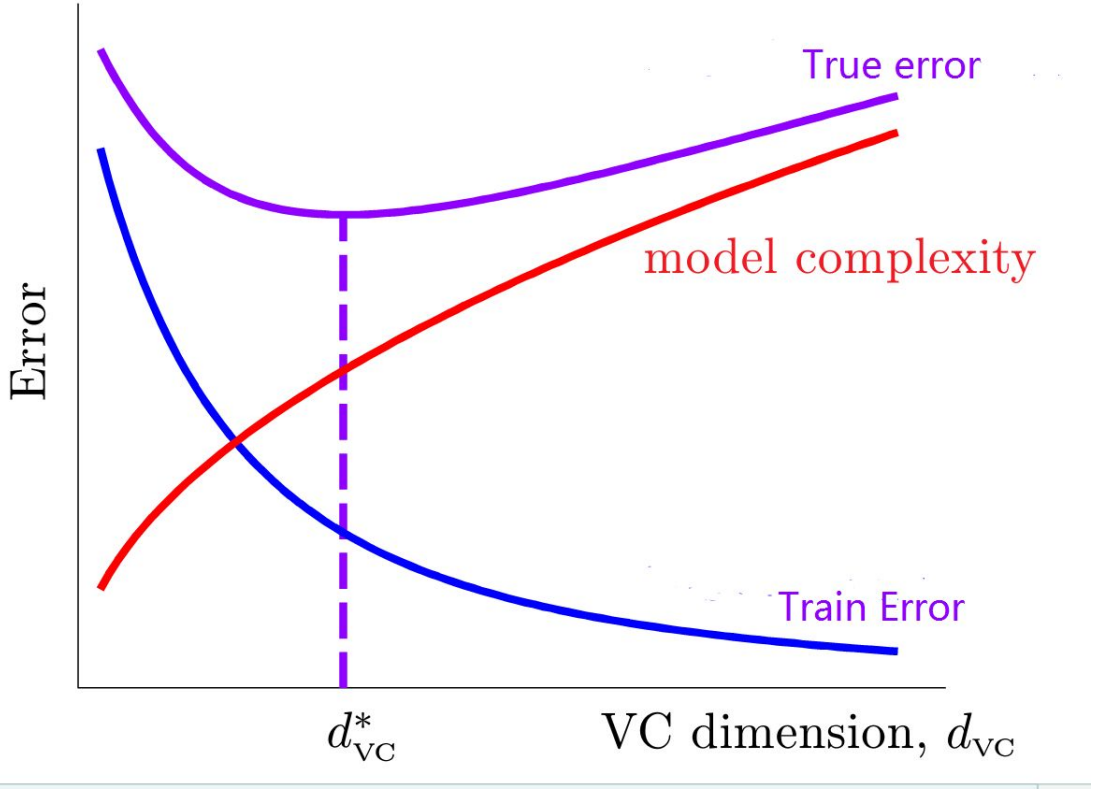
$$Error_{true}(h) > Error_{true}(h')$$

Which is represented by "good on training examples and bad on test examples"

Tradeoff of VC Dimension

- Model complexity increase as VC increase.

- The bigger VC is, train error more likely close to 0, the greater the upper bound between E_{train} and E_{true}



Structural Risk Minimization

- SRM = ERM + f(VC, m)
- where f(VC,m) is the confidence risk function

$$\mathrm{f}(VC,m) \prec \frac{VC}{\mathrm{m}}$$

- So we finally want Min($E_{train}(g)$ + f(VC,m))

$$\mathrm{E}_{true}(h) < \mathrm{E}_{train}(h) + \varepsilon = \mathrm{E}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln\frac{2m}{VC(H)}+1\right)+\ln\frac{4}{\delta}}{m}}$$

Structural Risk Minimization

Aim at choosing an h to minimize the bound on $E_{true}(h)$, a **trade-off** between hypothesis space complexity and empirical error $E_{train}(g)$

Model selection by SRM corresponds to finding the model simplest in terms of order and best in terms of empirical error on the data

Other Model Selection Criterion

- AIC (Akaike Information Criterion)

- BIC (Bayesian Information Criterion)

AIC

- AIC (Akaike Information Criterion)

$$\mathrm{AIC} = LogLikehood(data \mid MLE\ \ params) - (params\ number)$$

- MLE = Maximum Likelihood Estimation

- Take into account the R-squared of model(模型拟合度) and model complexity by measuring the number of parameters

BIC

- BIC (Bayesian Information Criterion)

$$\mathrm{BIC} = LogLikehood(data \mid MLE\ params) - \frac{(params\ number)}{2}\log m$$

- MLE = Maximum Likelihood Estimation

- Take into account the R-squared of model(模型拟合度) , the size of training example and model complexity

Conclusion

- There is a ideal hypothesis g, which works well on future data (Learnable)

- VC dimension is a measure of model complex by shattering.
- the bigger VC goes,
- the model more complex
- the upper bound on E_{train} and E_{true} increase

- Model selection is tradeoff between simple model and good performance on training data

The End...

Q&A