Introduction to Learning Theory

Questions suitable for Theoretical Analysis

- How hard is a given learning task?
- How many training examples are needed?
- How large is the risk of failing?

Concept Learning

Concept Learning

Learning of a boolean function from examples

Categories

- “Nice weather”
- “Dog”
- “Motor vehicle”
- “Criminal offence”

Subsets of a superset X

Terminology

c The concept to learn

$$c(x) \to \mathcal{F}/\mathcal{T}, \quad x \in X$$

h Hypothesis, Result of the learning (“guessed c”)

$$h(x) \to \mathcal{F}/\mathcal{T}, \quad x \in X$$

H Hypotheses space, All conceivable hypotheses (before data arrives)

$$h \in H$$

D Set of available training data

$$D \subseteq X$$

Terminology

Two kinds of training examples

Positive example:

$$x : c(x) = \mathcal{T}, \quad x \in D$$

Negative example:

$$x : c(x) = \mathcal{F}, \quad x \in D$$

Example of a *concept*

"Nice Weather"

Let each "weather instance" x_i be composed of three attributes:

$$x_1 = \ <\text{Sunny, Warm, Windy}>$$
$$x_2 = \ <\text{Cloudy, Warm, Calm}>$$
$$x_3 = \ \cdots$$

Generally: *Sky* $\times$ *Temperature* $\times$ *Wind*

Assume that the attributes can only take on certain discrete values:

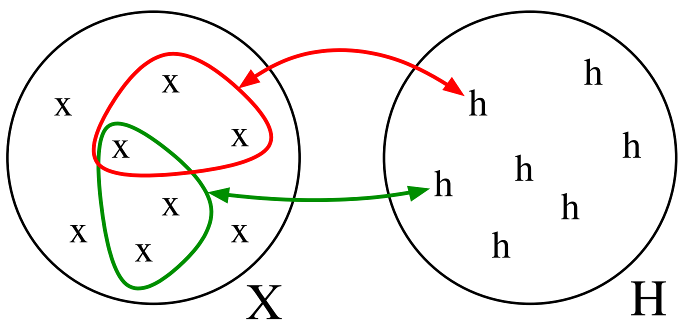
Sky	$\in$ { Sunny, Cloudy, Rainy }
Temp	$\in$ { Warm, Mild, Cold }
Wind	$\in$ { Windy, Calm }

Number of possible weathers: $|X| = 3 \cdot 3 \cdot 2 = 18$

Typical training samples

$x_1 =$ <Sunny, Warm, Windy> $\rightarrow$ Nice
$x_2 =$ <Sunny, Mild, Windy> $\rightarrow$ Nice
$x_3 =$ <Rainy, Cold, Windy> $\rightarrow$ Bad
$x_4 =$ <Sunny, Warm, Calm> $\rightarrow$ Nice

What does the hypotheses space H look like?



Each hypothesis h corresponds to one subset of X

How many hypotheses can we choose from?
How many subsets does X have?

$$|H| = 2^{|X|}$$

$$|H| = 2^{18} = 262144$$

Training data alone is not sufficient to isolate one hypothesis!

Inductive Bias

The assumptions the learner uses to generalize

Examples (from Wikipedia):

- Maximum conditional independence
- Minimum cross-validation error
- Maximum margin
- Minimum description length
- Minimum features
- Nearest neighbors

Assumptions:

- Concept Learning
- Training and test data from same distribution $\mathcal{D}$

What can go wrong?

- The result of leaning can be bad
  The resulting hypothesis makes too many errors
- Learning itself can fail
  The learning algorithm does not find any useful hypothesis

True Error

The probability that a given hypothesis gives the wrong answer

$$\mathrm{error}_{\mathcal{D}}(h) \equiv P_{x \in \mathcal{D}}\left[h(x) \neq c(x)\right]$$

How bad hypotheses are we prepared to accept?

Approximately Correct

A hypothesis h is called approximately correct if

$$\mathrm{error}_{\mathcal{D}}(h) < \epsilon$$

Quantification of the risk that learning does not find an approximately correct hypothesis

$$P_L\left[\mathrm{error}_{\mathcal{D}}(h) \geq \epsilon\right]$$

How often is it acceptable for learning to fail?

Probably Succeeds

The algorithm L is said to probably find a solution if

$$P_L\left[\mathrm{error}_{\mathcal{D}}(h) \geq \epsilon\right] < \delta$$

PAC-learning

Probably Approximately Correct

Given

- ϵ limit on the error
- δ limit on the risk
- n size of the examples

Efficiently PAC-learnable

A concept is said to be efficiently PAC-learnable if there exists an algorithm which runs in polynomial time in

$$n,\ \frac{1}{\epsilon} \text{ and } \frac{1}{\delta}$$

Analysis of a Consistent Learner

- Assumtion: no errors in training examples
- Examples are drawn from the distribution $\mathcal{D}$
- The solution is consistent with all training examples
- "Dangerous Hypotheses":

$$\mathrm{error}_{\mathcal{D}}(h) \geq \epsilon$$

We do not want learning to produce a dangerous hypothesis!

How large is the risk that a dangerous hypothesis is consistent with all training examples?

- Probability that one hypothesis h is contradicted by one example

$$\mathrm{error}_{\mathcal{D}}(h)$$

- Probability that h is not contradicted

$$1 - \mathrm{error}_{\mathcal{D}}(h)$$

- Risk that a *dangerous hypothesis* $(\mathrm{error}_{\mathcal{D}}(h) \geq \epsilon)$ is not contradicted by a randomly drawn example

$$\leq (1 - \epsilon)$$

- Risk that a *dangerous hypothesis* is not contradicted by one randomly drawn example

$$\leq (1-\epsilon)$$

- Risk that a *dangerous hypothesis* is not contradicted my m randomly drawn examples

$$\leq (1-\epsilon)^m$$

- How large is the risk that any dangerous hypothesis in H happens to be consistent with all examples:

$$\leq |H| \cdot (1-\epsilon)^m$$

$$\leq |H| \cdot e^{-\epsilon m}$$

How many training examples are needed?

How many examples m are needed to make the risk of ending up with a dangerous hypothesis less than δ?

$$\delta \geq |H| \cdot e^{-\epsilon m}$$

$$e^{\epsilon m} \geq \frac{|H|}{\delta}$$

$$m \geq \frac{1}{\epsilon}\left[\ln |H| + \ln \frac{1}{\delta}\right]$$

Problem with $|H|$

- Gives too pessimistic estimates
- Can't be used when $|H| = \infty$

Vapnik — Chervonenkis observation:

The important thing is not the *number of* hypotheses, but how they can form subsets in X

Scattering

A finite set S is scattered by the hypotheses H if every subset of S is described by a $h \in H$

The size of S is a measure of the expressive power of H

VC Dimension

$\mathrm{VC}(H)$
Size of the largest subset of X which can be scattered by H

Example:

H	Intervals on the real axis
X	Real numbers

- Can 2 points be scattered?
- Can 3 points be scattered?

Conclusion: $\mathrm{VC}(H) = 2$

Example:

H	Separating hyperplane
X	Points in $\Re^r$

- When $r = 1$

$$\mathrm{VC}(H) = 2$$

- When $r = 2$

$$\mathrm{VC}(H) = 3$$

- Generally

$$\mathrm{VC}(H) = r + 1$$