Computational Learning Theory

Part 1

Paul Scheible

Questions to Answer

- Sample Complexity
  - How many training examples are needed for a learner to converge?
- Computational Complexity
  - How much computational effort is needed for a learner to converge?
- Mistake Bound
  - How many training examples will the learner misclassify before converging?

Attributes for a Learning Problem

- Size of hypothesis space
- Accuracy to which the target concept must be approximated
- Probability that the learner will output a successful hypothesis
- Manner in which training examples are presented

Terminology

- X – Set of all possible instances over which the target function is defined
- x – A single instance of X
- C – Set of all target concepts
- c – A concept in C
- $c : X \rightarrow \{0, 1\}$

Terminology

- $c(x)$ – Value of a concept c given an instance x of X
- $\mathcal{D}$ – Probability distribution of X
- L – Learner
- H – Set of possible hypotheses considered by L
- h – A single hypothesis in H

Terminology

- D – Set of example instances
- ε – Bound on the error in the hypothesis
- δ – Probability of failure of the hypothesis

Probably Approximately Correct (PAC)

- Restricted to noise free training data
- Boolean valued concepts

Error in a Hypothesis

Definition

The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis *h* with respect to target concept *c* and distribution $\mathcal{D}$ is the probability that *h* will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Error of a Hypothesis

- Note difference between true and training error
- Note similarity between training and sample error

PAC Learnability

Definition

Consider a concept class C defined over a set of instances X of length n and a learner L using hypothesis space H. C is PAC-learnable by L using H if the following is true in polynomial time with respect to $1/\varepsilon$, $1/\delta$, n, and size(c):

$$\forall c \in C, \mathcal{D} \text{ over } X, 0 \leq \epsilon < \frac{1}{2}, 0 \leq \delta < \frac{1}{2},$$
$$Pr[L \rightarrow h \in H] \geq (1 - \delta) \text{ such that } error_{\mathcal{D}}(h) \leq \epsilon$$

PAC Learnability

- Show that some class C of target concepts are learnable in a polynomial number of examples
- Show that each example is processed in polynomial time
- Assumes H contains a hypothesis h with arbitrarily small error bound
- Difficult to prove if C is not known in advance

Sample Complexity for Finite Hypothesis Spaces

Version Space Definition

The version space, denoted $VS_{H,D}$, with respect to hypothesis space H and training examples D, is the subset of hypotheses from H consistent with training examples D.

$$VS_{H,D} = \{ h \in H \mid (\forall \langle x, c(x) \rangle \in D)(h(x) = c(x)) \}$$

Definition

Consider a hypothesis space *H*, target concept *c*, instance distribution $\mathcal{D}$, and a set of training examples *D* of *c*. The version space $VS_{H,D}$ is said to be **ε-exhaustive** with respect to *c* and $\mathcal{D}$, if every hypothesis *h* in $VS_{H,D}$ has error less than *ε* with respect to *c* and $\mathcal{D}$.

$$(\forall h \in VS_{H,D})\, error_{\mathcal{D}}(h) < \epsilon$$

Theorem 7.1

ε-exhausting the version space. If the hypothesis space *H* is finite, and *D* is a sequence of *m* >= 1 independently randomly drawn examples of some target concept *c*, then for any 0 <= ε <= 1, the probability that the version space $VS_{H,D}$ is not ε-exhausting (with respect to *c*) is less than or equal to

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

Sample Size

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

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

Sample Size

- Can be substantial overestimate
- Problem caused by size of H in equation

Agnostic Learning and Inconsistent Hypotheses

- $error_D(h)$ – Training error in hypothesis h
- Hoeffding bounds (additive Chernoff bounds)
- How many examples so that $error_{\mathcal{D}}(h) <= \varepsilon + error_D(h)$?

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$$

Boolean Conjunctions

- PAC-learnable

$$m \geq \frac{1}{\epsilon}\left(n \ln 3 + \ln\left(\frac{1}{\delta}\right)\right)$$

- Theorem 7.2 PAC-learnability of boolean conjunctions. The class *C* of conjunctions of boolean literals is PAC-learnable by the Find-S algorithm using $H = C$.

PAC-Learnability of Other Concept Classes

- Unbiased Learners
  - Not polynomially bounded in sample complexity based on Equation 7.2
- K-Term Disjunctive Normal Form
  - Polynomial sample complexity
  - Not polynomial in computational complexity

PAC-Learnability of Other Concept Classes

- K-Term Conjunctive Normal Form
  - Polynomial sample complexity
  - Polynomial computational complexity
  - K-term DNF can be reduced to K-term CNF