Probably Approximately Correct Learning - II

Prof. Dan A. Simovici

UMB

Trainig Error vs. Generalization Error

Let $\mathbf{s} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ be a sample. The *training error* or *empirical error* of a particular hypothesis H is the fraction of training examples it misclassifies:

$$\widehat{err}(H) = \frac{1}{m} \sum_{i=1}^{m} I_{H(\mathbf{x}_i) \neq y_i}$$

If $(\mathbf{x}, y) \sim \mathcal{D}$, the *true error* or the *generalization error* is

$$err(H) = P_{(\mathbf{x},y)\sim\mathcal{D}}[H(\mathbf{x}) \neq y].$$

The training error is a proxy for the generalization error.

A General Analysis of Classifier Errors

Success in learning depends on

- finding a classifier that fits the data well, that is, has low training error;
- the classifier must be simple;
- the learner must be provided with a sufficiently large training set.

The analysis does not depend on any probability distribution.

Trainig Error vs. Generalization Error

- When working with a single hypothesis H the training error is an unbiased estimator of the generalization error.
- With a large hypothesis space the algorithm will be biased towards hypotheses whose training errors are, by chance much lower than true errors.

Estimation of Generalization Error - I

- CENTRAL QUESTION: How much the training error $\widehat{err}(H)$ can differ from the true error $err(H)$ as a function of the number of training examples m?
- FUNDAMENTAL ASSUMPTION: Hypothesis H is selected before the training set is randomly chosen.

Estimation of Generalization Error - II

Equivalent problem: when a training example $(\mathbf{x}_i, y_i)$ is selected at random the probability $P(H(\mathbf{x}_i) \neq y_i)$ equals $p = err(H)$ and $P(H(\mathbf{x}_i) = y_i)$ equals $1 - p$. This can be restated in an experiment with a biased coin:

head	if $H(\mathbf{x}_i) \neq y_i$	p
tail	if $H(\mathbf{x}_i) = y_i$	1-p

The Coin Flipping Analogy

The estimation amounts now to the evaluation that the probability that the fraction of heads $\hat{p}$ in a series of m coin flippings will be different from p.

Hoeffding's Inequalities

Let $X_1, \ldots, X_m$ be m independent random variables ranging in the interval $[0, 1]$ and let A_m be the random variable

$$A_m = \frac{X_1 + \cdots + X_m}{m}.$$

Then, we have

$$P(A_m \geqslant E[A_m] + \epsilon) \leqslant e^{-2m\epsilon^2},$$

and

$$P(A_m \leqslant E[A_m] - \epsilon) \leqslant e^{-2m\epsilon^2}.$$

Also,

$$P(|A_m - E(A_m)| \geqslant \epsilon) \leqslant 2e^{-2m\epsilon^2}.$$

- $X_i = 1$ with the probability p (heads) and $X_i = 0$ with probability $1 - p$ (tails).
- A_m equals to $\hat{p}$, the fraction of heads obtained in m flips and $E(A_m) = p$. We have $A_m \leq p - \epsilon$ iff $\hat{p} \leq p - \epsilon$ iff n_h the number of heads is such that $n_h \leqslant (p - \epsilon)m$.
- The probability of at most $(p - \epsilon)m$ heads is at most $e^{-2m\epsilon^2}$.

The Learning Framework - I

- $$X_i = \begin{cases} 1 & \text{if } H(\mathbf{x}_i) \neq y_i, \\ 0 & \text{otherwise;} \end{cases}$$
- $E(A_n) = \widehat{err}(H)$;
- $E(A_m)$ is the generalization error;
- $P(err(H) \geqslant \widehat{err}(H) + \epsilon) \leq e^{-2m\epsilon^2}$.

The Learning Framework - II

Let

$$\delta = e^{-2m\epsilon^2},$$

so

$$\epsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

With the probability at least $1 - \delta$ we have

$$err(H) \leq \widehat{err}(H) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

If H has a low training error on a sufficiently large training set, then we can be confident that the true error of H is also low.

The Learning Framework - III

$$P(|err(H) - \widehat{err}(H)| \geqslant \epsilon)$$

is at most $2e^{-2m\epsilon^2}$, or, equivalently,

$$|err(H) - \widehat{err}(H)| \leqslant \sqrt{\frac{\frac{2}{\delta}}{2m}}$$

with a probability at least $1 - \delta$.

Finite Hypothesis Space Analysis

$\mathcal{H}$ is the space of hypotheses.

Theorem

Let $\mathcal{H}$ be a finite space of hypotheses and assume that a random training set of size m is chosen. Then, for any $\epsilon > 0$,

$$P\left((\exists H \in \mathcal{H}) : err(H) \geqslant \widehat{err}(H) + \epsilon\right) \leq |\mathcal{H}| e^{-2m\epsilon^2}.$$

Thus, with probability at least $1 - \delta$ we have:

$$err(H) \leq \widehat{err}(H) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2m}}.$$

Proof

- Hypothesis $H \in \mathcal{H}$ is chosen before observing the training set.
- If we fix any single hypothesis $H \in \mathcal{H}$, $P(err(H) - \widehat{err}(H) \geqslant \epsilon) \leq e^{-2m\epsilon^2}$.
- By union bound, the probability that this will happen for any hypothesis in $\mathcal{H}$ can be upper bounded by $|\mathcal{H}|e^{-2m\epsilon^2}$.

ϵ-Nets

Let $\mathcal{C}$ be a concept class and let $C \in \mathcal{C}$.

Definition

The *class of error regions* of C and $\mathcal{C}$ is the collection of sets

$$\mathrm{R}(C, \mathcal{C}) = \{C \oplus D \mid D \in \mathcal{C}\}.$$

Theorem

Let $\mathcal{C}$ be a collection of concepts, $\mathcal{C} \subseteq U$. If $K \in \mathcal{C}$, then $VCD(R(K,\mathcal{C})) = VCD(\mathcal{C})$.

Proof: Let S be a set. Define

$$f : \{S \cap C \mid C \in \mathcal{C}\} \longrightarrow \{S \cap D \mid D \in R(K,\mathcal{C})\}$$

as $f(S \cap C) = S \cap (C \oplus K)$ for every $C \in \mathcal{C}$. Observe that

$$f(S \cap C) = S \cap (C \oplus K) = (S \cap C) \oplus (S \cap K).$$

Thus, if $f(S \cap C_1) = f(S \cap C_2)$, the equality

$$(S \cap C_1) \oplus (S \cap K) = (S \cap C_2) \oplus (S \cap K)$$

implies $(S \cap C_1) = (S \cap C_2)$, so f is a bijections. Therefore, $\mathcal{C}$ shatters the set S if and only if $R(K,\mathcal{C})$ shatters that set.

A Further Refinement

For $\epsilon > 0$ define

$$\mathrm{R}_\epsilon(C, \mathcal{C}) = \{C \oplus D \mid D \in \mathcal{C} \text{ and } P(C \oplus D) \geqslant \epsilon\},$$

where P is a fixed probability on $\mathcal{P}(U)$.

Definition

A set S is an *ϵ-net* on for $\mathrm{R}(C, \mathcal{C})$ if for every $R \in \mathrm{R}_\epsilon(C, \mathcal{C})$ we have $S \cap R \neq \emptyset$.

Example

Let $U = [0,1]$, P be the uniform distribution on U and assume that $\mathcal{C}$ is

$$\mathcal{C} = \{\emptyset\} \cup \{[x,y] \mid x,y \in [0,1]\}.$$

If $C = \emptyset$, then $\mathrm{R}(C,\mathcal{C}) = \mathcal{C}$.
For any interval I included in $[0,1]$, $P(I)$ is the length of I.
The set of points

$$S = \left\{ n\epsilon \middle| 1 \leqslant n \leqslant \left\lceil \frac{1}{\epsilon} \right\rceil \right\}$$

is an ϵ-net for $\mathrm{R}(\emptyset,\mathcal{C})$ because the distance between two consecutive points of S is ϵ, $[x,y] \in \mathrm{R}_\epsilon(\emptyset,\mathcal{C})$ implies $P([x,y]) \geqslant \epsilon$ (and thus, $y - x \geqslant \epsilon$), so $S \cap [x,y] \neq \emptyset$.

Theorem

If a sample **s** *forms an* ϵ*-net for* $\mathrm{R}(C, \mathcal{C})$*, and a learning algorithm produces a hypothesis* $H \in \mathcal{C}$ *that is consistent with* **s***, then this hypothesis must have an error less than* ϵ*.*

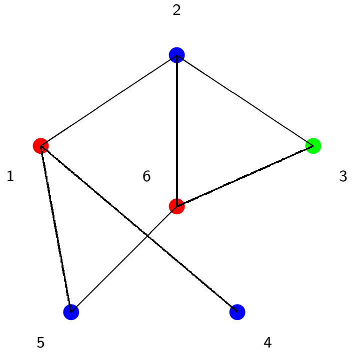
Proof: Note that $H \oplus C \in \mathrm{R}_\epsilon(C, \mathcal{C})$ because H was not hit by S and S is a ϵ-net for $\mathrm{R}(C, \mathcal{C})$, so we must have $H \oplus C \notin \mathrm{R}_\epsilon(C, \mathcal{C})$ and therefore, $err(H) \leqslant \epsilon$.

An NP-Complete Problem

The Graph 3-Coloring Problem: (G3CP) Given an graph $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, n\}$ is the vertex set and $E \subseteq V \times V$ is the edge set, determine if there exists a function $f : V \longrightarrow \{c_1, c_2, c_3\}$ such that for every edge $(i, j) \in E$, $f(i) \neq f(j)$.
This is an NP-complete problem, so a computationally intractable problem.

An Instance of G3CP

(Kearns and Vazirani)



$S^+_{\mathcal{G}}$

0	1	1	1	1	1
1	0	1	1	1	1
1	1	0	1	1	1
1	1	1	0	1	1
1	1	1	1	0	1
1	1	1	1	1	0

$S^-_{\mathcal{G}}$

0	0	1	1	1	1
0	1	1	0	1	1
0	1	1	1	0	1
1	0	0	1	1	1
1	0	1	1	1	0
1	1	0	1	1	0
1	1	1	1	0	0

3DNF Formulas

3DNF formulas are disjunctions of three monomials

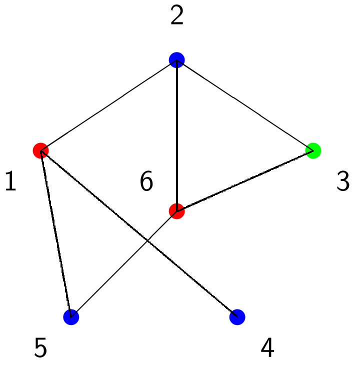
$$\phi = \mu_1 \vee \mu_2 \vee \mu_3,$$

where $\mu_i \in \mathsf{MON}_n$ for $1 \leqslant i \leqslant 3$.
The size of a formula ϕ is no larger than $6n$.
Claim: The graph $\mathcal{G}$ is 3-colorable if and only if $S_{\mathcal{G}} = S_{\mathcal{G}}^{+} \cup S_{\mathcal{G}}^{-1}$ is consistent with some 3DNF formula.

Claim Justification

Suppose that $\mathcal{G}$ is 3-colorable and choose a coloring of $\mathcal{G}$.
Let τ_K be a monomial that corresponds to the color K: τ_K is the conjunction of the variables that correspond to vertices not colored by K. Thus,

$$\tau_R = x_2x_3x_4x_5, \tau_B = x_1x_3x_6, \tau_G = x_1x_2x_4x_5x_6.$$



Every positive example in $S^+_{\mathcal{G}}$ satisfies one of the formulas τ_K.
No negative example in $S^-_{\mathcal{G}}$ satisfies any of the formulas τ_K.

Claim Justification (cont'd)

Suppose that $\tau_R \vee \tau_B \tau_G$ is consistent with $S_{\mathcal{G}}$. Define a coloring as follows: the color of i is K if $v(i)$ satisfies T_K (for $K \in \{R, B, G\}$) and is chosen arbitrary if more than one monomial is satisfied among the colors that correspond to these monomials.

- This is a *legal coloring:* if i and j are assigned the same color, say R, then both $v(i), v(j)$ satisfy τ_R. Since the i^{th} bit of $v(i)$ is 0 and the i^{th} bit of v_j is 1 it follows that neither x_i not $\overline{x_i}$ can appear in τ_R.
- Since $v(j)$ and $e(i,j)$ differ only in their i^{th} bits, if $v(j)$ satisfies τ_R, then so does $e(i,j)$, implying that $e(i,j) \not\in S_{\mathcal{G}}^{-}$, so $(i,j) \not\in E$.

Structural Risk Minimization

Find a hypothesis H for which one can guarantee the lowest probability of error for a given training sample

$$\mathbf{s} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{s}_m, y_m))$$

$$err(H) \leqslant widehaterr(H) + O\left(\frac{d \ln \frac{n}{d} - \ln \delta}{m}\right)$$

You should consult referrences [2] and [3] and [1].

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth.
Learnability and the Vapnik-Chervonenkis dimension.
Journal of the ACM, 36:929–965, 1989.

M. J. Kearns and U. V. Vazirani.
An Introduction to Computational Leaning Theory.
MIT Press, Cambridge, MA, 1997.

R. E. Schapire and Y. Freund.
Boosting – Foundations and Algorithms.
MIT Press, Cambridge, MA, 2012.