An Introduction to Statistical Learning Theory
Malik Magdon-Ismail

Benjamin Disraeli, Earl of Beaconsfield, (1804-1881)

There exists: Lies, Big Lies, and Statistics.
Learning

• No model Exists.

• Need to take “optimal” action based upon history.

Statistical

• World we live in has uncertainty.

• Need to embed in a theory that can account for uncertainty.
What Makes Statistical Learning Interesting?

March 14

March 16

March 21

March 23

March 30

April 1: $100
A “Non-Falsifiable” Strategy - The Postal Scam

Total Mailings = 64
Cost=$32 (@ $0.50 each)
Guaranteed Profit! (providing yield > \( \frac{1}{3} \)).
The Learning Problem

**Goal:** Find \( g(X, \alpha^*) \in \mathcal{L} \) that “approximates” \( Y \).

\[ P(X, Y) = P(X)P(Y|X) \] is not known, but is represented by an iid training sample

\[ (X_1, Y_1), \ldots, (X_N, Y_N) \]

**Ultimate test:**

\[ R(\alpha^*) = E_{P(Z)}[R(Z, g(Z, \alpha^*))] \]

\( R(Z, g(Z, \alpha^*)) \), measures the risk on an “unseen” point \( Z \).
The Postal Scam Again

**Data:** 5 Points: (Team 1, Team 2)

**Y:** +1 if Team 1 wins, -1 otherwise

**L:** All $2^5$ functions on 5 data points

**R(Z, g(Z, i))**: 1 if error, 0 otherwise

**R(i)**: True Probability of error

What is R for Function Output?

Don’t Know - Not Falsifiable.
Principle of Risk Minimization

Want Low Risk, $R$.

Stuck with $\mathcal{L}$

Pick function in $\mathcal{L}$ that has lowest Risk

$$\alpha^* = \arg\min_{\alpha} R(\alpha)$$
Empirical Risk Minimization (ERM)

- Approximate $\alpha^*$ by

$$\alpha_N = \arg\min_{\alpha} \frac{1}{N} \sum_{i=1}^{N} R(Y_i, g(X_i, \alpha))$$

- Hope $\alpha_N \approx \alpha^*$

- Postal Scam: $R_{emp} = 0$

**When can we be sure** $R_{emp} \approx R$

Postal Scam: If we can assure that few functions tried, then more “faith”:

$$2^{N-1} \gg Price$$
Interesting Questions

1. **Consistency**: When can we be sure that minimizing $R_{emp}$ corresponds to minimizing $R$? Does $R_{emp}$ reflect $R$?

2. **Convergence Rate**: How does the answer to (1) depend on $N$?

3. **Risk Bound**: Can one estimate / bound the true risk?

4. **Learning Algorithms**: Can one use the theory to construct good learning algorithms?
Consistency – What does it mean

Minimizing $R_{emp}$ corresponds to minimizing $R$.

\[
R(\alpha_N) \xrightarrow{N \to \infty} R(\alpha^*)
\]

\[
R_{emp}(\alpha_N) \xrightarrow{N \to \infty} R(\alpha^*)
\]

(convergence in probability)

Remember:

$\alpha^*$: The function that minimizes true risk

$\alpha^N$: The function that minimizes empirical risk
Trivial Consistency

A function with lower risk for every $Z$ will minimize both empirical and true risk.

We exclude such cases from the theory.
Consistency: Uniform Convergence

A theory of learning must be **worst case**: control the generalization of the worst function and one has “learning”

\[
P[\sup_{\alpha} (R(\alpha) - R_{emp}(\alpha)) > \epsilon] \xrightarrow{N \to \infty} 0
\]
for every \( \epsilon > 0 \)

**Intuition:** If the empirical risk is converging uniformly to the true risk, then picking based upon the empirical risk “converges” to picking based upon the true risk.
Consistency: Falsifiability is Enough

Suppose that for some \( N = d^* \), the number of possible classifications on any set of \( d^* \) points is less than \( 2^{d^*} \).

\[
N(Z_1, \ldots, Z_{d^*}) < 2^{d^*} \quad \forall (Z_1, \ldots, Z_N)
\]

Then, every set of data points can falsify the model.

The existence of such a \( d^* < \infty \) is a necessary and sufficient condition for uniform convergence, independent of \( P(X) \).
Intuition Behind Falsifiability

Scientist 2

Scientist 1

Scientist 3
Convergence with $N$

$$P[\sup_{\alpha} |R(\alpha) - R_{emp}| > \epsilon] \leq C \left((2N)^{d^*-1} + 1\right)e^{-N\epsilon^2}$$

- Exponential Convergence!
- $d^*$ plays a central role.

**Definition [VC-Dimension]:**

$$d_{VC} = \min(d^* - 1)$$

- Good learning models have small $d_{VC}$.
Bound on True Risk

With probability $> 1 - \eta$

$$R(\alpha) \leq R_{emp}(\alpha) + \Phi(N, d_{VC}, \eta)$$

where

$$\Phi(N, d_{VC}, \eta) = \frac{1}{N} \log \left( \frac{6(2N)^{d_{VC}} + 1}{\eta} \right)$$
Building Learning Algorithms

\[ d_1 < d_2 < d_3 < d_4 < d_5 \]

Structural Risk Minimization (SRM)
SRM – Example

In some canonically defined parameterized space, the norm of the parameters gives exactly such a structural risk hierarchy.

If the space is linear, then we get the celebrated support vector machine (SVM).
Why Are SVM’s Good

• They correspond to an SRM framework.

• Have intuitive geometric interpretation.

• The learning can be done exactly.

• Using kernel methods, arbitrarily complex structures can be accommodated, for example
  
  – Neural Networks

  – Polynomials of arbitrary degree

  – Radial Basis Functions
What’s Missing

We See how to get Consistency.

**Ultimate goal:** Low $R$

Don’t care if $R$ happens to be the minimum possible $R$ achievable in $\mathcal{L}$

**Necessary Condition:** Low $R$ functions must exist in $\mathcal{L}$.

How to guarantee this?
Consistency and Low $R$

In general No! (No Free Lunch (NFL))

Use knowledge about input domain to tailor $\mathcal{L}$ or the SRM - Heuristics, Hints, Invariances, . . . .
Conclusions

• One can “learn” consistently.

• One may or may not be able to learn “well”.

• One can use the theory to construct “sound” learning algorithms from the consistency point of view.

• Incorporate Prior Information to get best results – heuristics!

• Lots of Literature.

• Where did we really go wrong with the postal scam?

• Application to other areas, for example data mining.