MTAT.03.227 Machine Learning

Statistical Learning Theory and No Free Lunch Theorems

Sven Laur
University of Tartu
Statistical Learning Theory
Theoretical framework for ML

Let $L(a, b)$ be a loss function. Then the aim of most machine learning algorithms is to minimise risk

$$R(f) = \mathbb{E}_{x,y}(L(f(x), y))$$

over existing but unknown distribution. Usually, we seek a solution from a fixed function class $\mathcal{H}$ and try to minimise empirical risk

$$R_m(f) = \frac{1}{m} \cdot \sum_{i=1}^{m} L(f(x_i), y_i)$$

where samples are assumed to be iid.
Model bias and consistency

As the set of potential solutions is limited by $\mathcal{H}$, we might never discover the true function. A *model bias* for $\mathcal{H}$ and true distribution $\mathcal{D}$ is

$$\text{Bias}(\mathcal{H}|\mathcal{D}) = \min_{f \in \mathcal{H}} \mathbb{E}_{x,y \leftarrow \mathcal{D}}(L(f(x), y))$$

A machine learning algorithm $\mathcal{A}$ is *asymptotically consistent* if for any data distribution

$$\mathbb{E}_{\mathcal{D}}(R(f_m)) \xrightarrow{m} \text{Bias}(\mathcal{H}|\mathcal{D})$$

where $f_m$ is the function returned by $\mathcal{A}$ given $m$ samples and $\mathcal{H}$ is the function class from which $\mathcal{A}$ chooses its output.
Bias-variance-noise decomposition

Let $S$ denote training sample. Consider quadratic loss function then

$$
E_D((A_S(x) - y)^2) = E_D((A_S(x) - f(x) + f(x) - y)^2)
$$

$$
= E_D((A_S(x) - f(x))^2) + E_D((f(x) - y)^2)
$$

provided that $y = f(x) + \varepsilon$ and $\varepsilon$ is independent of $x$ and has zero mean. Let $f_* \in \mathcal{H}$ be the optimal function. Then the first term simplifies further

$$
E_D((A_S(x) - f(x))^2) = E_D((A_S(x) - \overline{A_S}(x) + \overline{A_S}(x) - f(x))^2)
$$

$$
= E_D((A_S(x) - \overline{A_S}(x))^2) + E_D((A_S(x) - f(x))^2)
$$

where $\overline{A_S}(x)$ is averaged prediction of $A_S(x)$ over all training sets $S$. 
Bias-variance trade-off

▷ Simple models have small variance but high bias.
▷ Complex models have small bias but high variance.
▷ We need to balance both components with regularisation.

Minimal training error can be viewed as estimate on model bias

\[ E_{tr} = \min_{f \in \mathcal{H}} R_m(f) \approx \min_{f \in \mathcal{H}} R(f) = \text{Bias}(\mathcal{H} | \mathcal{D}) . \]

Hence we somehow need to estimate the variance term in terms of model coefficients count or something similar.
Sampling bounds for training error

Let us consider finite set of functions \( \mathcal{H} = \{ f_1, \ldots, f_k \} \). The iid assumption allows us to find confidence intervals for \( R(f_j) \). They are of type

\[
\Pr \left[ |R_m(f_j) - R(f_j)| \geq \varepsilon \right] \leq c \cdot \exp \left( -\beta \varepsilon^2 \right) =: \delta
\]

for some constants \( c, \beta > 0 \). They follow from Chernoff, Hoeffding or McDiarmid inequalities. Now applying the union bound we get

\[
\Pr \left[ \exists j : |R_m(f_j) - R(f_j)| \geq \varepsilon \right] \leq |\mathcal{H}| \cdot \delta .
\]

This inequality bounds optimism \( \Delta = R_m(f_*) - R(f_*) \) for the proposed solution \( f_* \) that minimises training error. Since \( \delta \) decreases as a function of \( m \) the method is asymptotically consistent.
Tightness of inequality. Bad case

- Plot showing training error and functions.
- Plot showing best training error and optimism.
- Plot showing frequency and functions.
- Plot showing density and optimism.
Tightness of inequality. Good case
Effective dimension of function families

- A good function family has few functions that are near optimal.
- Very similar functions $\implies$ many near-optimal solutions.

Vapnik-Chervonenkis dimension is one way to characterise flexibility of $\mathcal{H}$

The **VC dimension** of a function class $\mathcal{H}$ if the largest number of samples $d$ for which one can find $x_1, \ldots, x_d$ such that any labelling $y_1, \ldots, y_d$ can be realised.

If the number of samples $m > d$ then only up to

$$G_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$$

labellings can be implemented. For $m > 2d$, $\mathcal{H}$ is really sparse net.
Uniform bound on the optimism

Suppose you have two independent $m$ element sets $S_1$ and $S_2$ and you optimise classifier on one and test it on other how to estimate optimism.

- Lets consider worst case $R^1_m(f) = 0$ on $S_1$ and $R^2_m(f) \geq \varepsilon/2$ on $S_2$.
- Then we somehow need that the combined data set should contain correct number of correct and correct number of incorrect predictions.
- There could be many labellings which would have the desired structure however not all of them are realisable by $H$ as

$$G_H(m) \leq \left(\frac{em}{d}\right)^d$$

- Now we have to estimate how many permutations of data elements provide desired empirical risk estimates. We get exponential bound.
Simple SLT bound

Combining results we get

$$\Pr \left[ \exists f \in \mathcal{H} : R_{1m}(f) = 0 \land R_{2m}(f) \geq \varepsilon/2 \right] \leq \left( \frac{em}{d} \right)^d 2^{-m\varepsilon/2}$$

something that converges to zero when \( m \) grows. As for the second empirical error is hold-out error it can be large only if \( R(f) \) is large.

More careful analysis gives you the bound

$$\Pr \left[ \forall f \in \mathcal{H} : R(f) \leq R_m(f) + c(\delta)\sqrt{\frac{\text{VCdim}(\mathcal{H})}{n}} \right] \leq 1 - \delta$$
More complex ideas

By construction

\[ R_m(f) = 1 + \frac{1}{m} \cdot \sum_{i=1}^{m} f(x_i)y_i \]

Hence we could bound the optimism by computing the upper bound

\[ R_m^2 - R_m^1 \leq \sup_{f \in \mathcal{H}} \left\{ \frac{1}{m} \cdot \sum_{i=1}^{m} f(x_i)y_i - \frac{1}{m} \cdot \sum_{i=m+1}^{2m} f(x_i)y_i \right\} \]

It turns out that considering a single function \( y \equiv 1 \) is enough to characterise the class \( \mathcal{H} \). Instead of taking maximum lets take average over distribution.
Alternatives to VC-dimension

There are many alternatives to VC-dimension:

- **Rademacher complexity**
  \[
  \mathcal{R}_m(\mathcal{H}|S) = \mathbb{E}\left[\sup_{f \in \mathcal{H}} \frac{2}{m} \cdot \sum_{i=1}^{m} \sigma_i f(x_i)\right]
  \text{ for } \sigma_i \in \{-1, 1\}
  \]
  \[
  \mathcal{R}_m(\mathcal{H}) = \mathbb{E}_{S \leftarrow \mathcal{D}}(\mathcal{R}_m(\mathcal{H}|S))
  \]

- **Covering numbers**

They are important since sometimes we cannot compute VC-dimension while the other complexity measures do exist and are easy to compute.
NFL Theorems
Why do we need NFL theorems?

- NFL theorems show that all machine learning methods are equally good if we do not make strong assumptions.
- Sampling bounds and SLT results are correct but do not tell exactly what we need and hope. NFL results shed a light to small print conditions.
- NFL theorems show that for each machine learning algorithm there exists a sample or sample class where it outperforms some other method.

We state NFL theorems in maximally simple way to show their banal nature.
Simplest version of NFL theorem

Consider a finite learning problem.

- Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be the training set.
- Let $\mathcal{T} = \{(x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)\}$ the remaining data.
- Let $\mathcal{A}_S(x)$ prediction of machine learning method on $x$.
- What is *out-of-training sample error*

$$R_{ot}(\mathcal{A}) = \sum_{i=m+1}^{m} [\mathcal{A}_S(x) \neq y_i]?$$

**Theorem.** If we assume that labels $y_i$ are computed as $f(x_i)$ for a function chosen uniformly form all possible functions then

$$E_f (R_{ot}(\mathcal{A})) = \frac{1}{2}.$$
NFL theorem for problem instances

- In real life tasks, some problems are harder than others.
- Let $\mathcal{F}$ denote probability distribution over all functions
- Which problem instances $\mathcal{F}$ the algorithm $\mathcal{A}$ can handle?

**Theorem.** Consider a uniform distribution over all possible problem instances. Then

$$\mathbb{E}_{\mathcal{F}} \left( \mathbb{E}_{f \leftarrow \mathcal{F}} \left( Rot(\mathcal{A}) \right) \right) = \frac{1}{2}.$$ 

As a result, no algorithm can be better than the other—there are some problem instances for which the other is better.
NFL theorem and SLT bounds

NFL theorem says that you cannot learn arbitrary function. On the same time, SLT bounds assure that $R_m(f) + \Delta$ is rather accurate estimate.

What happens then?

- We can assume that $n$ large enough so that $R(f) \approx R_{ot}(f)$.
- For most functions $R_m(f) + \Delta \approx 0.5$ and we do not care.
- For few functions $R_m(f) + \Delta \approx 0.0$
  - For half of these functions $R_{ot}(f) \approx 0.0$ and we are happy.
  - For half of these functions $R_{ot}(f) \approx 1.0$ and we are happy.

Small print notice. For fixed error estimate SLT bounds say nothing.

- Rejection of uninteresting results creates bias!
- We must additionally assume that the learning task is easy.