Machine Learning
EDS

Approximation-Estimation Error decomposition and first results on generalisation error of a learnt predictor

Janneke van Brummelen

Vrije Universiteit Amsterdam

Period 3.4
2023/2024

So far

- Because ideal predictors rely on the unknown joint distribution of features/outputs, they do not provide a practical way of making predictions
- We have introduced learning rules (e.g. ERM) to select predictors “best fitting” a given observed sample
- Will they provide accurate predictions for new unseen data?
- We need to quantify their generalisation abilities.

1. Approximation and Estimation errors
   - Approximation error
   - Estimation error
   - Decomposition of the excess risk

2. Upper-bounding the estimation error in ERM
   - Zero-error classification with finite models
   - Beyond the zero-error case: global upper-bounds
   - Non-deterministic case with finite models

Table of Contents

1 Approximation and Estimation errors
- Approximation error
- Estimation error
- Decomposition of the excess risk

2 Upper-bounding the estimation error in ERM
- Zero-error classification with finite models
- Beyond the zero-error case: global upper-bounds
- Non-deterministic case with finite models

Intuition

- When introducing predictors minimising the empirical risk, we straightaway saw that extreme overfitting could occur if no constraints were imposed.

- More generally, when selecting a predictor based on the sample, a structure has to be imposed: we search for predictors within certain families, certain models.

- What if a model that we impose is "far from reality"? Intuitively: even the best predictor within this model will have a poor performance.

The **discrepancy between** 1) the best performance any predictor within a given model can achieve, 2) the perf. of ideal predictors is called the approximation error.

Approximation error

Definition: Approximation error of a model

For a given model $S \subset \mathcal{F}$, the quantity denoted

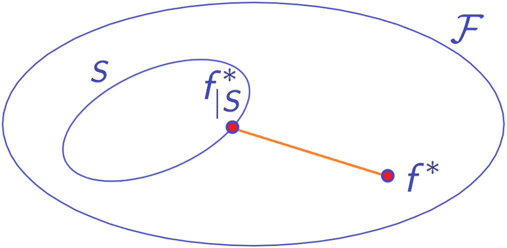
$$\ell(f^*, S) := \inf_{f \in S} \ell(f^*, f) = \inf_{f \in S} \mathcal{R}_P(f) - \mathcal{R}_P^* \geq 0,$$

is called the approximation error of the model S.
Any predictor in S achieving the infimum will be denoted $f^*_{|S}$.
If such a predictor exists, then by definition $\ell(f^*, S) = \ell(f^*, f^*_{|S})$.

Reminder

$\mathcal{R}_P^* := \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$ is Bayes risk,
i.e. the risk of ideal predictors, i.e. the lowest possible generalisation error achievable by any predictor $f \in \mathcal{F}$.

Approximation error

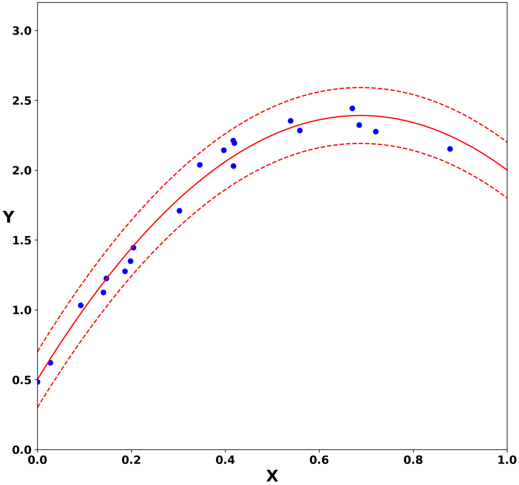


f^*: Bayes predictor $\qquad \mathcal{R}_P(f^*) = \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$

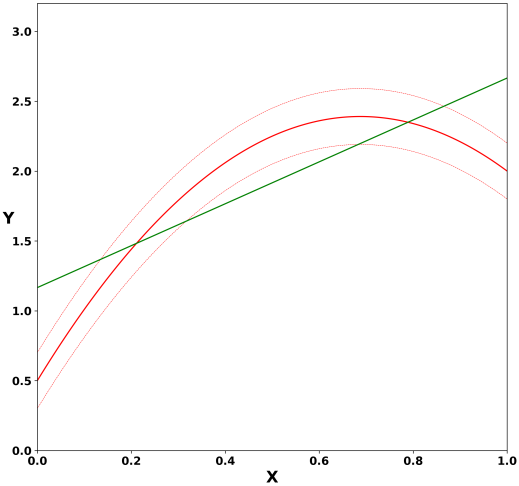
$f^*_{|S}$: theoretical best predictor within model S $\qquad \mathcal{R}_P(f^*_{|S}) = \inf_{f \in S} \mathcal{R}_P(f)$

$$\text{Approximation error} = \inf_{f \in S} \mathcal{R}_P(f) - \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$$

Illustration



Best linear approximation of a non-linear relationship



Approximation error

- The approximation error measures how accurately a given model is, in theory, able to capture a certain feature/output relationship.
- It is a property of the model and of the prediction problem addressed.
- It depends on the feature/output joint distribution P and on the model S, but not on the sample D_n.

Example: Neural Network as universal approximators

- Artificial neural networks constitute a class of popular models which had recent successes in pattern recognition and NLP
- Recall that neural networks have a long history:
    - 1943: McCulloch and Pitts propose a first formalisation of biology-inspired neural networks
    - 1958: Rosenblatt introduces the Perceptron
    - 1970s and 1980s: Emergence, formalisation and implementation of the back-propagation algorithm
    - Starting 2009-2012: Computational power and larger architectures (known as *deep* networks) enabled to reach human performance in several tasks

Neural networks are a family of nonlinear functions mapping features $x \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$ which are known since 1989 to feature a ***Universal Approximation*** property

Example: Neural Network as universal approximators

Single hidden-layer feedforward neural networks

Consider the regression framework with $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$. Formally, single hidden-layer feedforward NN are functions of the form:

$$NN(x) = \sum_{j=1}^{m} \beta_j \Psi\big(\boldsymbol{w}_j^\top x + b_j\big), \qquad \text{for } x \in \mathbb{R}^p,$$

- $m \geq 1$: number of units in the hidden layer (or *neurons*)
- $\boldsymbol{w}_1, \dots, \boldsymbol{w}_m \in \mathbb{R}^p$ and $b_1, \dots, b_m \in \mathbb{R}$ are weights defining an affine mapping of the inputs to the hidden layer
- $\Psi : \mathbb{R} \to [0, 1]$ is the activation function of the units, assumed here to be non-decreasing with $\Psi(x) \underset{x \to +\infty}{\longrightarrow} 1$, $\Psi(x) \underset{x \to -\infty}{\longrightarrow} 0$
- $\beta_1, \dots, \beta_m$ are weights mapping the hidden layer to the output of the neural network

Example: Neural Network as universal approximators

Proposition: NN are universal approximators
Hornik, Stinchombe and White (1989)

Let $p \geq 1$, Ψ be any activation function, and $f : \mathbb{R}^p \to \mathbb{R}$ be any continuous function.

Then, for any accuracy level $\varepsilon > 0$ and any $[a, b]^p \subset \mathbb{R}^p$, there exists a feedforward neural network NN with single hidden-layer and activation function Ψ such that:

$$\sup_{x \in [a,b]^p} |f(x) - NN(x)| < \varepsilon.$$

Neural networks with as little as one hidden layer are able to approximate any continuous function with arbitrary accuracy.

(See the paper if you are interested in the proof)

Example: Neural Network as universal approximators

Advnaced exercise: Approximation error of neural networks

Consider the regression framework $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}$. Let c be the quadratic cost, let P denote joint distribution of features/outputs and assume the following:

- The features are bounded: there is a bounded set $K \subset \mathbb{R}^d$ such that $\mathbb{P}(X \in K) = 1$.
- There exists a Bayes predictor which is continuous, i.e. there is a continuous predictor f^* such that $\mathcal{R}_P(f^*) = \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$.

Let S_{NN} be the set of feedforward neural networks with single-hidden layer. Show that in this context, the approximation error is zero:

$$\inf_{g \in S_{\mathrm{NN}}} \ell(f^*, g) = 0.$$

Are Universal Approximators a guarantee of success?

From Hornik, Stinchombe and White's (1989) article:

"*Any lack of success in applications [of neural networks] must arise from inadequate learning, insufficient numbers of hidden units or the lack of a deterministic relationship between input and target.*"

- "*lack of a deterministic relationship between input and target*": Bayes risk is high, i.e. even ideal predictors are not performing well
- "*insufficient numbers of hidden units*": chosen model S is too small
- *"inadequate learning"*: the selected predictor within model S is suboptimal

The last point introduces estimation error.

Table of Contents

1 Approximation and Estimation errors
- Approximation error
- Estimation error
- Decomposition of the excess risk

2 Upper-bounding the estimation error in ERM
- Zero-error classification with finite models
- Beyond the zero-error case: global upper-bounds
- Non-deterministic case with finite models

Intuition

- When a model has been chosen, we will select a predictor within this model according to some procedure or algorithm.
- There is no reason to believe that we will necessarily select a best performing predictor within this model: we might select a suboptimal one.
- Question: Will the predictor selected be close (in terms of performance) to the best predictor in the chosen model?

The **discrepancy between** 1) the performance of a sample-based predictor over a model, 2) the theoretical best performance of any predictor within this model is called the estimation error.

Estimation error

Definition: Estimation error

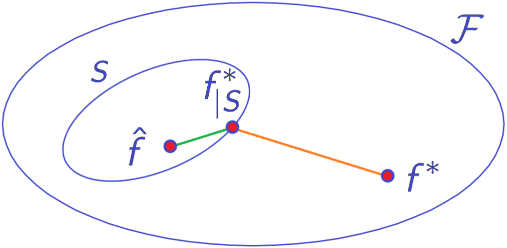
Let $S \subset \mathcal{F}$ be a model and $\hat{f}_S(D_n) \in S$ be a sample-based predictor (e.g. minimising the empirical risk over S). The quantity denoted

$$\mathcal{R}_P(\hat{f}_S(D_n)) - \inf_{f \in S} \mathcal{R}_P(f) = \ell(f^*, \hat{f}_S(D_n)) - \ell(f^*, S) \geq 0,$$

is called the estimation error of $\hat{f}_S(D_n)$. From now on D_n in $\hat{f}_S(D_n)$ is supressed and we write $\hat{f}_S$.

- The estimation error is related to the difficulty of finding/estimating a good predictor in the model S.
- More complex models makes it harder to find a good predictor.
- In regression with quadratic cost, the order of magnitude of the estimation error is typically *the number of parameters*.

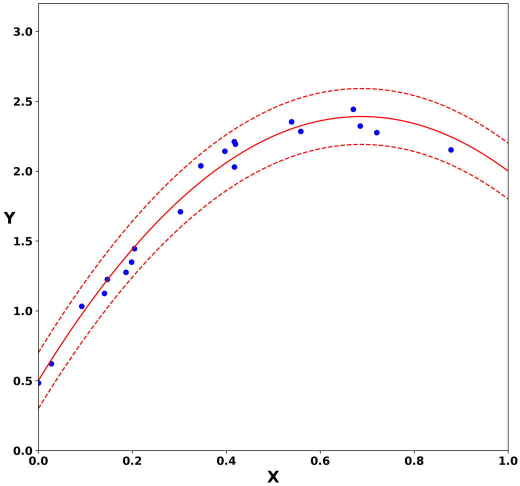
Estimation error



$\hat{f}$: Selected predictor within model S (e.g. by ERM)

$$\text{Estimation error} = \mathcal{R}_P(\hat{f}) - \inf_{f \in S} \mathcal{R}_P(f)$$

Illustration



ERM-selected linear predictor on the sample (dashed blue) vs theoretical best linear predictor (solid green)

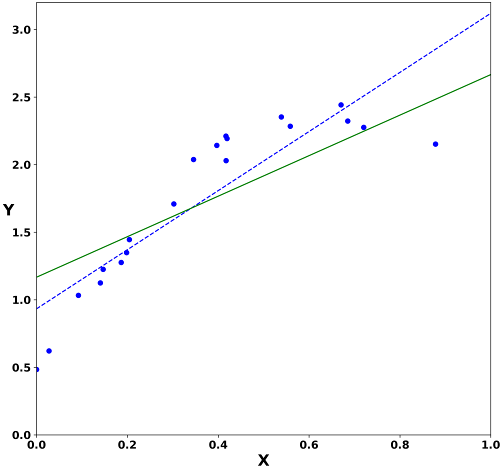


Table of Contents

1. Approximation and Estimation errors
   - Approximation error
   - Estimation error
   - Decomposition of the excess risk

2. Upper-bounding the estimation error in ERM
   - Zero-error classification with finite models
   - Beyond the zero-error case: global upper-bounds
   - Non-deterministic case with finite models

Approximation and Estimation error decomposition

Definition: Approximation and Estimation error decomposition

Let $S \subset \mathcal{F}$ be a model and $\hat{f}_S \in S$ be a sample-based predictor. Then, the excess risk of $\hat{f}_S$ can be written as:

$$\underbrace{\mathcal{R}_P(\hat{f}_S) - \mathcal{R}_P^*}_{\substack{\text{Excess risk of } \hat{f}_S \\ \text{compared to Bayes risk}}} = \underbrace{\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f)}_{\text{Estimation error}} + \underbrace{\inf_{f \in S} \mathcal{R}_P(f) - \mathcal{R}_P^*}_{\text{Approximation error}}$$

or equivalently

$$\underbrace{\ell(f^*, \hat{f}_S)}_{\substack{\text{Excess risk of } \hat{f}_S \\ \text{compared to Bayes risk}}} = \underbrace{\ell(f^*, \hat{f}_S) - \ell(f^*, S)}_{\text{Estimation error}} + \underbrace{\ell(f^*, S)}_{\text{Approximation error}}$$

Remarks

Remark: non-negativeness of the errors

As mentioned in their definitions, both the approximation and estimation errors are **non-negative**.

Remark: *Bias/Variance* decomposition

The previous decomposition is also sometimes referred to as a *bias/variance* decomposition:

The approximation error is then called *bias (of model S)*, and the estimation error is then called *variance*.

Trade-off between estimation and approximation

- The estimation error measures how far a selected (or *estimated*) predictor is from the best performance possible within the model S.

- When choosing a model S and looking for a sample-based predictor $\hat{f}_S \subset S$, we would like to minimise both errors.

Trade-off between estimation and approximation

- Intuitively: Bigger/more flexible models always allow to reduce approximation error:
  Let $S_1, S_2 \subset \mathcal{F}$ be two nested models such that $S_1 \subset S_2$. Then:
  $$\ell(f^*, S_2) \leq \ell(f^*, S_1).$$

But: Bigger/more flexible models might be complex, depend on many parameters (possibly infinitely many, e.g. non-parametric)

- Practically finding a best predictor within this model can be challenging: estimation error may be very high
- Might be prone to overfitting, mistaking noise in the data for structure

Trade-off between estimation and approximation

Approximation and estimation errors trade-off

Finding a good predictor involves adequatly choosing the model S to find a favorable trade-off between approximation and estimation errors.

- In some cases, it is possible to provide analytical expressions for the approximation error of a model.
- But in general, it is not easily accessible and other methods are used to realise the trade-off (e.g. penalisation of the empirical risk or cross-validation). See Chapter 4 of Mohri et al. (2018) for more details.

Ahead

- We will focus the estimation error in the ERM framework. Essentially, we will try to answer the question:

How close to the best performance possible within a given model can we hope a predictor selected by ERM to be?

- We will introduce several notions of flexibility (the standard term is *complexity*) of a model which will be useful to study the generalisation error of sample-based predictors.

Table of Contents

1 Approximation and Estimation errors
- Approximation error
- Estimation error
- Decomposition of the excess risk

2 Upper-bounding the estimation error in ERM
- Zero-error classification with finite models
- Beyond the zero-error case: global upper-bounds
- Non-deterministic case with finite models

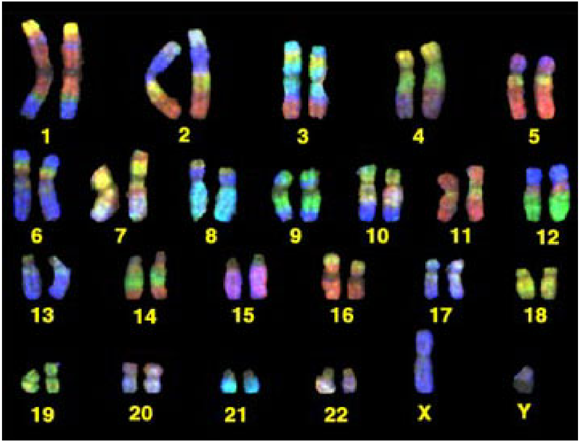
Zero-error classification

- Zero-error classification is one of the simplest frameworks in machine learning.
- Assume the label $Y \in \{0, 1\}$ is a deterministic function of the features $X \in \mathcal{X}$, i.e. there is a function $f^* \in \mathcal{F}$ such that
$$Y = f^*(X) \quad \text{almost surely.}$$
- Its generalisation error is zero and f^* is obviously a target function (an ideal predictor). With c the 0-1 cost:
$$\mathcal{R}_P(f^*) = \mathbb{E}\big[c\big(f^*(X), Y\big)\big] = \mathbb{P}\big(f^*(X) \neq Y\big) = 0.$$
- Although the label is a deterministic function of the features, the relationship may not be obvious neither theoretically, nor based on the data.

Zero-error classification: genetic example

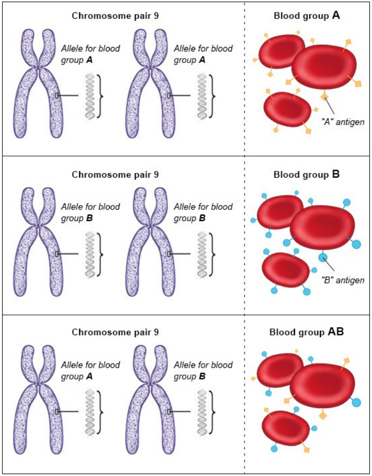
- Example: $X \in \mathcal{X}$ encodes the human genome and $Y \in \{0, 1\}$ encodes whether a person has a particular hereditary disease
- Note that the feature space is large but finite: the human genome is estimated to be composed of 20,000 genes, each having a finite number of distinct versions.
  $\Longrightarrow$ there is a (large but) finite numbers of possible predictors associating X and Y.
- With current genome sequencing techniques, it could be envisionable to gather enough data to study the relationship between X and Y.
- This can be framed as a zero-error classification problem over a finite model, for which we will show that ERM predictors can achieve a good performance.

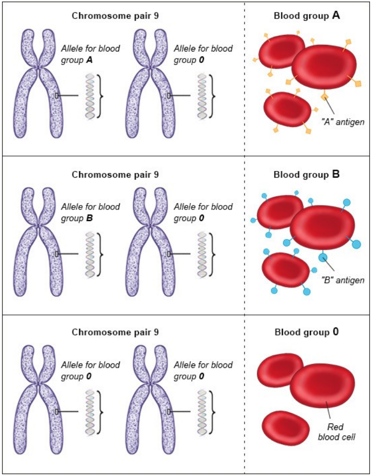
The human genome



Source: http://www.yorku.ca/kdenning/++2140%202006-7/2140-17oct2006.htm

ABO blood types : Alleles on Chromosome Pair 9





https://www.informedhealth.org/how-are-genes-passed-on.html

Finite and Infinite models

Definition: Finite and Infinite models

Let $\mathcal{X}$ be a feature space, $\mathcal{Y}$ be an output space, $\mathcal{F}$ the set of all predictors from $\mathcal{X}$ into $\mathcal{Y}$, and $S \subset \mathcal{F}$ a model.

If S contains a finite number of predictors, i.e. $\text{Card}(S) < +\infty$, then S is said to be a finite model.

If S contains an infinite number of predictors, i.e. $\text{Card}(S) = +\infty$, then S is said to be an infinite model.

Remark: Finite and Infinite models

Remark: Finite and Infinite models

The number of predictors a model contains is different in general from the *number of parameters* of the model.

For instance, the set of linear predictors from $\mathbb{R}^p$ into $\mathbb{R}$:

$$S_{\text{lin}} = \{f_{\boldsymbol{w}} : x \mapsto \boldsymbol{w}^\top x : \quad \boldsymbol{w} \in \mathbb{R}^p\},$$

is defined by p continous *parameters*, but contains an uncountably infinite number of predictors.

Upper-bound for zero-error classification with a finite model

Proposition: Upper-bound for zero-error classification

- Let $\mathcal{Y} = \{0, 1\}$ and P be a zero-error distribution in the sense that $Y = f^*(X)$ a.s. for some $f^* \in \mathcal{F}$.
- Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample, c the 0-1 cost.
- Let $S \subset \mathcal{F}$ be a finite model and $\hat{f}_S$ be an ERM learning rule over S.

Then, if $f^* \in S$, we have for any $n \geq 1$ and $\delta > 0$:

$$\mathbb{P}\left(\underbrace{\mathcal{R}_P^c(\hat{f}_S)}_{\substack{\text{Generalisation} \\ \text{error of } \hat{f}_S}} \leq \frac{\ln(\frac{1}{\delta}) + \ln(\operatorname{Card} S)}{n} \right) \geq 1 - \delta.$$

Recall that $\mathcal{R}_P^c(\hat{f}_S)$ is a random variable depending on the sample D_n.

Interpretation

Interpretation

The ERM predictor $\hat{f}_S$ is being computed based on the sample D_n. If we are lucky with the drawn sample, ERM may enable us to pick a good $\hat{f}_S$.

But we may get unlucky and only get observations dramatically oversampled from a certain region of the feature space $\mathcal{X}$. $\hat{f}_S$ will then generalise very poorly for predictions outside this region.

The proposition quantitifies the probability of $\hat{f}_S$ being a "good" predictor for a randomly drawn sample D_n, "good" in the sense that its generalisation error is below some value.

Interpretation: example

If we want $\hat{f}_S$ to have a generalisation error of at most $\varepsilon = 5\%$ in at least 99% of the cases when drawing a sample D_n at random, i.e. with probability $1 - \delta = 99\%$, then we seek

$$\mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) \leq \varepsilon\Big) \geq 1 - \delta,$$

with $\dfrac{\ln(\frac{1}{\delta}) + \ln(\mathrm{Card} S)}{n} \leq \varepsilon = 5\%$ and $\delta = 1\%$.

To achieve such requirements, the proposition tells us that we would need a sample of size n such that

$$n \geq 20\big(\ln(\mathrm{Card} S) + \ln 100\big).$$

Example: Learning a relation between genes and phenotype

- A person's genetic code can be described as a point $x \in \mathcal{X}$ in the set of all possible genomes.
- The human genome is characterised by $\sim$20,000 genes, each gene coming with a certain number of distinct possible versions called alleles, let's say at most p, for some $p \geq 1$.
- Focusing on persons for whom chromosomes go by pairs, a person's genome is thus characterised by the given of the two (possibly different) alleles of gene 1 it has, the two (possibly different) copies of gene 2 it has, etc.
- A person's genome could thus be encoded as a vector of size $40,000$ (2 alleles of each 20,000 genes), each entry being a number between 1 and p numbering the different alleles of each gene.

Example: Learning a relation between genes and phenotype

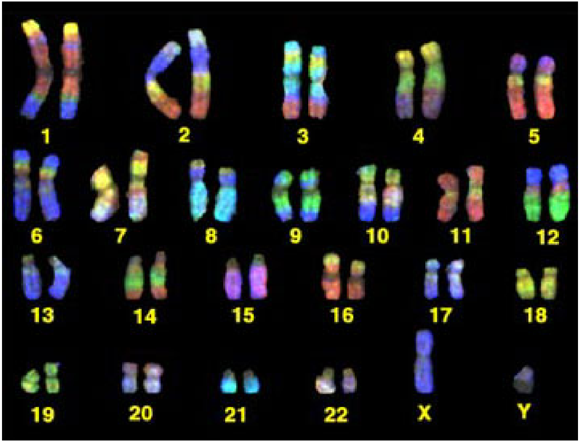
- A person's genome could thus be thought of as a point X in the (large but finite) set $\mathcal{X} = \{1, \ldots, p\}^{40000}$.
- Let's assume that a certain disease is suspected to be triggered ($Y = 1$) deterministically by a certain precise combination of at most 100 genes:

for a certain number q, $1 \leq q \leq 100$, certain distinct gene indexes $1 \leq g_1 < \ldots < g_q \leq 40000$, and certain versions $1 \leq v_1, \ldots, v_q \leq p$

$$Y = f^*(X) := \prod_{j=1}^{q} \mathbb{1}_{X_{g_j} = v_j}, \qquad \text{for } X \in \{1, \ldots, p\}^{40000}$$

The number q, the indexes, g_i's and the versions v_i's are unknown and we would like to find them.

The human genome



Source: http://www.yorku.ca/kdenning/++2140%202006-7/2140-17oct2006.htm

Example: Learning a relation between genes and phenotype

Question

How many genome examples n would we need in order to pin down the relationship between the genome and the disease with 95%-accuracy and 99%-confidence?

Framed as such, we are in a zero-error classification framework with (large but) finite model.

- Let S be the model containing all the functions of the above form.
- What is the cardinal of the set S? With some combinatorics, one can show (check it) that

$$\mathrm{Card} S = \sum_{q=1}^{100} p^q \frac{40000!}{(40000-q)!q!} = \sum_{q=1}^{100} p^q \binom{40000}{q}$$

- We would like $\hat{f}_S$ to have a generalisation error of at most $\varepsilon = 5\%$ with probability at least $1 - \delta = 99\%$ (in 99% of the cases when drawing a sample D_n at random)

- Then, by the proposition, the sample size n should satisfy

$$\frac{\ln \frac{1}{\delta} + \ln(\mathrm{Card}\, S)}{n} \leq \varepsilon,$$

- Assuming each gene has at most $p = 100$ possible versions, then $\mathrm{Card}\, S \approx 1.5 \cdot 10^{502}$, so $\ln(\mathrm{Card}\, S) \approx 1156$ and the sample size should be at least

$$n \geq \frac{1}{\varepsilon}\big[\ln(\mathrm{Card}\, S) - \ln \delta\big] \approx 23000.$$

Not such a big sample size: some companies proposing genome sequencing services had already more than 10 million clients.

Proof of the proposition (1/5).

Strategy of the proof:

1. Notice that f^* is a particular minimiser of the empirical risk, achieving zero empirical risk over the sample.
2. Deduce that any other empirical risk minimiser, in particular $\hat{f}_S$, must also achieve zero empirical risk.
3. Consider the probability $\mathbb{P}\big(\mathcal{R}_P(\hat{f}_S) \geq \varepsilon\big)$, i.e. the probability that the generalisation error of $\hat{f}_S$ is greater than some ε.
4. Bound this quantity by the probability that any other predictor $f \in S$ could have both zero empirical risk over the sample and generalisation risk greater than ε.

Proof (2/5).

Because $f^* \in S$ and we are in the zero-error case:

$$\inf_{f \in S} \mathcal{R}_P(f) = \mathcal{R}_P(f^*) = 0.$$

As $f^*(X) = Y$ a.s., it is immediately clear that $\widehat{\mathcal{R}}_n(f^*) = 0$ a.s. and thus, as risk is non-negative, $f^* \in S$ minimises the empirical risk over S.

Because $\hat{f}_S$ is by definition a minimiser of the empirical risk over S, it necessarily holds as well that

$$\widehat{\mathcal{R}}_n(\hat{f}_S) = 0, \quad a.s.$$

Proof (3/5).

Thus, for any $\varepsilon \geq 0$ we have on the one hand by the union bound:

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) \geq \varepsilon\Big) &\leq \mathbb{P}\Big(\{\exists f \in S :\ \widehat{\mathcal{R}}_n(f) = 0 \text{ and } \mathcal{R}_P(f) \geq \varepsilon\}\Big) \\
&= \mathbb{P}\Big(\bigcup_{f \in S} \{\widehat{\mathcal{R}}_n(f) = 0 \text{ and } \mathcal{R}_P(f) \geq \varepsilon\}\Big) \\
&\leq \sum_{f \in S} \mathbb{P}\big(\widehat{\mathcal{R}}_n(f) = 0 \text{ and } \mathcal{R}_P(f) \geq \varepsilon\big) \\
&= \sum_{f \in S} \mathbb{P}\big(\widehat{\mathcal{R}}_n(f) = 0\big) \mathbb{1}_{\mathcal{R}_P(f) \geq \varepsilon}.
\end{aligned}$$

Proof (4/5).

Furthermore:

$$\begin{aligned}\mathbb{P}\big(\widehat{\mathcal{R}}_n(f)=0\big) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{f(X_i)\neq Y_i}=0\right)\\ &= \mathbb{P}\big(\{\forall i=1,\dots,n:\ f(X_i)=Y_i\}\big)\\ &= \prod_{i=1}^{n}\mathbb{P}\big(f(X_i)=Y_i\big)\\ &= \prod_{i=1}^{n}\big(1-\mathcal{R}_P(f)\big)=\big(1-\mathcal{R}_P(f)\big)^n.\end{aligned}$$

Proof (5/5).

Hence

$$\begin{aligned}\mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) \geq \varepsilon\Big) &\leq \sum_{f \in S} \mathbb{P}\big(\widehat{\mathcal{R}}_n(f) = 0\big) \mathbb{1}_{\mathcal{R}_P(f) \geq \varepsilon} \\ &= \sum_{f \in S} \underbrace{\big(1 - \mathcal{R}_P(f)\big)^n \mathbb{1}_{\mathcal{R}_P(f) \geq \varepsilon}}_{\leq (1-\varepsilon)^n} \\ &\leq (1-\varepsilon)^n \operatorname{Card} S.\end{aligned}$$

Using the general inequality $1 + u \leq e^u$ for all $u \in \mathbb{R}$, we deduce:

$$\mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) \geq \varepsilon\Big) \leq e^{-\varepsilon n} \mathrm{Card} S.$$

Thus $\mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) \leq \varepsilon\Big) \geq 1 - e^{-\varepsilon n}\mathrm{Card} S$. Defining $\delta := e^{-\varepsilon n}\mathrm{Card} S$ and substituting ε in the probability concludes the proof.

Table of Contents

1 Approximation and Estimation errors
- Approximation error
- Estimation error
- Decomposition of the excess risk

2 Upper-bounding the estimation error in ERM
- Zero-error classification with finite models
- Beyond the zero-error case: global upper-bounds
- Non-deterministic case with finite models

Beyond the zero-error case

- In general, we are not in a zero-error framework
- There is likely some irreducible noise in the data that even ideal predictors cannot predict
- Can we quantify the probability that an ERM predictor will be close, in terms of generalisation error, to the best performance possible within a given model?

A global upper-bound of the estimation error in ERM

Proposition: A global upper-bound of the estimation error

Let $S \subset \mathcal{F}$ be a model (not necessarily finite) and $\hat{f}_S$ be an empirical risk minimiser over the model S. Then:

$$\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right|$$

Proof (1/1).

For any predictor $g \in S$, we have

$$\begin{aligned}
&\mathcal{R}_P(\hat{f}_S) - \mathcal{R}_P(g) \\
&\quad = \underbrace{\mathcal{R}_P(\hat{f}_S) - \widehat{\mathcal{R}}_n(\hat{f}_S)}_{\leq \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} + \underbrace{\widehat{\mathcal{R}}_n(\hat{f}_S) - \widehat{\mathcal{R}}_n(g)}_{\leq 0} + \underbrace{\widehat{\mathcal{R}}_n(g) - \mathcal{R}_P(g)}_{\leq \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|} \\
&\quad \leq 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right|
\end{aligned}$$

By taking the supremum over $g \in S$ on both sides of the inequality, and since only the term $-\mathcal{R}_P(g)$ depends on g:

$$\sup_{g \in S} \left[\mathcal{R}_P(\hat{f}_S) - \mathcal{R}_P(g) \right] = \mathcal{R}_P(\hat{f}_S) - \inf_{g \in S} \mathcal{R}_P(g) \leq 2 \sup_{f \in S} |\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)|.$$

Interpretation

Remark: Interpretation of the previous bound

The term $2\sup_{f\in S}\left|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)\right|$ is interpreted as a global measure of the complexity of the model S. This comes from the observations that:

- It is an increasing function of the model: for $S_1 \subset S_2$,

$$\sup_{f\in S_1}\left|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)\right| \leq \sup_{f\in S_2}\left|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)\right|$$

- It measures the maximum discrepancy possible between generalisation error and empirical error for predictors in S.

Table of Contents

1 Approximation and Estimation errors
- Approximation error
- Estimation error
- Decomposition of the excess risk

2 Upper-bounding the estimation error in ERM
- Zero-error classification with finite models
- Beyond the zero-error case: global upper-bounds
- Non-deterministic case with finite models

Estimation error upper-bound for ERM with finite models and bounded cost function

Proposition

- Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an iid sample
- $S \subset \mathcal{F}$ be a finite model
- c be a bounded cost function $c(y, y') \leq C$ for all $y, y' \in \mathcal{Y}$
- $\hat{f}_S$ be a predictor minimising the empirical risk over S.

Then, for any confidence level $\delta > 0$, and sample size $n \geq 1$

$$\mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) \leq \inf_{f \in S} \mathcal{R}_P(f) + C\sqrt{\frac{2\ln\frac{2}{\delta} + 2\ln\left(\text{Card } S\right)}{n}}\right) \geq 1 - \delta$$

Recall that the generalisation error $\mathcal{R}_P(\hat{f}_S)$ is a random variable because the predictor $\hat{f}_S$ depends on the sample D_n.

Proof (1/6).

Strategy of the proof:

1. Use the global bound from the previous proposition
2. Notice that for any $f \in S$, $\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)$ is a sum of independent, bounded, centered random variables
3. Apply Hoeffding's inequality
4. Rearrange terms

Proof (2/6).

Because $\hat{f}_S$ is an empirical risk minimiser over S, we have the upper-bound:

$$\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right|,$$

and it is sufficient to upper-bound the right-hand side.

For any fixed predictor $f \in S$:

$$\begin{aligned} \widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) &= \frac{1}{n} \sum_{i=1}^{n} c\big(f(X_i), Y_i\big) - \mathbb{E}\big[c\big(f(X), Y\big)\big] \\ &= \sum_{i=1}^{n} \frac{1}{n} \Big(c\big(f(X_i), Y_i\big) - \mathbb{E}\Big[c\big(f(X_i), Y_i\big)\Big] \Big), \end{aligned}$$

Hence, $\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)$ is the centered sum of n independent random variables, each with values in the interval $[0, C/n]$.

Proof (3/6).

Lemma: Hoeffding's inequality

Let $Z_1, \ldots, Z_n$ be independent bounded random variables such that for all $i = 1, \ldots, n$, $a_i \leq Z_i \leq b_i$ almost surely for some real constants a_i, b_i. Define

$$S_n = \sum_{i=1}^{n} \Big(Z_i - \mathbb{E}[Z_i] \Big).$$

Then for any $\varepsilon > 0$

$$\mathbb{P}\Big(\big| S_n \big| \geq \varepsilon \Big) \leq 2 \exp\left(- \frac{2\varepsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

Proof: Omitted (see solutions of problem 3 of problem set 1)

Proof (4/6).

Using Hoeffding's inequality with $Z_i = \frac{c(f(X_i),Y_i)}{n} \in [0, C/n]$, we get for all $z \geq 0$

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\right| \geq z\right) \leq 2\exp\left(\frac{-2z^2}{\sum_{i=1}^n (C/n)^2}\right) = 2\exp\left(\frac{-2z^2 n}{C^2}\right).$$

By the union bound:

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in S}\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\right| \geq z\right) &= \mathbb{P}\left(\bigcup_{f\in S}\left\{\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\right| \geq z\right\}\right)\\
&\leq \sum_{f\in S}\mathbb{P}\left(\left|\widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f)\right| \geq z\right)\\
&\leq \sum_{f\in S} 2\exp\left(\frac{-2z^2 n}{C^2}\right),
\end{aligned}$$

Proof (5/6).

Thus:

$$\mathbb{P}\left(\sup_{f \in S} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}_P(f) \right| \geq z \right) \leq 2 \, \mathrm{Card} S \ \exp\left(\frac{-2z^2 n}{C^2} \right).$$

Because

$$\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right|,$$

it holds for all $z \geq 0$ that

$$\begin{aligned} \mathbb{P}\left(\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \geq 2z \right) &\leq \mathbb{P}\left(2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| \geq 2z \right) \\ &\leq 2 \, \mathrm{Card} S \ \exp\left(\frac{-2z^2 n}{C^2} \right) \end{aligned}$$

Proof (6/6).

Therefore:

$$1 - \mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2z\Big) \leq 2 \operatorname{Card} S \ \exp\left(\frac{-2z^2 n}{C^2}\right)$$
$$\iff \mathbb{P}\Big(\mathcal{R}_P(\hat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f) \leq 2z\Big) \geq 1 - 2 \operatorname{Card} S \ \exp\left(\frac{-2z^2 n}{C^2}\right).$$

Finally, defining

$$\delta := 2 \operatorname{Card} S \ \exp\left(\frac{-2z^2 n}{C^2}\right),$$

and substituting for z in the left-hand side yields the conclusion.

Interpretation

Interpretation

In this setting there is no deterministic relationship between features and output.

The proposition quantifies the probability of $\hat{f}_S$ being a "good" predictor for a randomly drawn sample D_n, "good" in the sense that its generalisation error does not exceed the lowest achievable risk over S by more than some specified value.

The interpretation is thus very similar to that for the zero-error case, but the result is less favourable regarding the performance of ERM predictors.

Interpretation: a much weaker guarantee

- In the zero-error case, the requirement to guarantee that $\hat{f}_S$ has a generalisation error of at most $\varepsilon = 5\%$ in at least 99% of the cases when drawing a sample D_n at random ($\delta = 1\%$) was

$$n \geq \frac{1}{\varepsilon}\big(\ln(\mathrm{Card}\, S) + \ln(1/\delta)\big) = 20\big(\ln(\mathrm{Card}\, S) + \ln 100\big)$$

- However, in the setting of the last proposition, asking for similar requirements leads to

$$C\sqrt{\frac{2\ln\frac{1}{\delta} + 2\ln\left(2\mathrm{Card}\, S\right)}{n}} \leq \varepsilon$$

that is (using that $C = 1$ for the 0-1 cost):

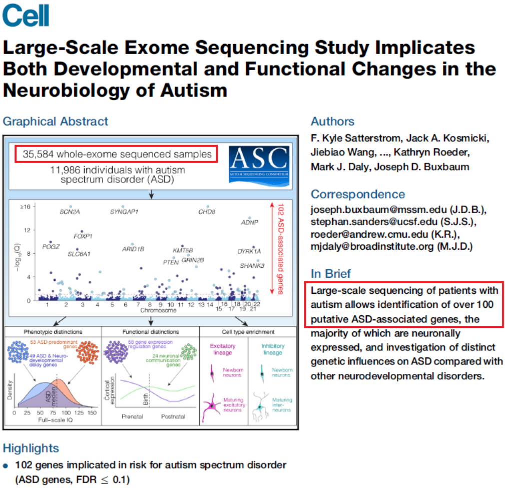
$$n \geq \frac{2}{\varepsilon^2}\big(\ln(2\mathrm{Card}\, S) + \ln(1/\delta)\big) = 800\big(\ln(2\mathrm{Card}\, S) + \ln(100)\big)$$

Interpretation: a less favourable guarantee

A much weaker bound

- To achieve the *same generalisation guarantees*, the sample needs to be more than 40 times larger than in the zero-error case.
- In the genetics example given before, this would mean that to identify a potentially non-deterministic relationship between gene combinations and disease (for instance if non-accounted environmental factor played a role), we would need more than 929 000 examples instead of 23 000.

Published in *Cell* (2020): 102 genes linked to autism identified based on $n \approx 35,000$



Cell

Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism

Graphical Abstract

Authors

F. Kyle Satterstrom, Jack A. Kosmicki, Jiebiao Wang, ..., Kathryn Roeder, Mark J. Daly, Joseph D. Buxbaum

Correspondence

joseph.buxbaum@mssm.edu (J.D.B.),
stephan.sanders@ucsf.edu (S.J.S.),
roeder@andrew.cmu.edu (K.R.),
mjdaly@broadinstitute.org (M.J.D.)

In Brief

Large-scale sequencing of patients with autism allows identification of over 100 putative ASD-associated genes, the majority of which are neuronally expressed, and investigation of distinct genetic influences on ASD compared with other neurodevelopmental disorders.

Highlights

- 102 genes implicated in risk for autism spectrum disorder (ASD genes, FDR $\leq$ 0.1)

Wrap-up

We have looked at the decomposition of excess risk into:

- Approximation error, discrepancy performance ideal predictor and theoretically best predictor in model,
- Estimation error, discrepancy between performance of sample-based predictor and theoretically best predictor in model.

We have established learning guarantees for:

- Zero-error classification for finite models,
- Non-deterministic setting with bounded cost and finite models.

Next week:

- We will see how to quantify complexity of infinite models,
- Learning guarantee for infinite models.

Assignment

- You can now make questions 1 and 2 of Assignment Part II
- I encourage you to already start working on it (deadline of Part II is March 18th at 23:59)
- **Important:** recall that the deadline of Assignment Part I is this Thursday (February 29th) at 23:59