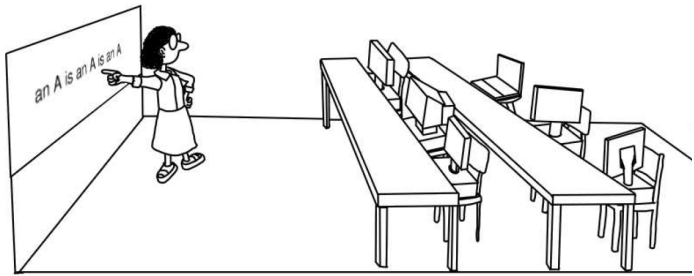
EE3-23: Machine Learning

Krystian Mikolajczyk & Deniz Gunduz

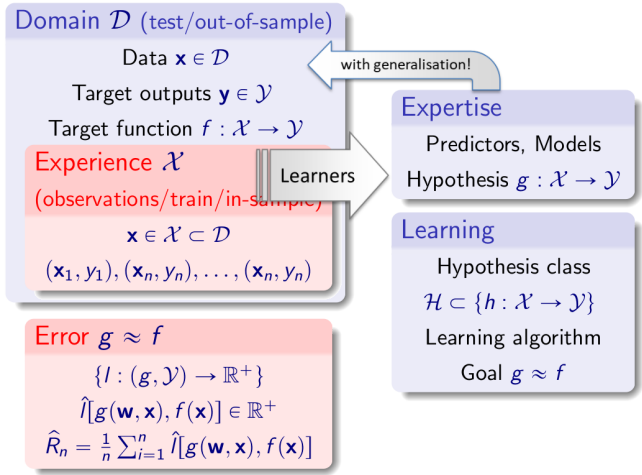
Department of Electrical and Electronic Engineering
Imperial College London



Part 2 Summary

- Feasibility of learning
- Hoeffding's inequality
- Target distribution and error cost
- Multiple hypothesis
- Growth function
- VC inequality
- Bias-variance trade-off

How can we learn?



$$P\left(|p_{event} - e_{event}| > \varepsilon\right) \leqslant \delta$$

Hoeffding's inequality

$$P\left(|R(h) - \widehat{R}_n(h)| > \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}$$

How can we learn?

- Is finding unknown target $f : \mathcal{X} \to \mathcal{Y}$ possible?
  - $N + 1$ sample can contradict the found target function f.
- Is learning possible?

$$P\left(|R(h) - \widehat{R}_n(h)| > \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}$$

Hoeffding's inequality  Vapnik-Chervonenkis inequality

Learning / generalisation theory

$$P\left(\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)|\right) > \varepsilon) \leqslant 4 \underbrace{m_{\mathcal{H}}(2n)}_{\leqslant (2n+1)^{dVC(\mathcal{H})}} e^{-\varepsilon^2 n/8}$$

Relation of $P(event)$ and $\mathbb{E}[event]$

We expect $p_{event} \approx e_{event}$, where the true function $p_{event} = P(event)$ and expectation $e_{event} = \mathbb{E}[event]$. Is this true?

- They can be very different.
- Likely to be true! e.g. Polls.

Hoeffding's inequality

For $\varepsilon > 0$,

$$P(e_{event} - p_{event} > \varepsilon) \leqslant \delta$$

$$P(e_{event} - p_{event} > \varepsilon) \leqslant e^{-2\varepsilon^2 n} \qquad \text{(one-sided)}$$

$$P(|e_{event} - p_{event}| > \varepsilon) \leqslant 2e^{-2\varepsilon^2 n} \qquad \text{(two-sided)}$$

The statement $e_{event} = p_{event}$ is Probably Approximately Correct (PAC).

Relation to learning

For a fixed hypothesis $h \in \mathcal{H}$:

- Training (in sample) error (empirical risk): $e_{event} = \widehat{R}_n(h)$
  - $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(h(\mathbf{x}_i) \neq f(\mathbf{x}_i)\right)$;
- Test (out of sample) error (risk): $p_{event} = R(h)$.
  - $R(h) = P\left(h(\mathbf{x}_i) \neq f(\mathbf{x}_i)\right)$;
  - $R(h) = \mathbb{E}\left[\widehat{R}_n(h)\right]$.

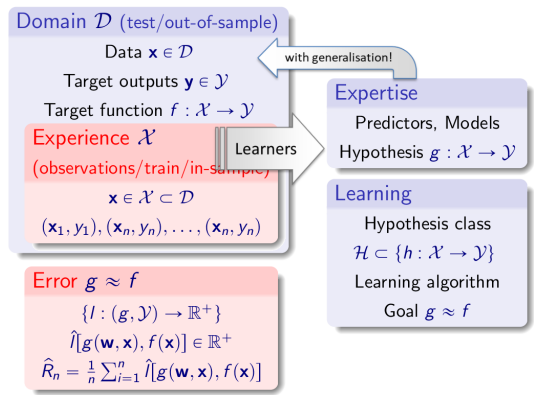
By Hoeffding's inequality

$$P\left(|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}.$$

What assumptions do we make?

Extensions

1. i.i.d.
2. Cost of error
3. Target distribution: is target function a function?
4. Fixed Hypothesis: does Hoeffding work for multiple h?



The i.i.d. assumption

- Input: $\mathbf{x} \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Target function $f : \mathcal{X} \to \mathcal{Y}$
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

  Assume the x_i are drawn independently from a distribution $P(\mathcal{X})$.

  i.i.d.: independent and identically distributed

Learning

- Hypothesis class: $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$
- Find $g \in \mathcal{H}$ such that $g \approx f$: $P(g(x) \neq f(x))$ is small where $\mathbf{x} \sim P(\mathcal{X})$.

Target distribution: Error Measures/Loss Functions

- How to quantify $h \approx f$?
- Usually pointwise error: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$

$$\ell(h(\mathbf{x}), f(\mathbf{x}))$$

Defined by the user or convenience!

- Examples:

squared error	$\ell(\hat{y}, y) = (\hat{y} - y)^2$
binary error	$\ell(\hat{y}, y) = \mathbb{I}(\hat{y} \neq y)$

- Training error: $\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)$
- Test error: $R(h) = \mathbb{E}[\ell(h(\mathbf{x}), y)]$

Target distribution: Error cost

Two types of error:

		f = +1	f = -1
h	+1	no error	false accept
	-1	false reject	no error

How do we penalize them?

- Equally: $\mathbb{I}(h(x) \neq f(x))$

		f = +1	f = -1
h	+1	0	1
	-1	1	0

- Aggressive: false negative is expensive

		f = +1	f = -1
h	+1	0	1
	-1	100	0

- Risk averse: false positive is expensive

		f = +1	f = -1
h	+1	0	1000
	-1	1	0

Target distribution: Learning problem

- Instead of assuming deterministic $y = f(\mathbf{x})$, y may be probabilistic: $y \sim P(y|\mathbf{x})$
  - allow the same inputs to have different labels
- The data points $(\mathbf{x}, y)$ are generated from $P(\mathbf{x}, y) = P(y|\mathbf{x})P(\mathbf{x})$: $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$
- Noise interpretation:

  $$y = f(\mathbf{x}) + \textit{noise} \text{ where } f(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] \text{ and } \mathbb{E}[\textit{noise}|\mathbf{x}] = 0.$$

  Example: $y = w^\top \mathbf{x} + N$ where $N \sim \mathcal{N}(0, \Sigma)$ is independent of $\mathbf{x}$.
- Deterministic is a special case: $\textit{noise} = 0$ and $P(y|\mathbf{x})$ is concentrated on the single point $f(\mathbf{x})$.

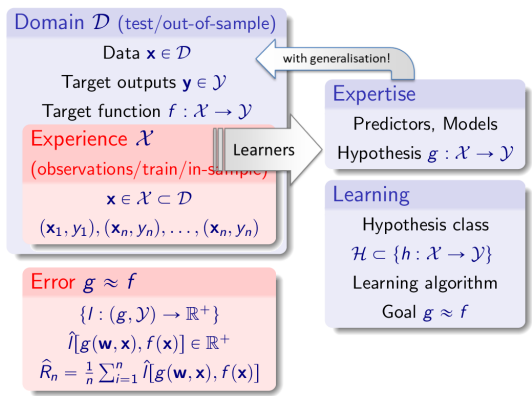
Learning problem

Learning $P(y|\mathbf{x})$.

Target distribution: Optimal Decisions

Optimal h depends on the noise and the loss:

- Squared error: $\ell(\hat{y}, y) = (\hat{y} - y)^2$
  - $g(\mathbf{x})$ minimizes $\mathbb{E}\left[(y - h(\mathbf{x}))^2 | x\right]$
  - $g(\mathbf{x}) = \mathbb{E}\left[y|\mathbf{x}\right]$
- Binary error: $\ell(\hat{y}, y) = \mathbb{I}\left(y \neq \hat{y}\right)$
  - $g(\mathbf{x}) = \text{argmax}_y\, P(y|\mathbf{x})$



Learning Setup with $\mathbf{x} \sim P$ and $P(y|\mathbf{x})$

- Input: $\mathbf{x} \in \mathcal{X}$
- Output: $y \in \mathcal{Y}$
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n) \sim P$

(P is the joint distribution of $(\mathbf{x}, y)$)

Learning

- Hypothesis class: $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- Find $g \in \mathcal{H}$ such that $g \approx P(y|\mathbf{x})$

How should we choose $\mathcal{H}$?

$$g = \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ R(h) = \mathbb{E}\left[\ell(h(\mathbf{x}), y)\right] \right\}$$

Different aggregation (not expectation) may be needed, e.g., when P is too imbalanced.

Relation to Hoeffding (fixed hypothesis)

For a fixed hypothesis $h \in \mathcal{H}$:

- Training (in sample) error (empirical risk): $e_{event} = \widehat{R}_n(h)$
  - $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(h(\mathbf{x}_i) \neq f(\mathbf{x}_i)\right)$;
- Test (out of sample) error (risk): $p_{event} = R(h)$.
  - $R(h) = P\left(h(\mathbf{x}_i) \neq f(\mathbf{x}_i)\right)$;
  - $R(h) = \mathbb{E}\left[\widehat{R}_n(h)\right]$.

By Hoeffding's inequality

$$P\left(|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}.$$

What assumptions do we make?

Single vs multiple hypotheses

Letting $\mathcal{H} = \{h_1, \ldots, h_M\}$:

$$
\begin{aligned}
& P\left(\max_{h \in \mathcal{H}} |\widehat{R}_n(h) - R(h)| > \varepsilon\right) \\
& = P\left(|\widehat{R}_n(h_1) - R(h_1)| > \varepsilon \text{ or } \ldots \text{ or } |\widehat{R}_n(h_M) - R(h_M)| > \varepsilon\right) \\
& \leqslant \sum_{m=1}^{M} P\left(|\widehat{R}_n(h_m) - R(h_m)| > \varepsilon\right) \\
& \leqslant 2Me^{-2\varepsilon^2 n}
\end{aligned}
$$

For any g, selected in any way based on the data

$$P\left(|\widehat{R}_n(g) - R(g)| > \varepsilon\right) \leqslant 2Me^{-2\varepsilon^2 n} .$$

Generalisation for hypotheses class (multiple hypotheses)

For all $h \in \mathcal{H}$, simultaneously

$$P\left(|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 2Me^{-2\varepsilon^2 n}$$

Define $\delta = 2Me^{-2\varepsilon^2 n} \quad \Rightarrow \quad \varepsilon = \sqrt{\frac{\log \frac{2M}{\delta}}{2n}}$, and so:

For all $h \in \mathcal{H}$, simultaneously with probability at least $1 - \delta$,

$$|\widehat{R}_n(h) - R(h)| \leqslant \sqrt{\frac{\log \frac{2M}{\delta}}{2n}} \ .$$

Bound for the difference between error on training data and error on test data, for any give n h

Empirical Risk Minimization (ERM)

Let $h^* \in \mathcal{H}$ be the optimal hypothesis in $\mathcal{H}$: $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$.

Choose g to be the best hypothesis with the smallest empirical error (the empirical risk minimizer):

$$g = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}_n(h) \ .$$

Risk of the empirical risk minimizer

With probability at least $1 - \delta$,

$$R(g) - R(h^*) \leqslant \sqrt{\frac{2 \log \frac{2M}{\delta}}{n}} \ .$$

Feasibility of learning: summary so far

- Learning an arbitrary unknown function: not possible
  - $N + 1$ data sample
  - instead learn: Target distribution $P(y|\mathbf{x})$ with data distribution $\mathbf{x} \sim P(\mathcal{X})$
- Learning under probabilistic assumptions – i.i.d. sample

Training error: $\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(h(x_i) \neq f(x_i)\right)$

Test error: $R(h) = P\left(h(x_i) \neq f(x_i)\right)$

Guarantees:

- For any fixed $h \in \mathcal{H}$,

$$P\left(|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}$$

- For any $g \in \mathcal{H}$ which may depend on the sample $\left(\text{e.g., } g = \text{argmin}_h \widehat{R}_n(h)\right)$,

$$P\left(|\widehat{R}_n(g) - R(g)| > \varepsilon\right) \leqslant 2|\mathcal{H}|e^{-2\varepsilon^2 n}$$

- Can we learn infinite function classes?

Example: Linear classification

- Features: $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$

  $\mathbf{x} = (x_0, x_1, \ldots, x_d) \in \mathbb{R}^{d+1}$ (with $x_0 = 1$).
- Labels: $y \in \{+1, -1\}$.
- Data points: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
- Hypothesis class:

  $h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$, with $\mathbf{w} = (w_0, w_1, \ldots, w_d)$

Perceptron finds $g \in \mathcal{H}$ such that $g(\mathbf{x}_i) = y_i$ for all $i = 1, \ldots, n$

assuming it exists

$$g \in \operatorname*{argmin}_{h \in \mathcal{H}} \underbrace{\sum_{t=1}^{n} \mathbb{I}\left(h(\mathbf{x}_t) \neq y_t\right)}_{\widehat{R}_n(h)}$$

minimize $\widehat{R}_n(h_{\mathbf{w}})$ in $\mathbf{w}$

Empirical Risk Minimization

Is $|\mathcal{H}|$ finite? Does the theory apply?

Overlapping hypotheses - real case scenario

Hypotheses are overlapping!

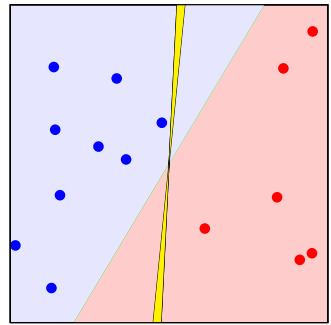
If $h_1 \approx h_2$ then:

$$\widehat{R}_n(h_1) \approx \widehat{R}_n(h_2)$$

$$R(h_1) \approx R(h_2)$$

Thus

$$|\widehat{R}_n(h_1) - R(h_1)| \approx |\widehat{R}_n(h_2) - R(h_2)|$$



$|\widehat{R}_n(h_1) - R(h_1)| > \varepsilon$ often implies $|\widehat{R}_n(h_2) - R(h_2)| > \varepsilon$

$h_1 \neq h_2$ if $\exists \mathbf{x}_i \in \mathcal{X} : h_1(\mathbf{x}_i) \neq h_2(\mathbf{x}_i) \rightarrow$ dichotomy

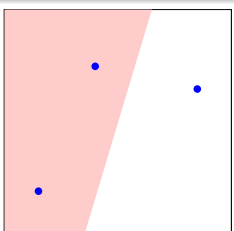
Growth Function - Perceptron

Use data samples $\mathbf{x}$ instead of entire input space $\mathcal{X}$.

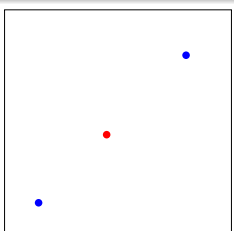
How many ways can we partition data points? Number of dichotomies?

number of hypotheses $|\mathcal{H}(\mathcal{X})| \gg |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)|$ number of dichotomies

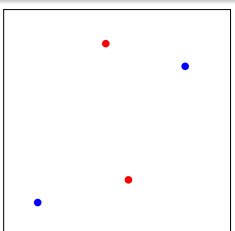
Growth function: $m_{\mathcal{H}}(n) = \max_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n)| \leqslant 2^n$ shattered



$n = 3$, in general position



$n = 3$, colinear



$n = 4$, in general position

$m_{\mathcal{H}}(3) = 8$

$m_{\mathcal{H}}(4) = 14 < 16$

Break point (capacity of $\mathcal{H}$, $\exists k : m_{\mathcal{H}(k)} \nexists (\mathbf{x}_1, \ldots, \mathbf{x}_k) \in \mathcal{X}$: shattered by $\mathcal{H}$)

Growth Function - Perceptron

Hoeffding's inequality

$$P\left(|\widehat{R}_n(g) - R(g)| > \varepsilon\right) \leqslant 2Me^{-2\varepsilon^2 n}$$

Replace $M = \infty$ with $m_{\mathcal{H}}(k) \rightarrow P\left(|\widehat{R}_n(g) - R(g)| > \varepsilon\right) \leqslant 2m_H(k)e^{-2\varepsilon^2 n}$

- no break point $\rightarrow m_{\mathcal{H}}(n) = 2^n$
- any break point $\exists k \rightarrow m_{\mathcal{H}}(k)$ is **polynomial** in n.
  $m_{\mathcal{H}}(k) = a_k n^k + a_{k-1} n^{k-1} +, \ldots, a_1 n + a_0$
- e^{-n} and large n data points will then reduce the whole probability
  Generalisation possible with probability assurance!
- no need to know k, only that it exists
- use $m_{\mathcal{H}}(2n)$, why? $|\widehat{R}_n(g) - R(g)| \approx |\widehat{R}_{n^{(1)}}(g) - \widehat{R}_{n^{(2)}}(g)|$

Generalisation Bounds

Extended Hoeffding's Inequality

$$P\left(\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-R(h)|>\varepsilon\right)\leqslant 2\underbrace{m_{\mathcal{H}}(n)}_{incorrect:R(h)missing}e^{-2\varepsilon^2 n}$$

Vapnik-Chervonenkis Inequality

$$P\left(\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-R(h)|>\varepsilon\right)\leqslant 4\underbrace{m_{\mathcal{H}}(2n)}_{\leqslant(2n+1)^{d_{VC}(\mathcal{H})}}e^{-\varepsilon^2 n/8}$$

where $d_{VC}(\mathcal{H})=\max\{n : m_{\mathcal{H}}(n)=2^n\}$ is the VC-dimension

The most important statement in theoretical machine learning

Vapnik-Chervonenkis Inequality

$$P\left(\sup_{h\in\mathcal{H}}|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 4 \underbrace{m_{\mathcal{H}}(2n)}_{\leqslant(2n+1)^{d_{VC}(\mathcal{H})}} e^{-\varepsilon^2 n/8}$$

where $d_{VC}(\mathcal{H}) = \max\{n : m_{\mathcal{H}}(n) = 2^n\}$ is the VC-dimension

- $m_{\mathcal{H}}(n+1) = 2^n$, $n+1$ is first break point, $n+2$ is also a break point
- max points that can be shattered $\approx$ "effective number of parameters"
- order of the polynomial that bounds $\mathcal{H}$: $m_{\mathcal{H}} \leqslant \sum_{i=0}^{d_{VC}} \binom{n}{i} \approx n^{d_{VC}}$
- independent of learning algorithm because $g \in \mathcal{H}$
- independent of input distribution p on $\mathcal{X}$ i.e. $\mathbf{x} \sim p$, only n matters
- independent of target distribution $P(y|\mathbf{x})$
- it concerns g and $\mathcal{H}$ and $(\mathbf{x}_i, \ldots, \mathbf{x}_n) \sim p$
- if d_{VC} is finite $\Rightarrow g \in \mathcal{H}$ will generalise with probability δ

Generalisation Bounds

Vapnik-Chervonenkis Inequality:

$$P\left(\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-R(h)|>\varepsilon\right)\leqslant 4\underbrace{m_{\mathcal{H}}(2n)}_{\leqslant(2n+1)^{d_{VC}(\mathcal{H})}}e^{-\varepsilon^2 n/8}$$

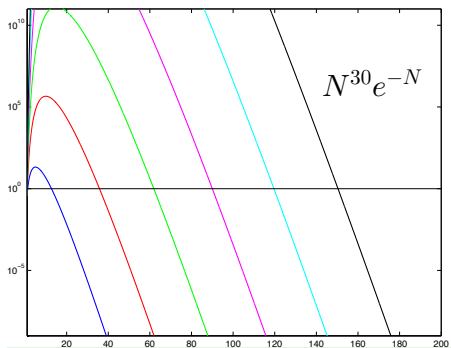
where $d_{VC}(\mathcal{H})=\max\{n : m_{\mathcal{H}}(n)=2^n\}$ is the VC-dimension $\approx$ "effective number of parameters".

Example: for linear classification in d dimension, $d_{VC}=d+1$, $(w_0,\ldots,w_d)$

Figure: relation $n^{d_{VC}(\mathcal{H})}e^{-n}$.

Change $\mathcal{H}$ or n to make $P(.)$ small! N is proportional to d_{VC}

Rule of thumb: $n\geqslant 10d_{VC}(\mathcal{H})$.

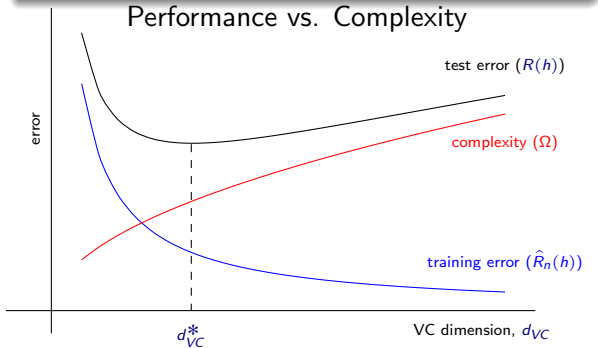


Generalisation Error

With probability at least $1-\delta$, for all $h \in \mathcal{H}$ simultaneously,

$$R(h) \leqslant \widehat{R}_n(h) + \underbrace{\sqrt{\frac{8 d_{VC}(\mathcal{H})}{n} \log(2n+1) + \frac{8}{n} \log \frac{4}{\delta}}}_{\Omega(n,\mathcal{H},\delta)}$$

$$R(h) \leqslant \widehat{R}_n(h) + \Omega(n, \mathcal{H}, \delta)$$



Polynomial Bounds on growth function

Different bounds lead to different approximations of $\Omega(n, \mathcal{H}, \delta)$

Polynomial Bounds on growth function

- $m_{\mathcal{H}}(k)$ is **polynomial** in n, with break point $k+1$, $d_{VC} = k$
- $m_{\mathcal{H}}(n,k) = a_k n^k + a_{k-1} n^{k-1} +, \ldots, a_1 n + a_0$
  - inconvenient to use, so "nicer" bounds needed
- Order of the polynomial that bounds $\mathcal{H}$: $m_{\mathcal{H}}(n, d_{VC}) \leqslant \sum_{i=0}^{d_{VC}} \binom{n}{i} \approx n^{d_{VC}}$
- If $d_{VC}(n) < \infty$, then for all n:

$$m_{\mathcal{H}}(n) \leqslant n^{d_{VC}} + 1 \leqslant (n+1)^{d_{VC}}$$

and for all $n \geqslant d_{VC}$, an improved bound is:

$$m_{\mathcal{H}}(n) \leqslant \left(\frac{ne}{d_{VC}}\right)^{d_{VC}} \leqslant n^{d_{VC}} + 1$$

Practical ML scenario

HMRC is considering using ML to identify suspicious tax return cases. Overall, it estimates about $\$4.4 \cdot 10^9$ in taxes is lost due to tax evasion each year. The average cost of investigating a taxpayer is approximately $\$10^4$. There are approximately 10 million taxpayers submitting their own tax returns, which are in structured form consisting of 100 fields, that can be converted into real value numbers. There are approx 4400 tax evasion cases every year and their records from the past 10 years are available. HMRC will find ML useful if it can guarantee that the test error will not differ from training by more than 20% with 99% certainty.

- Identify relevant ML components and formulate it as an ML problem.
- What is required to guarantee that the predictor meets HMRC criteria?

Practical ML scenario

$\mathbf{x}_i \in \mathbb{R}^{100}$ – there are $n_p = 44000$ positive data points and $100M$ in total

To learn, we should choose similar number of negative examples from $100M$, e.g. $n_n = 44000$, thus $n = 88000$

$f(\mathbf{x}_i) = y$, $y \in \{-1, 1\}$ – binary classification problem

$\widehat{R}_n(h) = \frac{1}{n} \sum_n \mathbb{I}\left(g(x_i) \neq y\right)$ – simple error function

H – hypothesis class with $d_{VC} \leqslant 8800$, eg. polynomial of degree k and linear classifier

ERM $g = \operatorname{argmin}_{h \in H} \widehat{R}_n(h)$ – algorithm to find the best predictor g, so PLA

Better loss:

false negative – cost of not finding evasion: $= (4.4 \cdot 10^9)/(4.4 \cdot 10^3) \approx 10^6$

false positive – cost of investigating a tax return $= 10^4$

false negative leads to 100 times higher cost than false positive

therefore better loss: $\mathbb{I}\left(g(x_i) \neq y\right) = 100$ if $y = 1$ otherwise $\mathbb{I}\left(g(x_i) \neq y\right) = 1$ if $y = -1$

Practical ML scenario

error $\varepsilon = 0.2$, $n = 88000$, $P(\cdot) = 0.99$,

choose H with d_{VC} according to VC inequality

From VC inequality, the test error can be larger than the training error with probability 0.01

$$P\left(|\widehat{R}_n(h) - R(h)| > \varepsilon\right) \leqslant 4n^{d_{VC}} e^{-\varepsilon^2 n/8} = 4 \cdot 88000^{d_{VC}} e^{-0.2^2 88000/8} \approx 0.01$$

hence $d_{VC} \leqslant 39$.

No need to derive $d_{VC} = \ldots$, try a few numbers 5, 50, etc and you see if you need to reduce or increase.

Bias-Variance Trade-Off

VC analysis: test error $\leqslant$ training error + complexity penalty

Another approach: **bias-variance** analysis:

test error = bias + variance

- Bias: how well can $\mathcal{H}$ approximate f? (as before)
- Variance: how well can we select a good $h \in \mathcal{H}$?

Setup:

- e.g. $\mathbf{x}$ - patient record, $\mathcal{D}$ - a hospital, y - cost prediction.
- $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $y_i = f(\mathbf{x}_i) \in \mathbb{R}$.
- Test error within $\mathcal{D}$ (squared):
  $R(g^{(\mathcal{D})}) = \mathbb{E}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \middle| \mathcal{D}\right] = \mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$

Bias-Variance Analysis

Test error within $\mathcal{D}$:

$$R(g^{(\mathcal{D})}) = \mathbb{E}_x \left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

Expected test error (over many $\mathcal{D}_1, \ldots, \mathcal{D}_K$) and $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^{(\mathcal{D})}$:

$$\begin{aligned} \mathbb{E}\left[R(g^{(\mathcal{D})}) \right] &= \mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{\mathbf{x}} \left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \right] \\ &= \mathbb{E} \left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_x \left[\mathbb{E}_{\mathcal{D}} \left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \right] \end{aligned}$$

The Average Hypothesis

Concentrate on $\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$ for a given $\mathbf{x} \in \mathcal{X}$ (patient x!).

Average hypothesis over many datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$

The best possible : $\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right] \approx \frac{1}{K}\sum_1^K g^{(\mathcal{D}_k)}(\mathbf{x})$

Expected error for patient $\mathbf{x}$ with costs predicted by many $g^{(\mathcal{D})}(\mathbf{x})$

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] &= \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x}))^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\Big[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2 + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \\
&\qquad + 2(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))(\bar{g}(\mathbf{x}) - f(\mathbf{x}))\Big] \\
&= \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right] + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \\
&\qquad + 2\underbrace{\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))(\bar{g}(\mathbf{x}) - f(\mathbf{x}))\right]}_{2(\bar{g}(\mathbf{x}) - \bar{g}(\mathbf{x}))(const)}
\end{aligned}$$

Bias and Variance

$$\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right]}_{\text{var}(\mathbf{x})} + \underbrace{(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2}_{\text{bias}(\mathbf{x})} .$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[R(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})\right] \\
&= \text{bias} + \text{var} .
\end{aligned}$$

$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2\right]$

- how far $\bar{g}(\mathbf{x})$ from $f(\mathbf{x})$
- large if $\mathcal{H}$ is small
- small if $\mathcal{H}$ is large

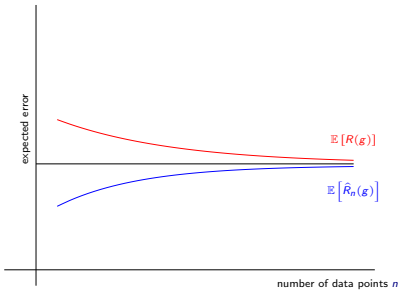
$\text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})^2\right]\right]$

- how far $g^{(\mathcal{D})}(\mathbf{x})$ from $\bar{g}(\mathbf{x})$
- small if $\mathcal{H}$ is small
- large if $\mathcal{H}$ is large

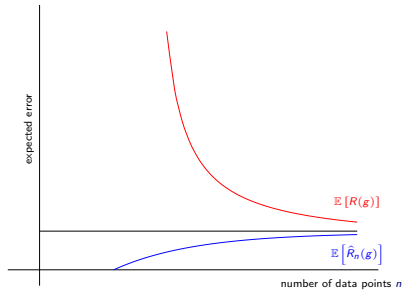
Complexity-Performance Trade-off

Match the model complexity to the data not to the target complexity!

Learning curves:

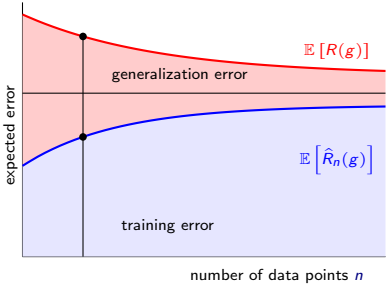


simple model

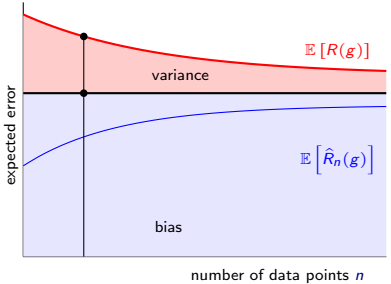


complex model

VC vs Bias-Variance



VC analysis



bias-variance

- best approximation in between $\mathbb{E}[R(g)]$ and $\mathbb{E}\left[\widehat{R}_n(g)\right]$
- in VC the error is on the training sample
- bias based on the best approximation $\bar{g}(x)$ (over all $^{(\mathcal{D})}$)
- bias constant, only depends on $\mathcal{H}$ not on n

Terminology

$\mathbb{R}$	the set of real numbers
$\mathbb{R}_+$	the set of non-negative real numbers
$\mathbb{N}$	the set of natural numbers
$\mathbf{x}, \mathbf{w}$	(column) vectors
$\mathbf{x}$	typically input data
$\mathbf{w}$	typically hypothesis (model) parameters
$\|\mathbf{x}\|_2^2$	ℓ_2 norm of $\mathbf{x} = \mathbf{x}^\top \mathbf{x}$
$\|\mathbf{x}\|_2$	ℓ_2 norm of $\mathbf{x} = \sqrt{\mathbf{x}^\top \mathbf{x}}$
$\|\mathbf{x}\|_1$	ℓ_1 norm of $\mathbf{x} = \sum_i \|x_i\|$
$\mathbb{E}_{\mathbf{x}}\left[g(\mathbf{x})\right]$	expected value of $g(\mathbf{x})$ over $\mathbf{x}$
$\frac{1}{n}\sum_i g_i(\mathbf{x})$	an estimate of expected value with prob. by Hoeffding
$\operatorname{argmin}_{\lambda \in \Lambda} g(\lambda)$	argument λ for which $g(\lambda)$ reaches minimum
$\operatorname{argmax}_{\lambda \in \Lambda} g(\lambda)$	argument λ for which $g(\lambda)$ reaches maximum

$\log$	the natural logarithm
$\mathbb{P}$, P	probability
h	hypothesis, predictor
$\mathcal{H}$	hypothesis class
g	best hypothesis trained on data
$h \sim g$	h is similar to g, close approximation of g
$\mathbf{x} \sim P$	$\mathbf{x}$ is sampled from P, i.i.d. according to P
$\mathbb{I}(x)$	$\mathbb{I}(x) = 1$ if $x =$ *true*, $\mathbb{I}(x) = 0$ if $x =$ *false*
$R()$	true error on all data, unknown
$\widehat{R}()$	empirical error on training data
$\breve{R}()$	validation error on validation data
$\mathcal{L}_n()$	loss on available data
$f(\mathbf{x})$	target function, unknown, modelled by g

Terminology

- Training set: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Test set: $\mathcal{D}' = \{(x'_1, y'_1), \ldots, (x'_m, y'_m)\}$.
- Loss function: $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

	statistics	learning theory	machine learning
$\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$	in-sample error E_{in}	empirical risk $\widehat{R}_n, L_n, \mathcal{L}_n$	training error
$\mathbb{E}\left[\ell(h(x), y)\right]$ $\mathbb{E}\left[\ell(g^{(\mathcal{D})}(x), y) \middle\vert \mathcal{D}\right]$	out-of-sample error E_{out}	risk, generalization error $R, L, \mathcal{L}$	(true) test error
$\mathbb{E}\left[\ell(g^{(\mathcal{D})}(x), y)\right]$	expected out-of-sample error	expected risk	expected test error
$\frac{1}{n}\sum_{i=1}^{m} \ell(h(x'_i), y'_i)$	test error E_{test}	empirical test error $\widehat{R}'_m$	(empirical) test error

Relations:

- $R(h) = \mathbb{E}\left[\widehat{R}_n(h)\right]$ + high prob. by Hoeffding
- $R(g^{(\mathcal{D})})$ and $\widehat{R}_n(g^{(\mathcal{D})})$: typically $\mathbb{E}\left[\widehat{R}_n(g^{(\mathcal{D})}) \middle\vert \mathcal{D}\right] \neq R(g^{(\mathcal{D})})$ but h.p. by VC inequality
- $R(h) = \mathbb{E}\left[\widehat{R}'_m(h)\right]$ and $R(g^{(\mathcal{D})}) = \mathbb{E}\left[\widehat{R}'_m(g^{(\mathcal{D})}) \middle\vert \mathcal{D}\right]$; h.p. by Hoeffding (in both cases!)

Part 2 Summary

- Feasibility of learning
- Hoeffding's inequality
- Target distribution and error cost
- Multiple hypothesis
- Growth function
- VC inequality
- Bias-variance trade-off