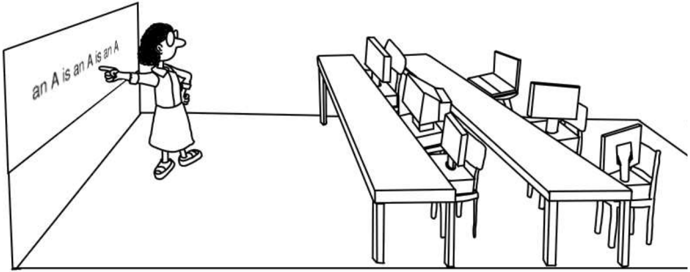
Machine Learning

Krystian Mikolajczyk & Deniz Gunduz

Department of Electrical and Electronic Engineering
Imperial College London



Machine Learning - Part 2.1 Summary

Department of Electrical and Electronic Engineering

Imperial College London

- Feasibility of learning
- Hoeffding's inequality
- Target distribution and error cost

Some simple terms

- A is impossible: $P(A) = 0$
- A is certain: $P(A) = 1$
- A is almost certain $P(A) \approx 1$
- A is probable $P(A) > 0 \ \wedge P(A) < 1$
- A is correct if error $\varepsilon_A = 0$
- A is approximately correct $\varepsilon_A \approx 0$, very small
- A is probably (how certain?) correct $P(\varepsilon_A = 0) \approx 1$
- A is probably approximately correct P.A.C. $P(\varepsilon_A) > B$ with $B \approx 1$

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
$\mathrm{argmin}_{\lambda\in\Lambda}\, g(\lambda)$	argument λ for which $g(\lambda)$ reaches minimum
$\mathrm{argmax}_{\lambda\in\Lambda}\, g(\lambda)$	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
$\mathbb{I}\left(x\right)$	$\mathbb{I}\left(x\right) = 1$ if $x = \mathit{true}$, $\mathbb{I}\left(x\right) = 0$ if $x = \mathit{false}$
$R()$	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

How can we learn?

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

Hoeffding's inequality

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

Domain $\mathcal{D}$ (test/out-of-sample)

Data $\mathbf{x} \in \mathcal{D}$

Target outputs $\mathbf{y} \in \mathcal{Y}$

Target function $f : \mathcal{X} \to \mathcal{Y}$

Experience $\mathcal{X}$ (observations/train/in-sample)

$\mathbf{x} \in \mathcal{X} \subset \mathcal{D}$

$(\mathbf{x}_1, y_1), (\mathbf{x}_n, y_n), \ldots, (\mathbf{x}_n, y_n)$

with generalisation!

Learners

Expertise

Predictors, Models

Hypothesis $g : \mathcal{X} \to \mathcal{Y}$

Learning

Hypothesis class

$\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$

Learning algorithm

Goal $g \approx f$

Error $g \approx f$

$\{l : (g, \mathcal{Y}) \to \mathbb{R}^+\}$

$\hat{l}[g(\mathbf{w}, \mathbf{x}), f(\mathbf{x})] \in \mathbb{R}^+$

$\widehat{R}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{l}[g(\mathbf{w}, \mathbf{x}), f(\mathbf{x})]$

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

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

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

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

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

Hoeffding's inequality

For $\varepsilon > 0$,

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

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

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

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): $p_{event} = R(h)$.
  - $R(h) = P\left(h(\mathbf{x}_i) \neq f(\mathbf{x}_i)\right)$;
  - $R(h) = \mathbb{E}\left[\widehat{R}_n(h)\right]$.

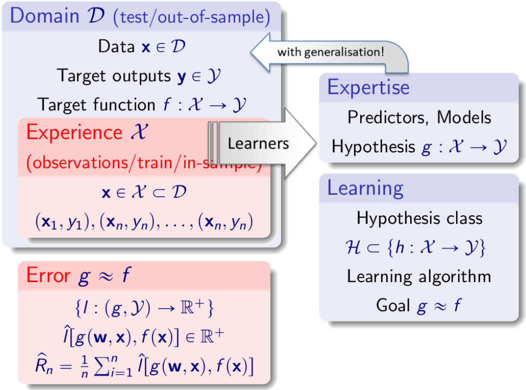
By Hoeffding's inequality

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

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 $\quad \ell(\hat{y}, y) = (\hat{y} - y)^2$

  binary error $\quad \ell(\hat{y}, y) = \mathbb{I}(\hat{y} \neq y)$

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

Target distribution: Error cost

Two types of error:

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

How do we penalize them?

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

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

- Aggressive: false negative is expensive

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

- Risk averse: false positive is expensive

		f	
		+1	-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:

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

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

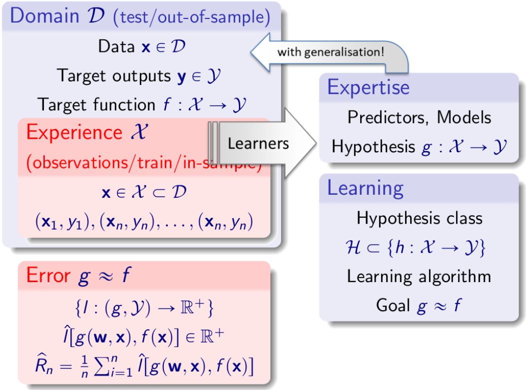
Learning problem

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 (e.g., $g = \operatorname{argmin}_h \widehat{R}_n(h)$),

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

- Can we learn infinite function classes?