Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

ERM over finite hypothesis classes

Outline

- Empirical risk minimisation: Overfitting if we optimise over all h
- Hypothesis class $\mathcal{H}$: Examples and ERM over $\mathcal{H}$
- Assume $\mathcal{H}$ is a finite set
  - We will derive a generalisation error bound for ERM solution $\widehat{h}$

$$L^*_{\mathcal{D}} \quad \leq \quad \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \quad \leq \quad L_{\mathcal{D}}(\widehat{h}) \quad \leq \quad \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{2\ln(|\mathcal{H}|) + c}{m}}$$

$\ln(\cdot) =$ natural logarithm, $c =$ some additional term

Empirical risk minimisation

- Empirical risk (training error) of predictor h:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x), y)$$

  - $L_S(h)$ is estimate of the generalisation error $L_{\mathcal{D}}(h)$ $\qquad \mathbb{E}_S[L_S(h)] = L_{\mathcal{D}}(h)$

- Empirical risk minimisation (ERM): $\underset{h \in \mathcal{Y}^{\mathcal{X}}}{\text{minimise}}\ L_S(h)$

  - Proxy for minimising $L_{\mathcal{D}}(h)$
  - ERM solution $\widehat{h}$ has small $L_S(\widehat{h})$, but may have high $L_{\mathcal{D}}(\widehat{h})$

Understanding the failure of ERM

- Example of ERM solution (learning by memorisation):

$$h_{mem}(x) = \begin{cases} 1 & \text{if } (x,1) \in S \\ 0 & \text{otherwise} \end{cases}$$

- Recall h_{mem} can be poor solution if $|\mathcal{X}| \gg m$

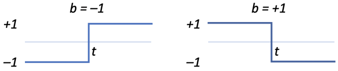
$$L_S(h_{mem}) = 0 \qquad \text{but} \qquad \mathrm{L}_{\mathcal{D}}(h_{mem}) \text{ can be large}$$

- Question: Recall that we stated $\mathbb{E}_S[L_S(h)] = L_{\mathcal{D}}(h)$
  - Why is the same not true for h_{mem}?

Understanding the failure of ERM

- Why is the *memorised* solution poor?
  - h_{mem} overfits the training sample S
  - Search space $\mathcal{Y}^{\mathcal{X}}$ has all functions (including complex ones that overfit)
- Solution: Reduce the search space to some $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$
  - $\mathcal{H}$ is called hypothesis class / concept class
  - Example: Linear classifiers, decision stumps
  - Decision stumps over $\mathcal{X} = \mathbb{R}$

$$\mathcal{H} = \{b \cdot \mathrm{sign}(t - x) \;:\; t \in \mathbb{R}, b \in \{\pm 1\}\} \;\subset\; \{-1, 1\}^{\mathbb{R}}$$



How good can ERM solution be?

- ERM over $\mathcal{H}$: $\underset{h\in\mathcal{H}}{\text{minimise}}\, L_S(h)$
- Recall: Our goal is to find h with small $L_\mathcal{D}(h)$
- Smallest generalisation error we can have from ERM: $L_\mathcal{D}(\mathcal{H}) := \min_{h\in\mathcal{H}} L_\mathcal{D}(h)$
- Smallest possible generalisation error (Bayes risk): $L^*_\mathcal{D} := \min_{h\in\mathcal{Y}^\mathcal{X}} L_\mathcal{D}(h)$
- Let $\widehat{h}$ be the solution of ERM over $\mathcal{H}$

$$L_\mathcal{D}(\widehat{h}) \;\geq\; L_\mathcal{D}(\mathcal{H}) \;\geq\; L^*_\mathcal{D}$$

How good can ERM solution be?

- How much worse is $L_{\mathcal{D}}(\widehat{h})$ compared to $L^*_{\mathcal{D}}$?

$$L_{\mathcal{D}}(\widehat{h}) - L^*_{\mathcal{D}} \;=\; \underbrace{L_{\mathcal{D}}(\mathcal{H}) - L^*_{\mathcal{D}}}_{\substack{\text{approximation error}\\ \text{or, inductive bias}}} \;+\; \underbrace{L_{\mathcal{D}}(\widehat{h}) - L_{\mathcal{D}}(\mathcal{H})}_{\text{estimation error}}$$

- This decomposition is true for any learning paradigm

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} J(h) \qquad\qquad J(h) = L_S(h) \text{ for ERM}$$

Approximation error / Inductive bias

$$L_{\mathcal{D}}(\widehat{h}) - L^*_{\mathcal{D}} = \underbrace{L_{\mathcal{D}}(\mathcal{H}) - L^*_{\mathcal{D}}}_{\substack{\text{approximation error} \\ \text{or, inductive bias}}} + \underbrace{L_{\mathcal{D}}(\widehat{h}) - L_{\mathcal{D}}(\mathcal{H})}_{\text{estimation error}}$$

- Minimum error that occurs because we restrict our search to $\mathcal{H}$
- Inductive bias = Bias induced by making model assumptions ($\mathcal{H}$ is linear classifier)
- Cannot be controlled by learning algorithm
- Bias decreases if $\mathcal{H}$ is increased

$$\mathcal{H} \subset \mathcal{H}' \implies L_{\mathcal{D}}(\mathcal{H}) - L^*_{\mathcal{D}} \geq L_{\mathcal{D}}(\mathcal{H}') - L^*_{\mathcal{D}}$$

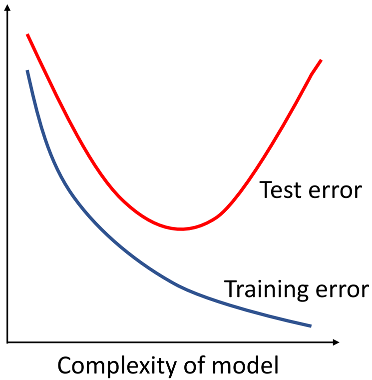
Estimation error

$$L_{\mathcal{D}}(\widehat{h}) - L^*_{\mathcal{D}} = \underbrace{L_{\mathcal{D}}(\mathcal{H}) - L^*_{\mathcal{D}}}_{\substack{\text{approximation error} \\ \text{or, inductive bias}}} + \underbrace{L_{\mathcal{D}}(\widehat{h}) - L_{\mathcal{D}}(\mathcal{H})}_{\text{estimation error}}$$

- Error of $\widehat{h}$ with respect to best possible $h \in \mathcal{H}$
- Could be controlled by learning algorithm
- Estimation error for ERM typically increases if $\mathcal{H}$ is larger

Overfitting and underfitting / Bias-variance trade-off

- What happens if we increase $\mathcal{H}$?
  - Allow $\mathcal{H}$ to have more complex predictors
  - Linear classifier $\rightarrow$ Quadratic decision boundary $\rightarrow \dots$
- Bias reduces $\implies$ training error reduces (overfitting)
  Estimation error increases
- Test error / generalisation error:
  - Large for small $\mathcal{H}$ due to large bias (underfitting)
  - Large for large $\mathcal{H}$ as estimation error high (overfitting)



Generalisation error bound

- $\widehat{h}$ or $\widehat{h}_{ERM} = \arg\min_{h \in \mathcal{H}} L_S(h)$

- How good (or bad) is the generalisation error of $\widehat{h}$?

$$L^*_{\mathcal{D}} \leq L_{\mathcal{D}}(\mathcal{H}) \leq L_{\mathcal{D}}(\widehat{h}) \leq ??$$

- To derive an upper bound, assume $\mathcal{H}$ is finite

- Two settings:
  - Simple: No randomness in label, $y = h_0(x)$ for some $h_0 \in \mathcal{H}$
  - General: Random label, $(x, y) \sim \mathcal{D}$ . . . also includes $y = h_0(x)$ where $h_0 \notin \mathcal{H}$

Bound on generalisation error: Simple case

$y = h_0(x)$ with $h_0 \in \mathcal{H} \quad \Longrightarrow \quad L_{\mathcal{D}}(\mathcal{H}) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$ and $L_S(\widehat{h}) = \min_{h \in \mathcal{H}} L_S(h) = 0$

Theorem ERMH.1 (Generalisation error bound in simple case)

Assume $\mathcal{H}$ is finite, $\ell =$ 0-1 loss and $\widehat{h} =$ ERM solution

For $\epsilon \in (0,1)$,

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(\widehat{h}) > \epsilon\right) \leq |\mathcal{H}| e^{-m\epsilon}$$

Equivalent statement:

For $\delta \in (0,1)$,

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(\widehat{h}) > \frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{m}\right) \leq \delta$$

Equivalent statement:

For $\delta \in (0,1)$,

$$L_{\mathcal{D}}(\widehat{h}) \leq \frac{\ln(|\mathcal{H}|) + \ln(\frac{1}{\delta})}{m} \qquad \text{with probability} \geq 1 - \delta$$

The above result in simple words

- Set $\delta = 0.01 \quad \implies \quad \ln(\frac{1}{\delta}) = 4.6 < 5$
- Consider T independent runs of following thought experiment (T is large)
  - Sample $S \sim \mathcal{D}^m$, and solve ERM to get $\widehat{h}$
  - Compute $L_{\mathcal{D}}(\widehat{h})$ (equivalently, compute test error on infinitely many samples)
- Out of T runs: $\quad L_{\mathcal{D}}(\widehat{h}) \leq \dfrac{\ln(|\mathcal{H}|) + 5}{m} \quad$ holds in $0.99T$ runs
- If we have only one run (typical happens in practice):
  - The bound will typically be true

Equivalence of the three statements

- Verify 3^{rd} statement is rephrasing of 2^{nd}
- Try getting 2^{nd} statement from 1^{st}
- Solution: Set $\delta = |\mathcal{H}|e^{-m\epsilon}$, and write ϵ in terms of δ
- Due to equivalence, we prove only the first statement

Proof of generalisation error bound

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\widehat{h}) > \epsilon\right) \leq |\mathcal{H}|e^{-m\epsilon}$$

Let $\mathcal{H}_{bad} = \{h \in \mathcal{H} \ : \ L_{\mathcal{D}}(h) > \epsilon\}$

$$L_{\mathcal{D}}(\widehat{h}) > \epsilon \quad \implies \quad \widehat{h} \in \mathcal{H}_{bad} \quad \implies \quad \text{there is } h \in \mathcal{H}_{bad} \text{ such that } L_S(h) = 0$$

Try to write above in terms of events and their probabilities

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\widehat{h}) > \epsilon\right) &\leq \mathbb{P}_{S\sim\mathcal{D}^m}(\text{there is } h \in \mathcal{H}_{bad} \text{ such that } L_S(h) = 0) \\ &\leq \mathbb{P}_{S\sim\mathcal{D}^m}\left(\bigcup_{h\in\mathcal{H}_{bad}} \{L_S(h) = 0\}\right)\end{aligned}$$

Proof: Union bound

Recall for events A, B: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

Using above, $$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\bigcup_{h\in\mathcal{H}_{bad}} \{L_S(h) = 0\}\right) \leq \sum_{h\in\mathcal{H}_{bad}} \mathbb{P}_{S\sim\mathcal{D}^m}\big(L_S(h) = 0\big)$$

Proof: Independence of training samples

Recall for independent events A, B: $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) \cdot \mathbb{P}(B)$

Bounding $\mathbb{P}_{S\sim\mathcal{D}^m}\big(L_S(h) = 0\big)$ for $h \in \mathcal{H}_{bad}$

$$L_S(h) = 0 \quad \Longrightarrow \quad \ell(h(x_i), y_i) = 0 \text{ for every } (x_i, y_i) \in S$$

Hence

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\big(L_S(h) = 0\big) &= \mathbb{P}_{S\sim\mathcal{D}^m}\left(\bigcap_{i=1}^{m}\{\ell(h(x_i), y_i) = 0\}\right)\\ &= \prod_{i=1}^{m}\mathbb{P}_{(x_i,y_i)\sim\mathcal{D}}\big(\ell(h(x_i), y_i) = 0\big)\end{aligned}$$

Proof: ℓ is 0-1 loss

$\ell(h(x_i), y_i)$ is Bernoulli with

$$\begin{aligned}\mathbb{P}_{(x_i,y_i)\sim\mathcal{D}}\big(\ell(h(x_i), y_i) = 1\big) &= \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) \\ &= L_{\mathcal{D}}(h) \\ &> \epsilon & \ldots \text{ for } h \in \mathcal{H}_{bad}\end{aligned}$$

Thus, $\mathbb{P}_{(x_i,y_i)\sim\mathcal{D}}\big(\ell(h(x_i), y_i) = 0\big) \leq 1 - \epsilon$

Proof: Final steps

$$
\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\widehat{h}) > \epsilon\right) &\leq \sum_{h\in\mathcal{H}_{bad}} \mathbb{P}_{S\sim\mathcal{D}^m}\big(L_S(h) = 0\big) && \ldots \text{ union bound} \\
&\leq \sum_{h\in\mathcal{H}_{bad}} \prod_{i=1}^{m} \mathbb{P}_{(x_i,y_i)\sim\mathcal{D}}\big(\ell(h(x_i), y_i) = 0\big) && \ldots \text{ independent samples} \\
&\leq \sum_{h\in\mathcal{H}_{bad}} (1-\epsilon)^m \\
&= |\mathcal{H}_{bad}|(1-\epsilon)^m \\
&\leq |\mathcal{H}|(1-\epsilon)^m && \ldots \text{ since } \mathcal{H}_{bad} \subset \mathcal{H} \\
&\leq |\mathcal{H}|e^{-m\epsilon} && \ldots 1-\epsilon \leq e^{-\epsilon} \text{ for all } \epsilon
\end{aligned}
$$

General case: True labels can be random

We use the following result to derive a generalisation error bound

Theorem ERMH.2 (Uniform convergence of $L_S(\cdot)$ for finite $\mathcal{H}$)

Let $\epsilon \in (0,1)$, $\mathcal{H} \subset \{-1,+1\}^{\mathcal{X}}$ *and we measure risk with respect to 0-1 loss.*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\max_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right) \leq 2|\mathcal{H}|e^{-2m\epsilon^2}$$

Equivalent statement: *Let* $\delta \in (0,1)$. *With probability* $\geq 1-\delta$,

$$\max_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}}$$

Generalisation error from uniform convergence

$\widehat{h}$ = solution of ERM

With probability at least $1-\delta$, the following hold simultaneously:

$$L_{\mathcal{D}}(\widehat{h}) - L_S(\widehat{h}) \leq \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}}$$

$$L_S(h) - L_{\mathcal{D}}(h) \leq \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}} \qquad \text{for every } h \in \mathcal{H}$$

Using above, we can show following generalisation error bound

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + \sqrt{\frac{2\ln(|\mathcal{H}|) + 2\ln(\frac{2}{\delta})}{m}} \qquad \text{with probability } 1-\delta$$

Generalisation error from uniform convergence

With probability $1-\delta$,

$$
\begin{aligned}
L_{\mathcal{D}}(\widehat{h}) &\leq L_S(\widehat{h}) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}} \\
&= \min_{h\in\mathcal{H}} L_S(h) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}} && \ldots \widehat{h} \text{ minimises } L_S(h) \\
&\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}} && \ldots 2^{nd} \text{ statement of previous slide} \\
&= L_{\mathcal{D}}(\mathcal{H}) + \sqrt{\frac{2\ln(|\mathcal{H}|) + 2\ln(\frac{2}{\delta})}{m}}
\end{aligned}
$$

Uniform convergence of empirical risk

Theorem ERMH.3 (Uniform convergence of $L_S(\cdot)$ for finite $\mathcal{H}$)

Let $\epsilon \in (0,1)$, $\mathcal{H} \subset \{-1,+1\}^{\mathcal{X}}$ and we measure risk with respect to 0-1 loss.

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\max_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right) \leq 2|\mathcal{H}|e^{-2m\epsilon^2}$$

- We prove above statement
- Earlier slide mentioned an equivalent statement
    - Verify that they are equivalent

Two useful probability inequalities

Theorem ERMH.4 (Tail bound for maximum (consequence of union bound))

Let $Z_1, \ldots, Z_n$ *be* n *random variables.*

$$\mathbb{P}\left(\max_{1\le i\le n} Z_i > \epsilon\right) \le \sum_{i=1}^{n} \mathbb{P}(Z_i > \epsilon)$$

Theorem ERMH.5 (Hoeffding's inequality)

Let $Z_1, \ldots, Z_n$ *be* n *independent random variables such that* $\mathbb{P}(Z_i \in [a_i, b_i]) = 1$ *for all* i*.*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i])\right| > \epsilon\right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Try to prove uniform convergence using above

Proof of uniform convergence

$$
\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^m}\left(\max_{h\in\mathcal{H}}|L_S(h)-L_{\mathcal{D}}(h)|>\epsilon\right)\\
&\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}\left(|L_S(h)-L_{\mathcal{D}}(h)|>\epsilon\right) && \ldots \text{ union bound}\\
&\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}\left(\frac{1}{m}\left|\sum_{i=1}^m \Big(\ell(h(x_i),y_i)-L_{\mathcal{D}}(h)\Big)\right|>\epsilon\right) && \ldots \text{ definition of } L_S(h)\\
&\leq \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}\left(\frac{1}{m}\left|\sum_{i=1}^m \Big(\ell(h(x_i),y_i)-\mathbb{E}[\ell(h(x_i),y_i)]\Big)\right|>\epsilon\right) && \ldots \text{ definition of } L_{\mathcal{D}}(h)\\
&= \sum_{h\in\mathcal{H}} \mathbb{P}_{S\sim\mathcal{D}^m}\left(\left|\sum_{i=1}^m \Big(Z_i-\mathbb{E}[Z_i]\Big)\right|>m\epsilon\right) && \ldots \text{ define } Z_i=\ell(h(x_i),y_i)
\end{aligned}
$$

Use Hoeffding's inequality noting that $Z_1,\ldots,Z_m$ independent with $Z_i\in[0,1]$