Lecture 6. Generalisation with Countably Finite Model Class

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio

THE UNIVERSITY OF MELBOURNE

This lecture

- Motivation
- Finite model class
- Empirical risk and true risk
- Generalisation
  - Bounding test risk with high probability
- A proof-sketch for generalisation
  - Union bound
  - Hoeffding's inequality

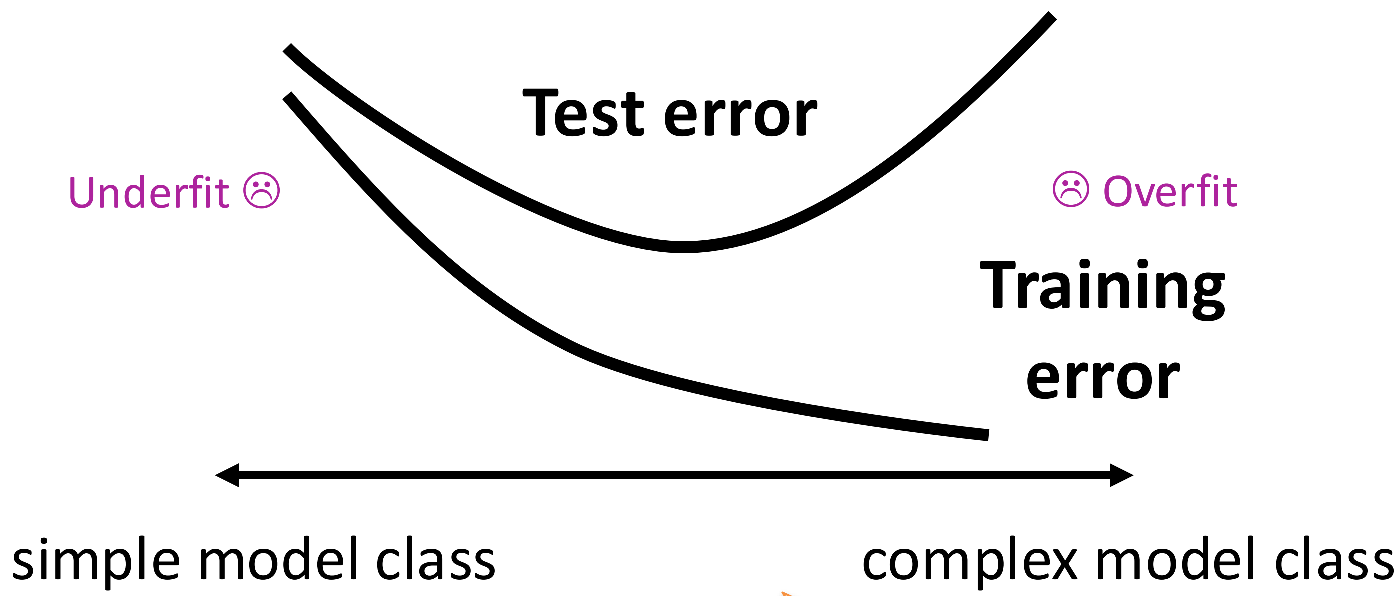
Motivation

…from previous lectures

Classification problems

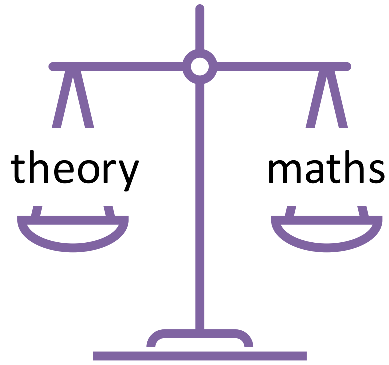
- There are two parts to any classification task
- Estimation: how to select the best classifier out of a particular set
  - E.g., logistic regression learns the “best” classifier from the set of linear classifiers
- Model selection: how to select the best set of classifiers
  - E.g., linear classifiers, quadratic classifiers, cubic classifiers
  - E.g., single-feature classifiers, multi-feature classifiers
- Both of these selections have to be made based on training data

Test error and training error



But how to measure model class complexity?

Generalisation and Model Complexity



- Theory we've seen so far (mostly statistics)
  - Asymptotic notions (consistency, efficiency)
  - Convergence could be really slow
  - Model complexity undefined
- Want: finite-sample theory
- Want: define model complexity and relate it to test error
  - Test error can't be measured in real life, but it can be provably bounded!
- Want: distribution-independent, learner-independent theory
  - A fundamental theory applicable *throughout ML*
  - Unlike bias-variance: distribution dependent, no model complexity,

Finite Model Class

Not the most general model class, but allows to get some initial intuition.

Countably Finite

- Countable set: we can count the elements of the set as we do with natural numbers
  * Countably finite set S with a finite number of elements $|S|$
    - $S = \{19,37,102\}$, $|S| = 3$
    - $S =$ the set of all odd natural numbers less than 10, $|S| = 5$
  * Countably infinite set with infinite number of elements
    - natural numbers $\mathbb{N}$, integers $\mathbb{Z}$, rational numbers $\mathbb{Q}$
    - the set of all odd natural numbers
- Uncountable set: we cannot count the elements of the set as we do with natural numbers
    - $\mathbb{R}$
    - $[0,1]$

A Finite Model Class: Single-feature

- Consider we have 2 features and a countably finite set $\mathcal{F}$ of classifiers, containing:

$$f(x) = \mathrm{sgn}(x_1) = \begin{cases} +1, \text{if } x_1 > 0 \\ -1, \text{if } x_1 \leq 0 \end{cases}$$
$$f(x) = \mathrm{sgn}(-x_1)$$
$$f(x) = \mathrm{sgn}(x_2)$$
$$f(x) = \mathrm{sgn}(-x_2)$$

- Here $|\mathcal{F}| = 4$

Another Finite Model Class: Multi-feature

- Consider we have 2 features and a countably finite set $\mathcal{F}$ of classifiers, containing:

$$f(x) = \text{sgn}(x_1 + x_2) = \begin{cases} +1, \text{if } x_1 + x_2 > 0 \\ -1, \text{if } x_1 + x_2 \leq 0 \end{cases}$$

$$f(x) = \text{sgn}(x_1 - x_2)$$
$$f(x) = \text{sgn}(-x_1 + x_2)$$
$$f(x) = \text{sgn}(-x_1 - x_2)$$
$$f(x) = \text{sgn}(x_1)$$
$$f(x) = \text{sgn}(-x_1)$$
$$f(x) = \text{sgn}(x_2)$$
$$f(x) = \text{sgn}(-x_2)$$

- Here $|\mathcal{F}| = 8$

Empirical Risk and True Risk

Which one do we really care about?

Empirical Risk

- Training data $\boldsymbol{D} = \{\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_n, y_n\}$
- The empirical risk of a classifier f for loss l is

$$\hat{R}_{\boldsymbol{D}}[f] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i, f(\boldsymbol{x}_i)\big)$$

aka training error for

$$l(y, y') = \begin{cases} 1, \text{if } y \neq y' \\ 0, \text{if } y = y' \end{cases}$$

- Given classifier f and n samples in $\boldsymbol{D}$, we can compute $\hat{R}_{\boldsymbol{D}}[f]$

Empirical Risk Minimisation

- Training data $\boldsymbol{D} = \{\boldsymbol{x}_1, y_1, \dots, \boldsymbol{x}_n, y_n\}$ is a random variable!
  - There is an unknown data distribution P
  - $(\boldsymbol{x}_i, y_i)$ i.i.d. with distribution P

Independent and identically distributed

- The empirical risk of a classifier f for loss l is

$$\hat{R}_{\boldsymbol{D}}[f] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i, f(\boldsymbol{x}_i)\big)$$

- ERM: $\hat{f}_{\boldsymbol{D}}$ minimises the empirical risk

$$\hat{f}_{\boldsymbol{D}} = \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_{\boldsymbol{D}}[f]$$

Go trough all the $|\mathcal{F}| = 8$ classifiers and choose the best for data D

True Risk

- The true risk is the expected value of the loss l
  - The empirical risk is an estimate (an average of a finite number of samples n) of the expected value
  - Intuitively speaking, the true risk is the empirical risk when using an infinite number of samples
- The true risk of a classifier f for loss l is

$$R[f] = \mathbb{E}\, l\big(Y, f(X)\big) = \int l\big(Y, f(X)\big) P(X,Y)\, dX\, dY$$

aka generalisation error (expected test error) for

$$l(y, y') = \begin{cases} 1, \text{if } y \neq y' \\ 0, \text{if } y = y' \end{cases}$$

- Given classifier f, we cannot compute $R[f]$ because the data distribution P is unknown

Empirical Risk and True Risk

- While we can only compute the empirical risk $\hat{R}_{\boldsymbol{D}}[f]$, we are truly interested on the true risk $R[f]$, because the true risk is the measure of how we will perform on unseen data
- Under-fitting: large empirical risk $\hat{R}_{\boldsymbol{D}}[f]$ and true risk $R[f]$
- Over-fitting: small empirical risk $\hat{R}_{\boldsymbol{D}}[f]$, large true risk $R[f]$

Generalisation

What can we say about something we cannot even compute (true risk)?

Generalisation Theorem

- For a finite model class $\mathcal{F}$, without knowing the data distribution P, with probability $\geq 1-\delta$ over the choice of the training set D of n i.i.d. samples

$$R[\hat{f}_D] \leq \hat{R}_D[\hat{f}_D] + \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}$$

We cannot compute $R[f]$, but we can bound it!

  * E.g., $|\mathcal{F}| = 8$, $\delta = 0.1$, with probability $\geq 1-\delta = 0.9$

$$R[\hat{f}_D] \leq \hat{R}_D[\hat{f}_D] + \sqrt{\frac{\log 8 + \log 10}{2n}}$$

Structural Risk Minimisation

- Choose the model class (e.g., single-feature classifiers, multi-feature classifiers) with best guarantee of generalisation:

$$\underbrace{\hat{R}_D[\hat{f}_D]} + \underbrace{\sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}}$$

Large for simple classifiers,
small for complex classifiers

Small for simple classifiers (small $|\mathcal{F}|$),
large for complex classifiers (large $|\mathcal{F}|$)

Large for small n (few samples),
small for large n (many samples)

Mini Summary

- Caveat: Bound is "with high probability" since we could be unlucky with the data
- Worst-case on distributions P: We don't want to assume something unrealistic about where the data comes from
- Some initial measure of model complexity $|\mathcal{F}|$
- Structural risk minimisation (trade-off)

A Proof-sketch for Generalisation

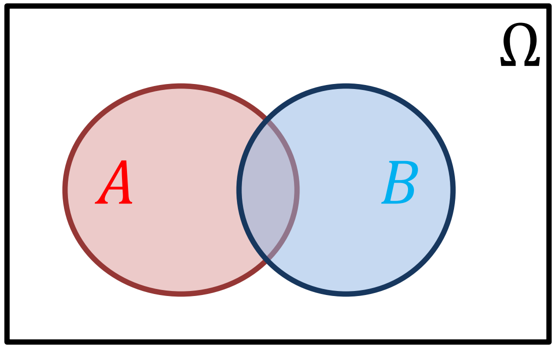
What can we say about something we cannot even compute (true risk)?

Tool 1: Union bound

- For events/conditions A, B

$$\Pr[A \text{ or } B] \leq \Pr[A] + \Pr[B]$$

- Proof-sketch:

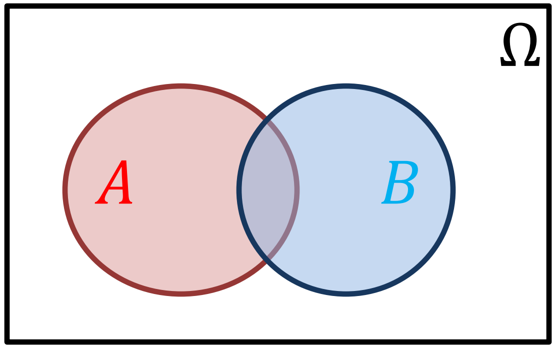


Tool 1: Union bound

- For events/conditions A, B

$$\Pr[A \text{ or } B] \le \Pr[A] + \Pr[B]$$

- Proof-sketch:



- For events/conditions $A_1, A_2 \ldots A_k$

$$\Pr[A_1 \text{ or} \ldots \text{or } A_k] \le \Pr[A_1] + \ldots + \Pr[A_k]$$

- Proof: $\Pr[A_1 \text{ or} \ldots \text{or } A_k] = \Pr[A_1 \text{ or } (A_2 \text{ or} \ldots \text{or } A_k)]$

Tool 2: Hoeffding's inequality

- The probability that the empirical average is far from the expectation is small.
- Many such concentration inequalities.
- Let $Z_1, \ldots, Z_n, Z$ be i.i.d. random variables with domain $[0,1]$. For a constant $\varepsilon > 0$

$$\Pr\left[\mathbb{E}Z - \frac{1}{n}\sum_{i=1}^{n} Z_i > \varepsilon\right] \leq e^{-2n\varepsilon^2}$$

(Proof outside scope of COMP90051)

Generalisation Theorem Proof-sketch

- Find ε such that with probability $\geq 1-\delta$

$$R[\hat{f}_{\boldsymbol{D}}] \leq \hat{R}_{\boldsymbol{D}}[\hat{f}_{\boldsymbol{D}}] + \varepsilon$$

$$\varepsilon = \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}$$

- Equivalent to

$$R[\hat{f}_{\boldsymbol{D}}] - \hat{R}_{\boldsymbol{D}}[\hat{f}_{\boldsymbol{D}}] \leq \varepsilon$$

Generalisation Theorem Proof-sketch

- Find ε such that with probability $\geq 1-\delta$
$$R\left[\hat{f}_{\boldsymbol{D}}\right] \leq \hat{R}_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right] + \varepsilon$$

- Equivalent to
$$\varphi_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right] = R\left[\hat{f}_{\boldsymbol{D}}\right] - \hat{R}_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right] \leq \varepsilon$$

- Using a worst case bound (the maximum over all f)
$$\varphi_{\boldsymbol{D}}\left[\hat{f}_{\boldsymbol{D}}\right] \leq \max_{f\in\mathcal{F}} \varphi_{\boldsymbol{D}}[f]$$

Generalisation Theorem Proof-sketch

- Find ε such that with probability $\geq 1-\delta$

$$R[\hat{f}_D] \leq \hat{R}_D[\hat{f}_D] + \varepsilon$$

- Equivalent to

$$\varphi_D[\hat{f}_D] = R[\hat{f}_D] - \hat{R}_D[\hat{f}_D] \leq \varepsilon$$

- Using a worst case bound (the maximum over all f)

$$\varphi_D[\hat{f}_D] \leq \max_{f\in\mathcal{F}} \varphi_D[f]$$

- If $\varphi_D[f] \leq \varepsilon$ for all classifiers $f \in \mathcal{F}$, then

$$\varphi_D[\hat{f}_D] \leq \max_{f\in\mathcal{F}} \varphi_D[f] \leq \varepsilon$$

Generalisation Theorem Proof-sketch

- Now, find ε such that with probability $\geq 1 - \delta$

$$\varphi_{\boldsymbol{D}}[f] \leq \varepsilon \text{ for all classifiers } f \in \mathcal{F}$$

- In other words

$$\Pr[\varphi_{\boldsymbol{D}}[f] \leq \varepsilon \text{ for all } f \in \mathcal{F}] \geq 1 - \delta$$

Generalisation Theorem Proof-sketch

- Now, find ε such that with probability $\geq 1-\delta$

$$\varphi_{\boldsymbol{D}}[f] \leq \varepsilon \text{ for all classifiers } f \in \mathcal{F}$$

- In other words

$$\Pr[\varphi_{\boldsymbol{D}}[f] \leq \varepsilon \text{ for all } f \in \mathcal{F}] \geq 1-\delta$$

- For any event/condition A, $\Pr[A] = 1 - \Pr[\text{not } A]$, thus

$$\Pr[\varphi_{\boldsymbol{D}}[f] > \varepsilon \text{ for some } f \in \mathcal{F}] \leq \delta$$

- This is called a *uniform deviation bound*

Generalisation Theorem Proof-sketch

- Now, find ε such that

$$\Pr[\varphi_D[f] > \varepsilon \text{ for some } f \in \mathcal{F}] \leq \delta$$

- By union bound

$$\Pr\left[R[f] - \hat{R}_D[f] > \varepsilon \text{ for some } f \in \mathcal{F}\right]$$

$$\leq \sum_{f \in \mathcal{F}} \Pr\left[R[f] - \hat{R}_D[f] > \varepsilon\right]$$

event/condition
A_f defined as
$R[f] - \hat{R}_D[f] > \varepsilon$

Generalisation Theorem Proof-sketch

- Now, find ε such that

$$\Pr[\varphi_{D}[f] > \varepsilon \text{ for some } f \in \mathcal{F}] \leq \delta$$

- By union bound

$$\Pr\left[R[f] - \hat{R}_{D}[f] > \varepsilon \text{ for some } f \in \mathcal{F}\right]$$

$$\leq \sum_{f \in \mathcal{F}} \Pr\left[R[f] - \hat{R}_{D}[f] > \varepsilon\right]$$

- By Hoeffding's inequality

$$\Pr\left[R[f] - \hat{R}_{D}[f] > \varepsilon\right] \leq e^{-2n\varepsilon^2}$$

$$R[f] = \mathbb{E}Z$$
$$Z = l\left(Y, f(X)\right)$$
$$\hat{R}_{D}[f] = \frac{1}{n}\sum_{i=1}^{n} Z_i$$
$$Z_i = l\left(y_i, f(\boldsymbol{x}_i)\right)$$

Generalisation Theorem Proof-sketch

- Putting the last two equations together

$\Pr\left[R[f] - \hat{R}_{\boldsymbol{D}}[f] > \varepsilon \text{ for some } f \in \mathcal{F}\right] \leq |\mathcal{F}|e^{-2n\varepsilon^2}$

- Set $\delta = |\mathcal{F}|e^{-2n\varepsilon^2}$

$$\varepsilon = \sqrt{\frac{\log|\mathcal{F}| + \log(1/\delta)}{2n}}$$

- and we proved our claim.

Discussion

- Hoeffding's inequality here only for bounded loss in $[0,1]$
  * Fancier concentration inequalities leverage variance or do not assume boundedness
- Uniform deviation is worst-case deviation between true risk and empirical risk, across the model class
  * Advantages: works for any learner, data distributionn
  * ERM on a very large over-parametrised $\mathcal{F}$ may approach the worst-case, but learners generally may not (custom analysis, data-dependent bounds, PAC-Bayes, etc.)
- Not i.i.d. data (Martingale theory, coloring numbers, etc.)

Another Model Class

- Finite model class
  - Bounding uniform deviation with union bound and Hoeffding's inequality
- Consider we have 2 features and an uncountable set $\mathcal{F}$ of classifiers, containing for all $w_1 \in \mathbb{R}$, $w_2 \in \mathbb{R}$:

$$f(x) = \mathrm{sgn}(w_1 x_1 + w_2 x_2)$$

Next time: Countably infinite model classes and uncountable model classes!