MATH466/MATH766
Math of machine learning
01/08 Lecture 1 introduction

References:

- Ch5 of Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville
- Ch2 of The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman

Todays contents:

- overview
- policies
- basic concepts in machine learning
- supervised learning overview

Important concepts:

- generalization gap
- overfitting and underfitting
- bias-variance decomposition

Recommend reading:

course policies: refer to syllabus for details

1. communication: canvas, email, in-person. questions, feedbacks and complains are valuable.
2. academic integrity, collaboration and AI tools.
3. grade: course evaluation (2%) + HW (38%) + 2 exams (30%) + course project (30%).
   midterm: will provide practice questions before the exam
   final: will provide practice questions before the exam or change the exam to a take-home open-book exam
   HW: write solutions independently; feedback is very welcome
4. individual course project:
   can be any topic related to machine learning
   several checkpoints are setted
   the learning process is the goal of the course project
   refer to the specific file for details
5. late policy:
   late day
   penalty
   where to ask for help
6. notes to graduate students:
   an additional assignment
   grade and withdraw

course overview:

brainstorm

This course will emphasize understanding machine learning methods through the mathematical derivation, computation and proof. Coding serves as a tool to help understanding.

Some basic concepts in machine learning

- The main components of machine learning are data/task, model/method and algorithm
  data/task: what do you want the machine to learn
  model/method: how do you formulate the learning procedure
  algorithm: how do you solve the learning model

- Based on the data/tasks provided, machine learning are roughly divided into 3 categories: supervised learning, unsupervised learning and reinforcement learning
  supervised learning: labeled data/learn the label. E.g. regression, classification
  unsupervised learning: unlabeled data. E.g. clustering, dimension reduction, manifold learning, data generating
  reinforcement learning: no data beforehand, environment provided, data are generated and learning happens while interacting with the environment.

- There are many other active research area in machine learning whose category is not very clear. We follow this categorization for teaching purpose, for example clearer structure, easier comparison and etc. Please feel free to explore more after class if you are interested and you are always welcome to email me/come to office hour when you have questions so that we can learn together

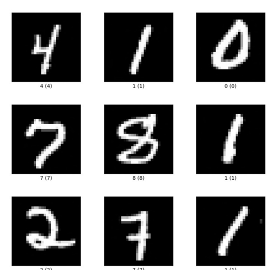
Supervised learning overview

1. Task

labeled data $\{(x_i, y_i)\}_{i=1}^n$,

$x_i \in \mathbb{R}^d$, inputs ;

$y_i \in \mathbb{R}$, labels / responses



MNIST
a dataset of handwritten digits

$x_i \in \mathbb{R}^{784}, y_i \in \{0,1,\cdots,9\}$

Goal: predict y given x

e.g. by $f(x)$

- classification. disct labels
  $y_i \in \{0,1,2,\cdots,9\}$.
- regression. cts labels
  $y_i \in \mathbb{R}$

Variables Table

Variable Name	Role	Type	Demographic	Description	Units
No	ID	Integer			
X1 transaction date	Feature	Continuous		for example, 2013.250=2013 March, 2013.500=2013 June, etc.	
X2 house age	Feature	Continuous			year
X3 distance to the nearest MRT station	Feature	Continuous			meter
X4 number of convenience stores	Feature	Integer		number of convenience stores in the living circle on foot	integer
X5 latitude	Feature	Continuous		geographic coordinate, latitude	degree
X6 longitude	Feature	Continuous		geographic coordinate, longitude	degree
Y house price of unit area	Target	Continuous		10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared	10000 New Taiwan Dollar/Ping

UCI Real estate valuation
a dataset used to study house unit area price in a district in China

$x_i \in \mathbb{R}^6, y_i \in \mathbb{R}$

Q: How to measure the performance of a learned model f ?

2. Measurement : {training, testing, population} error

generalization gap

- loss function: $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$

e.g. $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, $\ell(\hat{y}, y) = \begin{cases} 1, & \hat{y} \neq y \\ 0, & \hat{y} = y \end{cases}$

$\ell(\hat{y}, y) = |\hat{y} - y|$

- error / empirical risk: if the learned model is $y = f(x)$

$$R(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(\underbrace{f(x_i)}_{\text{prediction}}, \underbrace{y_i}_{\text{data}})$$

- training and testing:

assume training data and testing data

(x, y) follows the same distribution $P(x, y)$

use training set to find the model f

use testing set to measure if the model f is good.

training set

$\{(x_i^{tr}, y_i^{tr})\}_{i=1}^{n_{tr}}$

testing set

$\{(x_i^{te}, y_i^{te})\}_{i=1}^{n_{te}}$

model training never across this line →

training error $R_{tr}(f) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \ell(f(x_i^{tr}), y_i^{tr})$

testing error $R_{te}(f) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \ell(f(x_i^{te}), y_i^{te})$

We want both error to be small

If testing data are i.i.d sampling from $P(x,y)$,

then as $n_{te} \to +\infty$, $R_{te}(f)$ is an approximation of

population error $\quad R_p(f) = \int \ell(f(x), y)\, dP(x,y)$

- testing / population error are also called generalization error.
- generalization gap

$$R_p(f) - R_{tr}(f) \quad \text{or} \quad R_{te}(f) - R_{tr}(f)$$

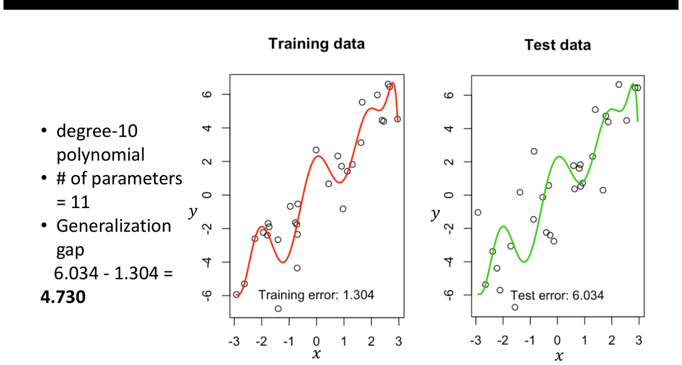
smaller generalization gap is better

3. common issues and the reason

3.1. overfitting and underfitting

overfitting $\qquad\qquad$ underfitting

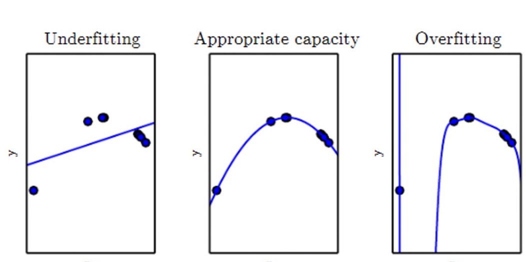
[redacted]



cr. Prof. Xiuyuan Cheng

cr. Prof. Xiuyuan Cheng

[redacted]



cr. Ch. 5 of Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville

[redacted]

underfitting : poor performance on both training and testing data.

← the model is too simple to capture data complexities

overfitting : large generalization gap

← the model overlearns the data

(noise, inaccurate data)

3.2. Model capacity and U-shape curve

Training : find a model f to fit the training data.

usually within a set of models $\mathcal{F}$, i.e. $f \in \mathcal{F}$

Model capacity : the ability of functions in $\mathcal{F}$ to fit data

e.g. the set of constant functions has a low capacity

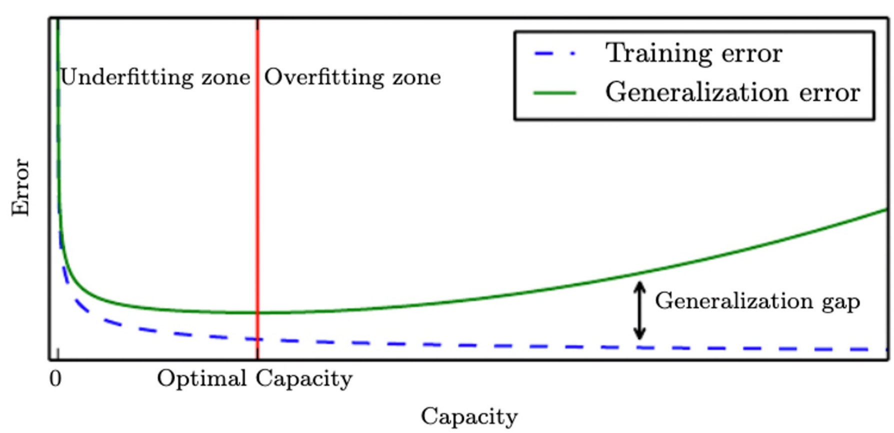
$$\mathcal{F} = \{ f : f(x) \equiv a \mid a \in \mathbb{R} \}$$

e.g. the set of all well-defined functions has a high capacity

higher model capacity $\xrightarrow{\text{usually}}$ lower training error

BUT higher model capacity is not always better

U-shape curve



cr. Ch. 5 of Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville

★ 3.3 Bias - Variance Decomposition

learned model f ← training data

random

⇒ f is also random

assume training process is deterministic

the randomness of f all comes from the randomness of training data

⇒ at a new input x with true response y

the squared error of prediction $(f(x)-y)^2$ is random

the randomness all comes from the randomness of training data

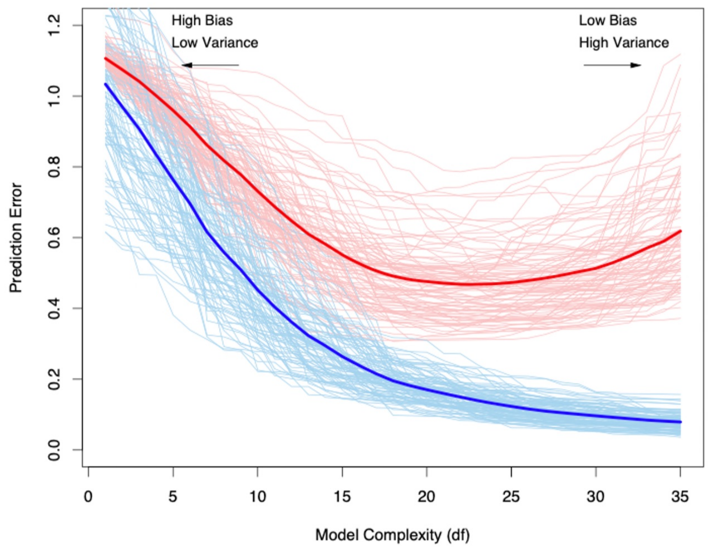
Consider the expectation of the squared error

$$\mathbb{E}_{tr}[(f(x)-y)^2]$$

$$= \underbrace{\mathbb{E}_{tr}[(f(x)-\mathbb{E}_{tr}f(x))^2]}_{\text{Variance}} + \underbrace{(\mathbb{E}_{tr}f(x)-y)^2}_{\text{Bias}^2}$$

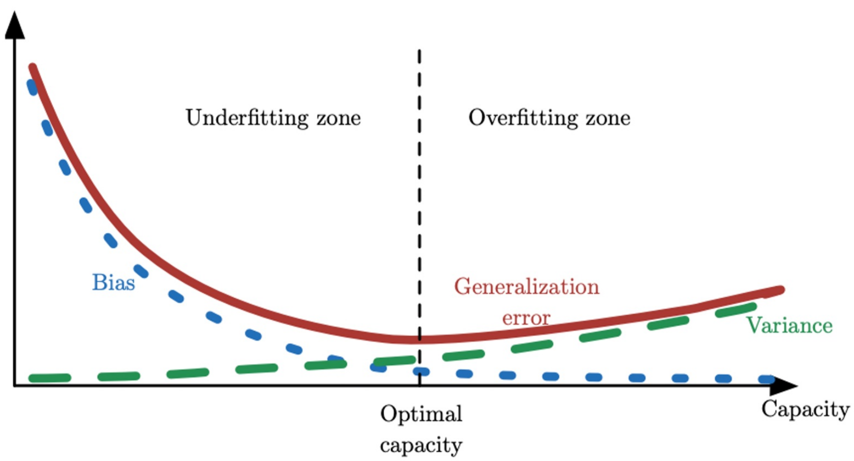
low capacity → high bias, low variance → underfitting

high capacity → high variance, low bias → overfitting



Blue: training
Red: testing

cr. Ch. 7.2 of The Elements of Statistical Learning. 2(1), 2009, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman



cr. Ch. 5 of Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville

Q: How to find the optimal capacity?

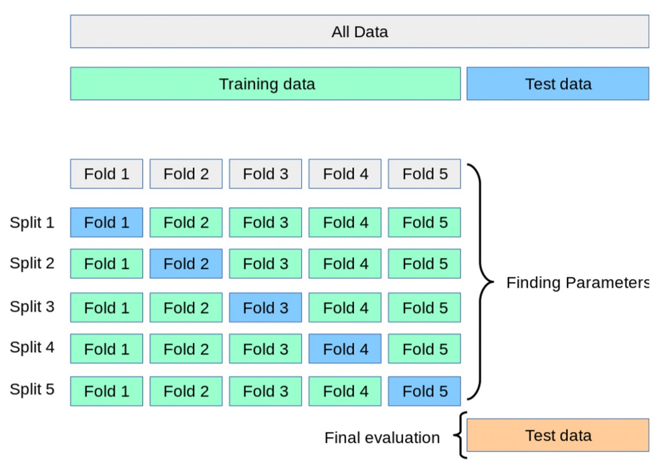
4. Solve the issue: model selection and cross-validation

- Validation

Given training set, "sacrifice" a part of samples as a trail to test the model and the part of data are called validation set.

The trail testing error computed on the validation set can be used to indicate testing error.

- Cross Validation (5-fold)



cr. sklearn 3.1. cross-validation