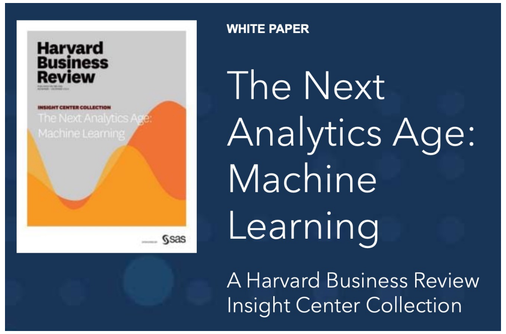
Machine Learning

Prof. Adil Khan

Who Am I?

- Adil Khan
- Professor in Innopolis University (IU), Russia
- Expertise: Machine Learning, Deep Learning
- Head of Machine Learning Lab at IU
- Director of the Institute of AI at IU



The top most emreging trend in Computing and Information Technology

Machine Learning

Large Scale Academic Research

One of the largest number of start-ups

Has changed business in almost every industry

Active Recruitment of ML Engineers

To Effectively use Machine Learning

- One must understand:

  - What is machine learning?
  - The available machine learning methods at your disposal
  - Characterestics of those methods
  - Circumstances under which a method would be most effective
  - Their theoratical underpinnings

Course Objectives

1. Teach you how to implement important machine learning methods and apply them for solving real-world problems
2. Provide you with both the theoratical and practical knowledge of machine learning methods
3. Supervised Methods (Linear, Polynomial and Logistic Regression, Decision Trees, ANNs, CNNs, etc.)
4. Unsupervised Methods (k-means, kmeans++, Hierarchical Clusternig, etc.)
5. Working with ML Models (Regularization, Dimensionality Reduction, Ensemble Learning)

Prerequisites

- Basics of
    - Python programming
    - Linear algebra
    - Calculus
    - Probability

Plan for Today

- Foundations of Machine Learning
  - What does it mean for machines to learn?
    - Task, Experience and Performance
  - Regression
  - Classification
  - Train and Test Error
  - Overfitting and Underfitting
  - Bias-Variance Tradeoff

What Does it Mean for a Machine to Learn?

- If you downloaded a copy of Wikipedia, has your computer really learned something?

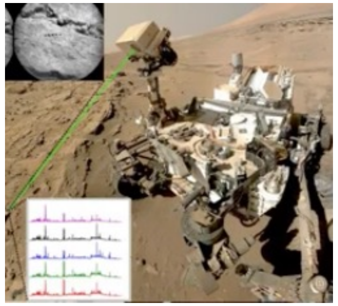
- Did it make your computer any smarter?

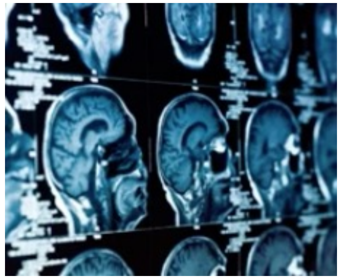
What is Machine Learning?

Machine Learning

- Computer programs that *improve* their *performance* at some *task* through *experience*

Usual Examples of Tasks where ML is being Used





Experience

Targets: For example, malware or Not

Data

$$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$$

Number of available examples

Predictors: For example, execution behavior of a program

$$\boldsymbol{x} \in \mathbb{R}^d$$

Examples of Predictors and Response

Input $(\boldsymbol{x})$	**Output $(\boldsymbol{y})$**	**Application**
Home Features	Price	Real Estate
Ad, User info	Click ad? (0/1)	Online Advertising
Image	Object (1, …, 1000)	Photo Tagging
Audio	Text Transcript	Speech Recognition
English	Chinese	Machine Translation
Image, Radar Info	Position of other cars	Autonomous Driving

Performance

- Needs to be defined according to the given task
- For example: the ratio of correctly identified malwares

Goal of Learning

- Learning or inferring a "functional" relationship between predictors and target

$$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$$

$$\boldsymbol{x} \in \mathbb{R}^d$$

$$y = f(x)$$

$\hat{f} \approx f$ *Goal of learning*

Putting It All Together (1)

$$D = \{(\boldsymbol{x}_i, y_i) \in X \times Y : 1 \leq i \leq N\}$$

Learning Algorithm

$$\hat{f}: X \rightarrow Y$$

Putting It All Together (2)

$$D = \{(\boldsymbol{x}_i, y_i) \in X \times Y : 1 \le i \le N\}$$

"Supervised" Learning Algorithm

$$\hat{f}: X \to Y$$

Example: Dataset 1

Dependent Variable
(Response)

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Independent Variables
(Predictors)

Exampe: Dataset 2

Dependent Variable (Response)

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445

Independent Variables (Predictors)

Compare the Response Variable

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445

$y = \{c_1, c_2, \cdots, c_k\}$

$y \in \mathbb{R}$

Classification and Regression

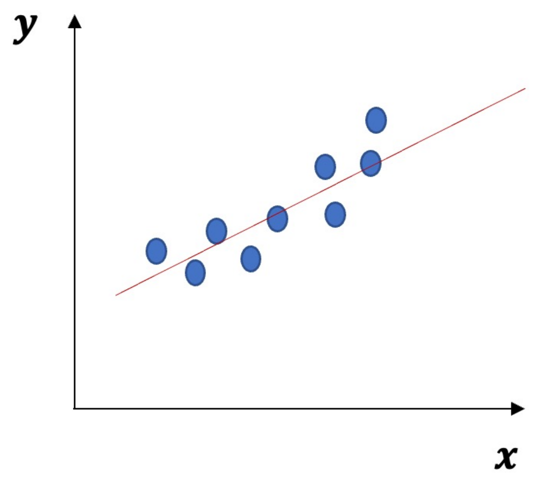
Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Classification

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445

Regression

Estimating $\boldsymbol{f}$



$y = w_o + w_1 x$

f

$$y = f(\boldsymbol{x}; parameters)$$

$$y = f(\boldsymbol{x}; \boldsymbol{w})$$

$$y = f(x; w_0, w_1) = w_0 + w_1 x$$

Assessing the Quality of Learning

- Let $T_r = \{x_i, y_i\}_{i=1}^N$ be the training data we used to estimate $\hat{f}(x)$.

- To assess the quality of estimate, we can compute

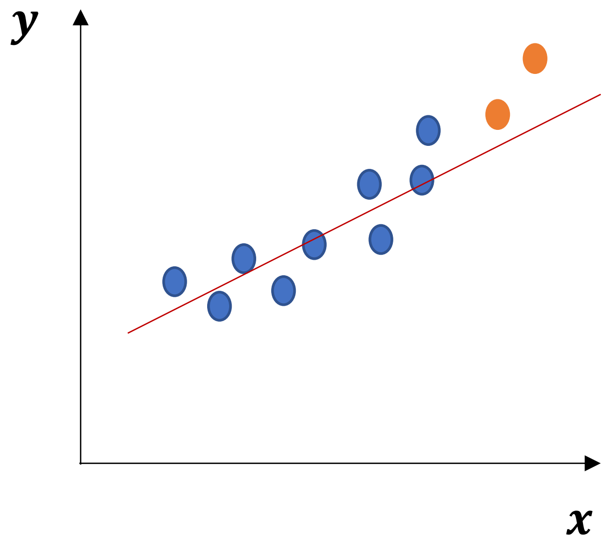
$$\mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i \in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$$

- But this is not a reliable approach

What can go wrong?

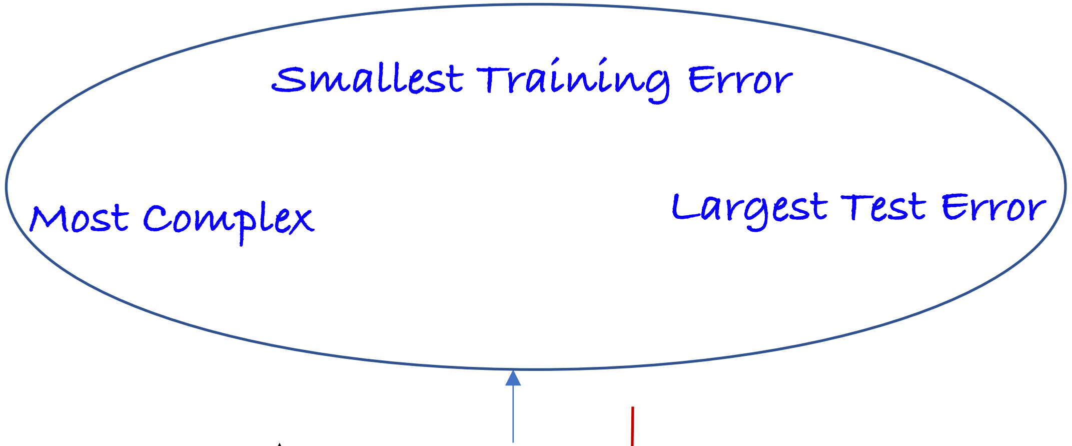
Training Data

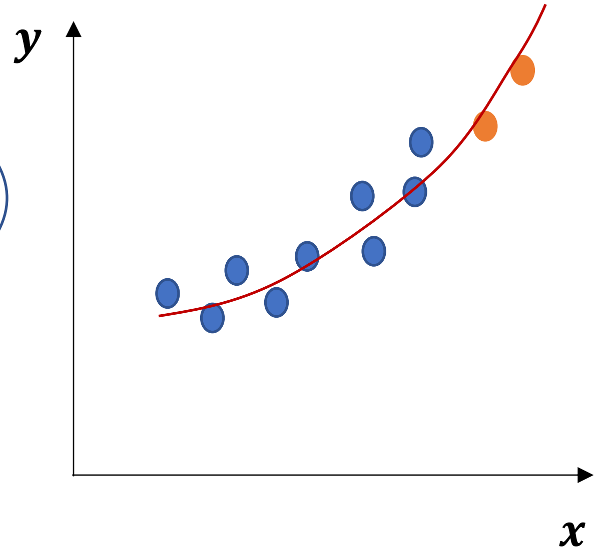
Unseen Test Data



$$y = \underbrace{w_o + w_1 x}_{f}$$

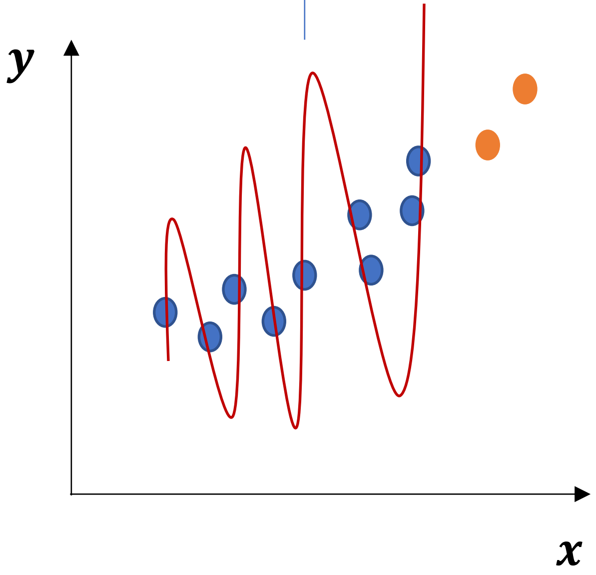
(a)





$$y = \underbrace{w_o + w_1 x + w_2 x^2}_{f}$$

(b)



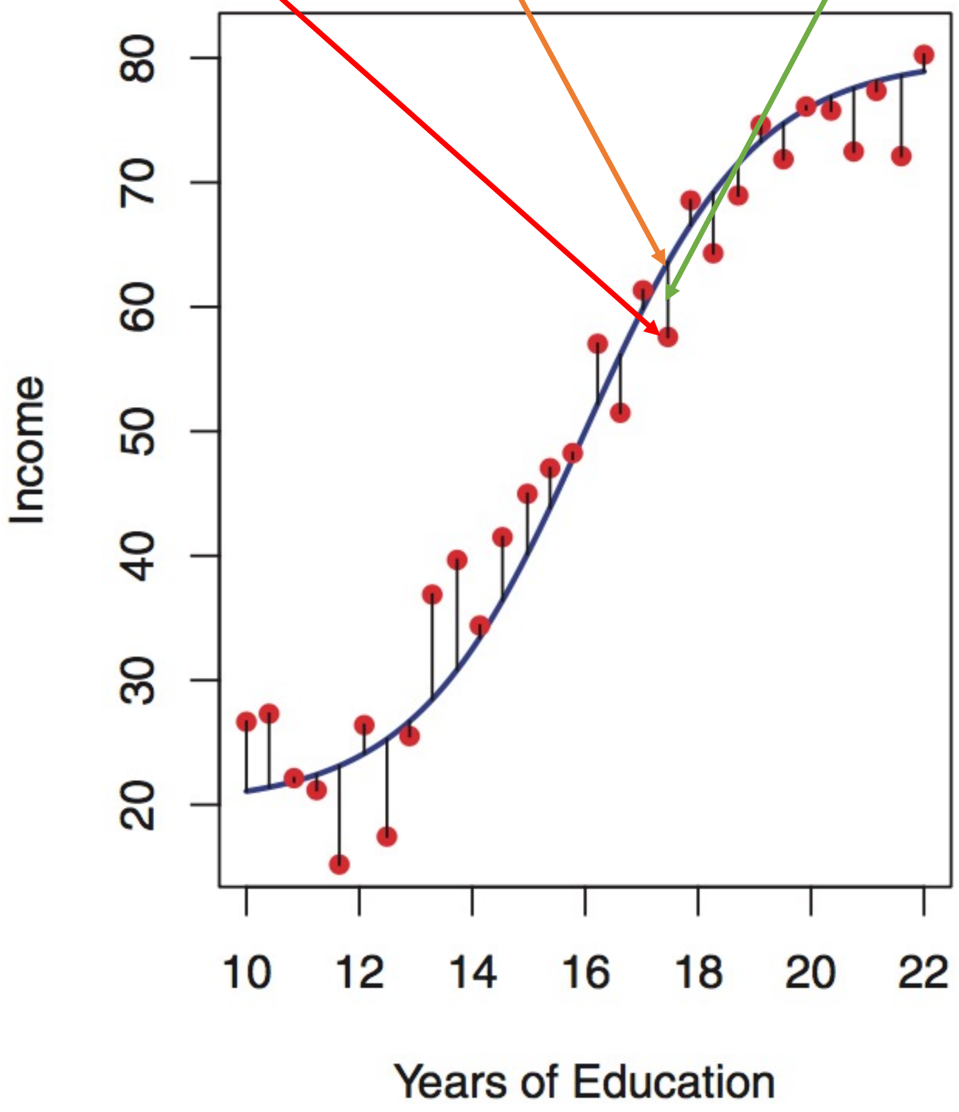
(c)

$$y = \underbrace{w_o + w_1 x + w_2 x^2 + w_3 x^8}_{f}$$

Why Does it Happen?

- Data is inherently **noisy**

$$y_i = f(x_i) + \varepsilon_i$$



Thus Use Separate Test Data

- Thus, if possible, we should try to use the test data $T_e = \{x_i, y_i\}_{i=1}^{M}$

$$\mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i \in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$$

Putting It All Together (3)

Training Phase

$$Tr = \{(\boldsymbol{x}_i, y_i) \in X \times Y : 1 \leq i \leq N\}$$

"Supervised" Learning Algorithm

$$\hat{f}: X \to Y$$

Evaluation Phase

$$Te = \{(\boldsymbol{x}_i, y_i) \in X \times Y : 1 \leq i \leq n\}$$

Learned Model

Important Take-away (1)

Goal of Machine Learning

Learn how do features relate to labels?

$$f: \boldsymbol{x} \to y$$

Features

Labels

$$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$$

For example: risk score of the patient

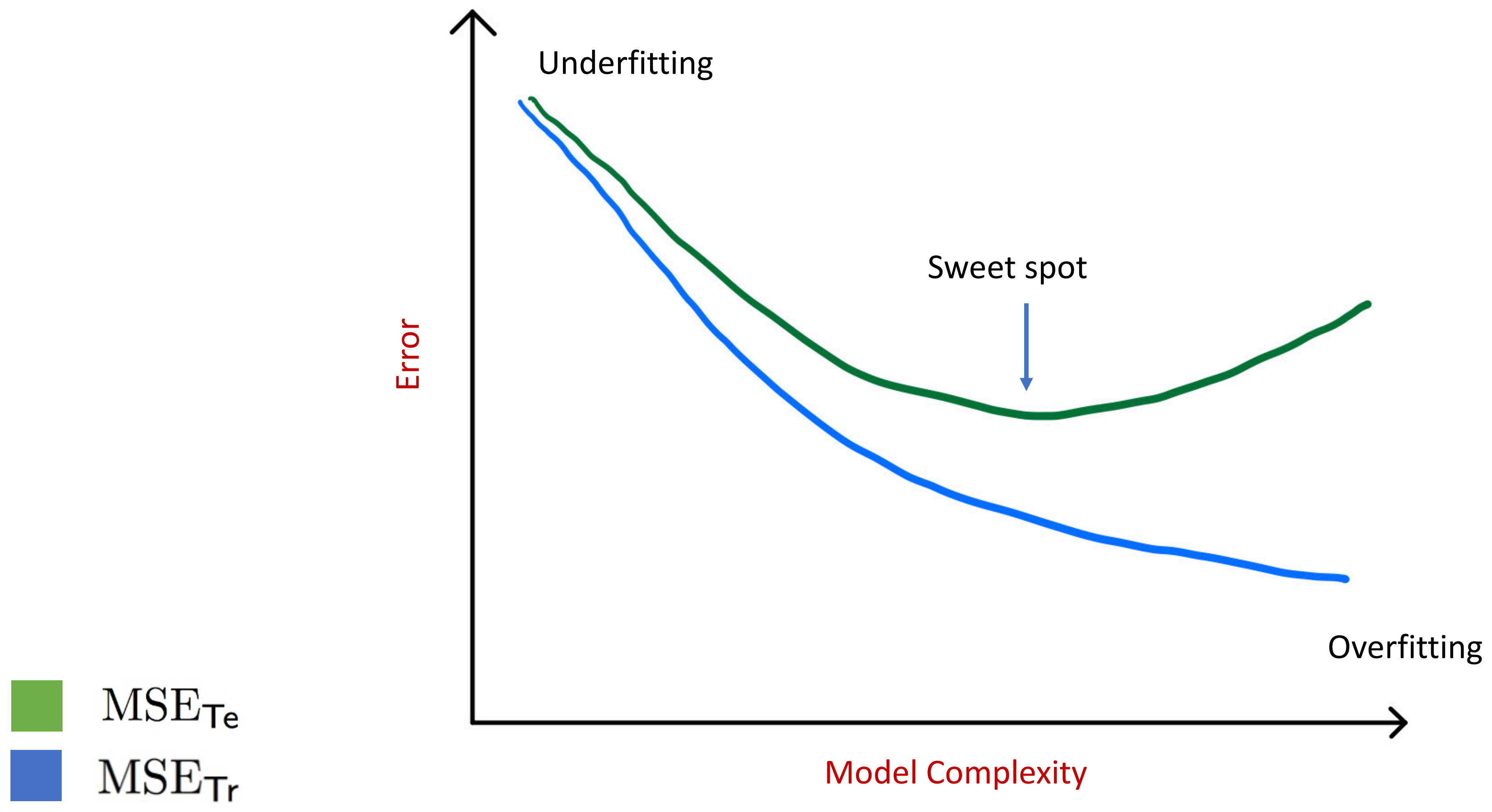
$y \in \mathbb{R}$ Regression

For example: COVID + OR COVID -

$y \in \{c_i\}_{i=1}^{k}$ Classification

For example: clinical Data of a patient

Important Take-away (2)



Bias Variance Tradeoff

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

Where (x_0, y_0) is a test observation

Typically, as the flexibility or complexity of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

Summary

- Course Introduction
- Overview of Machine Learning
  - Three components: Task, Experience, Performance
  - Predictors and Response
  - Goal of Learning
  - Classification and Regression
  - Parametric Models
  - Model Evaluation
  - Bias-Variance Tradeoff