Machine Learning Basics

Raghavendra Selvan
Erik Dam
Data Science Lab
Faculty of SCIENCE
raghav@di.ku.dk
@SuperVoxel

UNIVERSITY OF COPENHAGEN

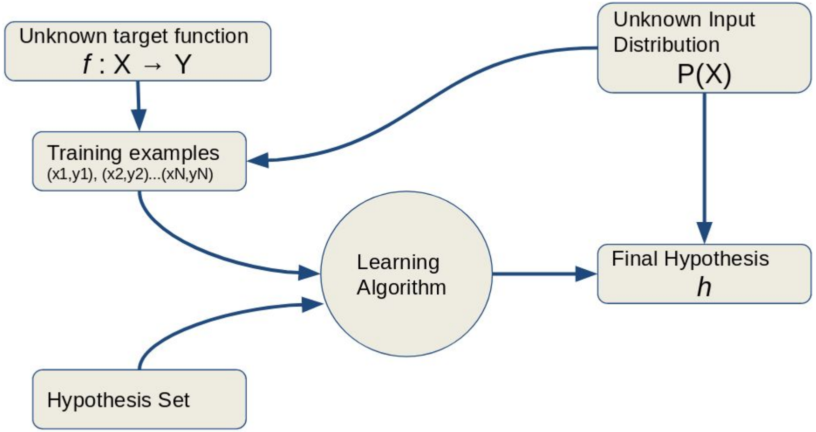
Overview

- Basics of Machine learning
- Types of learning
- Principles of Learning
- k-NN classifier

A learning algorithm

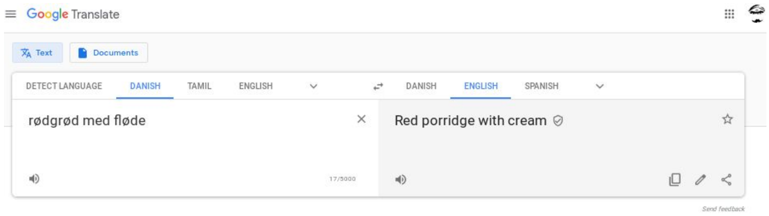
"A computer program is said to learn from experience **E** *with respect to some class of tasks* **T** *and performance measure* **P***, if its performance at tasks in* **T***, as measured by* **P***, improves with experience* **E***."*

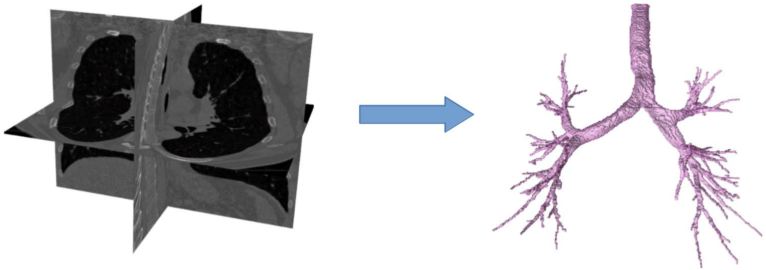
Mitchell, Tom M. Machine learning (1997)



The Task, T

- Classification
- Regression
- Transcription
- Machine translation
- Face recognition
- Anomaly detection
- Synthesis & sampling
- Denoising
- Density estimation
- Self-driving





The Performance measure, **P**

Not always straightforward but most common:

- Accuracy
- Error rates/ losses (0-1 loss)
- Log probability
- KL divergence

https://thispersondoesnotexist.com/

The Experience, **E**

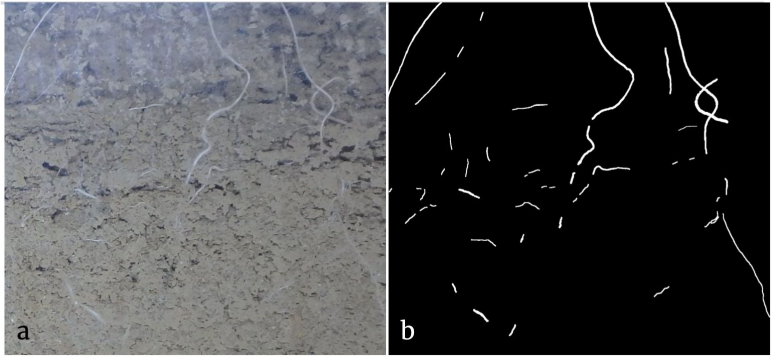
All the ways information can enter the model primarily as:

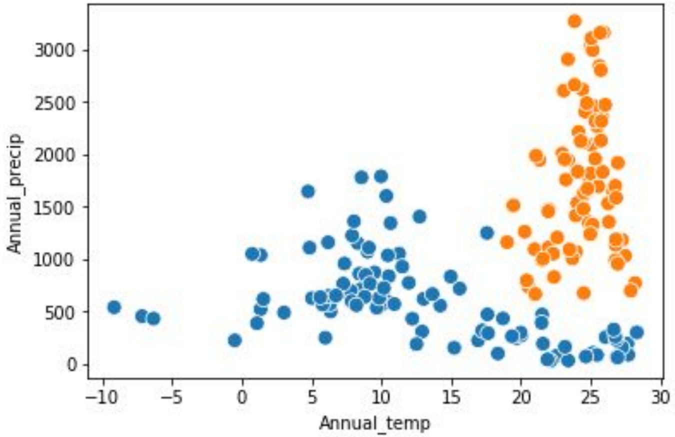
- Prior information
- Data/ supervision

More concrete classification of ML methods is based on **E**

Supervised Learning

- Strong labels for the entire dataset
- (Relatively) Easy to train
- Hard to obtain high quality labels
- Ex: Image Segmentation



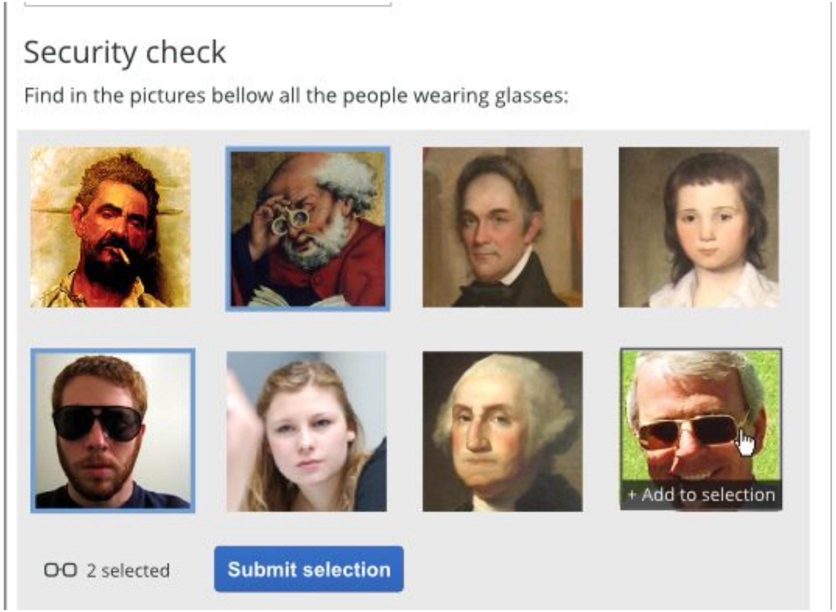


Unsupervised learning

- No labels.
- “Figure it out yourself” model
- Ex: Social networks, Gene expression networks

Semi-supervised learning

- Strong labels for some of the data
- Weak labels for all of the data
- Can be useful in cases where strong labels are hard!
- Ex: Captcha



Reinforcement learning

- Combination of strong and weak labels
- Online learning
- Constant learning
- Ex: Streaming services recommendation

AlphaGo

Task: Learning from YOUR data

1. socrative.com
2. Student login
3. Class name: **RAGHAV**

Principles of Learning

Generalization Error: Discrepancy between Training and Test performance

$$\mathbf{E}_{in}(h) = \frac{1}{n} \sum_{i=1}^{N} l(h(X_i), Y_i)$$

$$\mathbf{E}_{out}(h) = \mathbb{E}_{p(X,Y)}[l(h(X), Y)]$$

$$\mathcal{G}_{err} = \mathbf{E}_{out}(h) - \mathbf{E}_{in}(h)$$

Four horsemen of ML failure

1. Data assumptions
2. Data snooping
3. Underfitting
4. Overfitting

Data assumptions

1. **i.i.d**
    - **Identical:** Data is drawn from the same data distribution
    - **Independent:** Data points independent from each other
2. Sampling/Selection bias

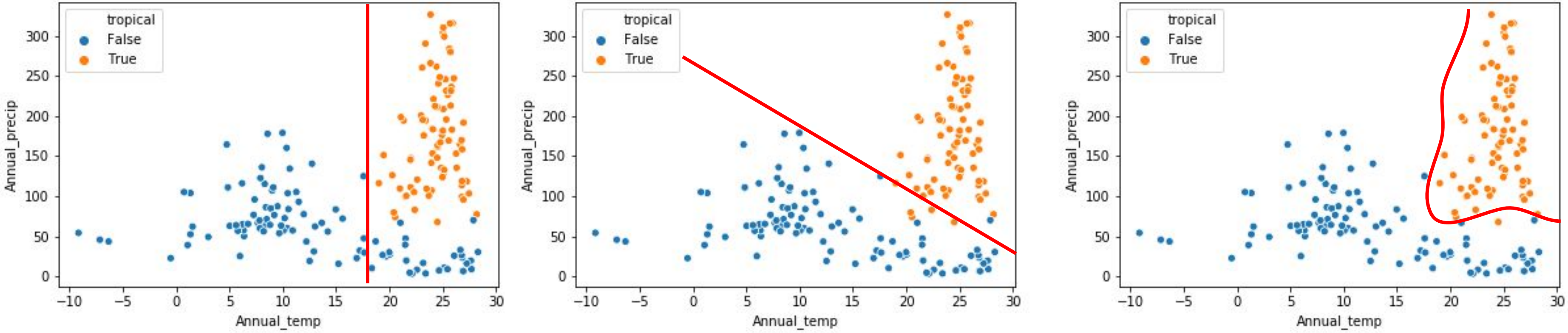
- If i.i.d assumption is violated does learning work?
- How can we overcome?

Data Snooping

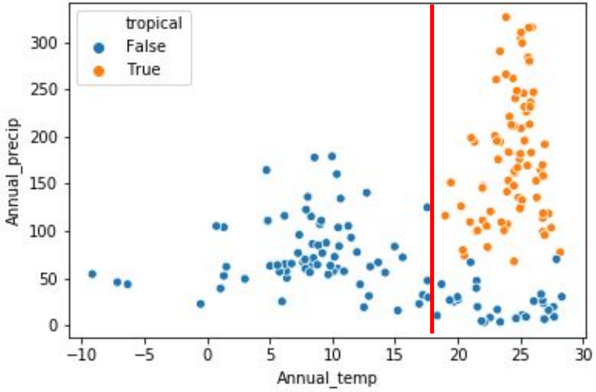
- Test data has informed the model selection
- Generalization suffers

“If you want an unbiased assessment of your learning performance, you should keep a test set in a vault and never use it for learning in any way”
Mostafa et al. Learning from data (book)

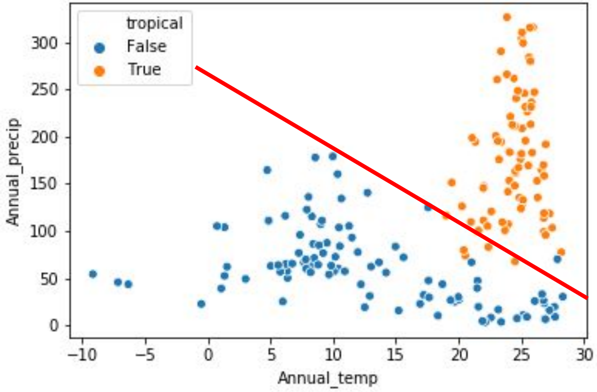
Underfitting & Overfitting



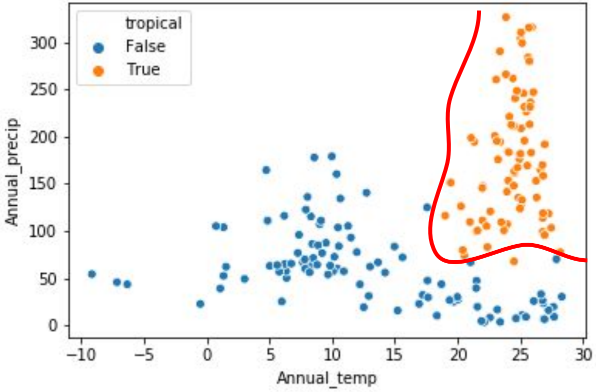
Underfitting & Overfitting



Underfitting



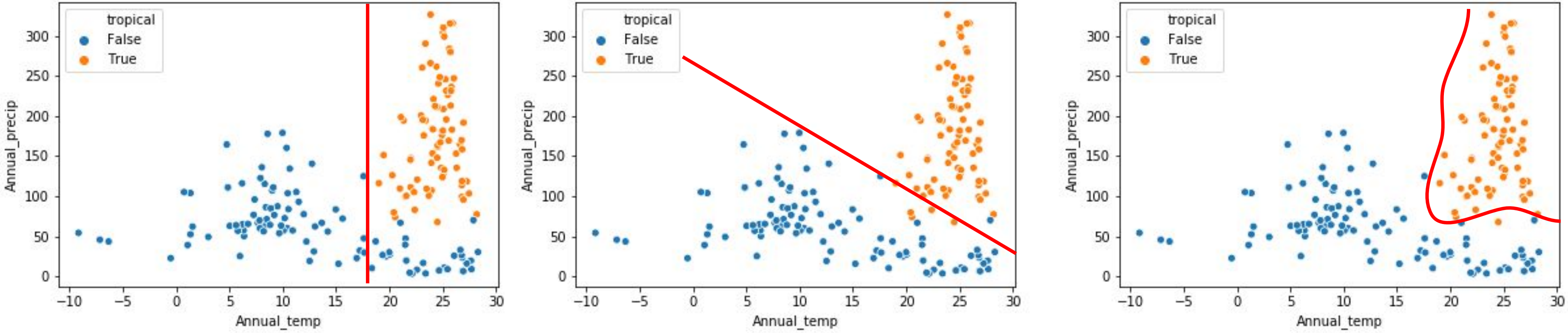
Appropriate capacity



Overfitting

Underfitting & Overfitting

- Models are chosen based on training error
- Test error ≥ Training error



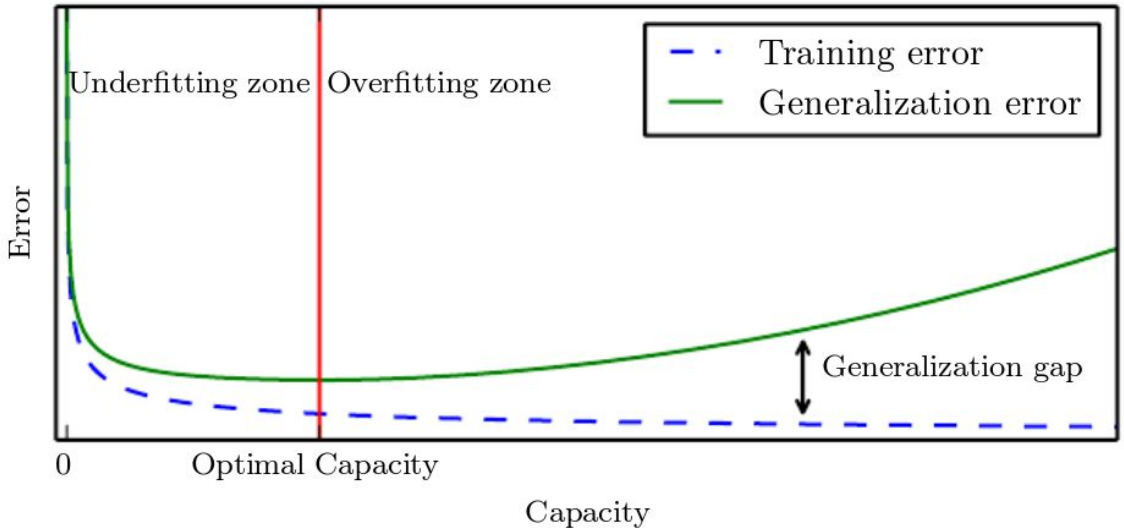
Underfitting

Appropriate capacity

Overfitting

Handling overfitting

- Representational capacity
  - **Occam's Razor:** *"The simplest model that fits the data is also the most plausible."*



Summary of Learning Principles

- Data is not ideal
- Lock away test data
- Low generalization error is the *Holy Grail* of all ML
- Model capacity is hard to decide, even with Occam's Razor
- Underfitting & Overfitting can hamper performance