Machine Learning

Zhou Zhao (赵洲)

College of Computer Science
Zhejiang University

zhaozhou@zju.edu.cn

Short Bio

- Dr. Zhou Zhao (赵洲)
  - zhaozhou@zju.edu.cn
- Professor at CS college (人工智能所).
  - 玉泉校区曹光彪楼主楼415室
- Research interests:
  - Machine learning
  - Data mining
  - Computer vision
  - …
- https://person.zju.edu.cn/zhaozhou

Course information (Cont'd)

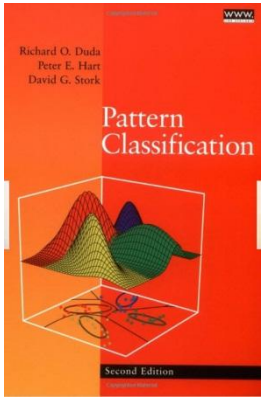
- Prerequisite:
  - Linear algebra, analysis, probability theory
  - Basic programming skills

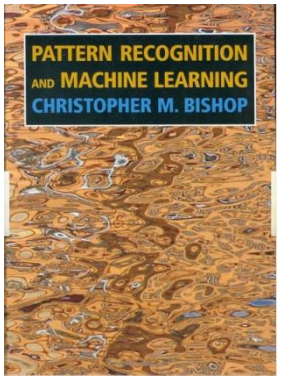
- Course textbook: No textbook is required. (Papers and other materials are available at the class web page)

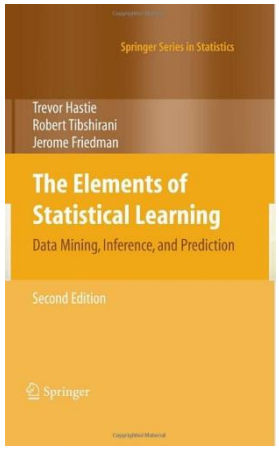
- Objective:
  - Basic understandings of some of the important machine learning methods.
  - Basic ability to use some machine learning techniques to solve real world problems.

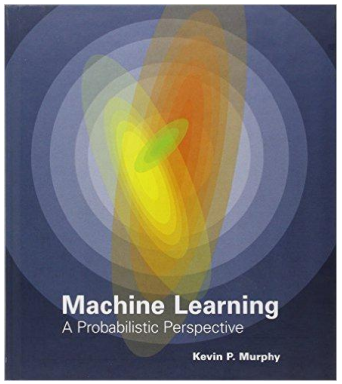
Reference Books

- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley, 2000
- C. M. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006
- T. Hastie, R. Tibshirani & J. Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** (2nd ed.), Springer, 2009
- Kevin Murphy, **Machine Learning: A Probabilistic Perspective**, The MIT Press, 2012









Evaluation

- Project and Presentation 90%
  - Presentation: 10%
  - Report: 20%
  - Programming Code: 60%
  - Report Deadline: **第15周周五晚上12点整**
- Attendance 10%
  - 签到20次，每次占0.5%

Report Requirement

- Good Presentation
- Good Survey
- Good Implementation
- Good Experimental Analysis
- Novel Ideas is much better (**but is not the requirement**)
- Report written using **Word** (>10 pages without reference)
- Code written by **Python (based on GPU or CPU)**

Topic

1. 语音识别 (ASR)
2. 语音合成（TTS）
3. 文字识别 (OCR)
4. 图片检测
5. 手语翻译/生成
6. 唇语翻译/生成
7. **图片分类 （http://yann.lecun.com/exdb/mnist/）（推荐Topic）**
8. 图片聚类
9. 图片分割
10. 机器翻译
11. 智能问答
12. 文本情感分类

Presentation Slot

- Send the email to RA (0920770@zju.edu.cn) to bid the presentation slot
  - e.g. prefers A > B > C > D

- Arrange the presentation slot based on your preference and the timestamp of the email
  - e.g. 16-th Mon
  - e.g. 16-th Feb

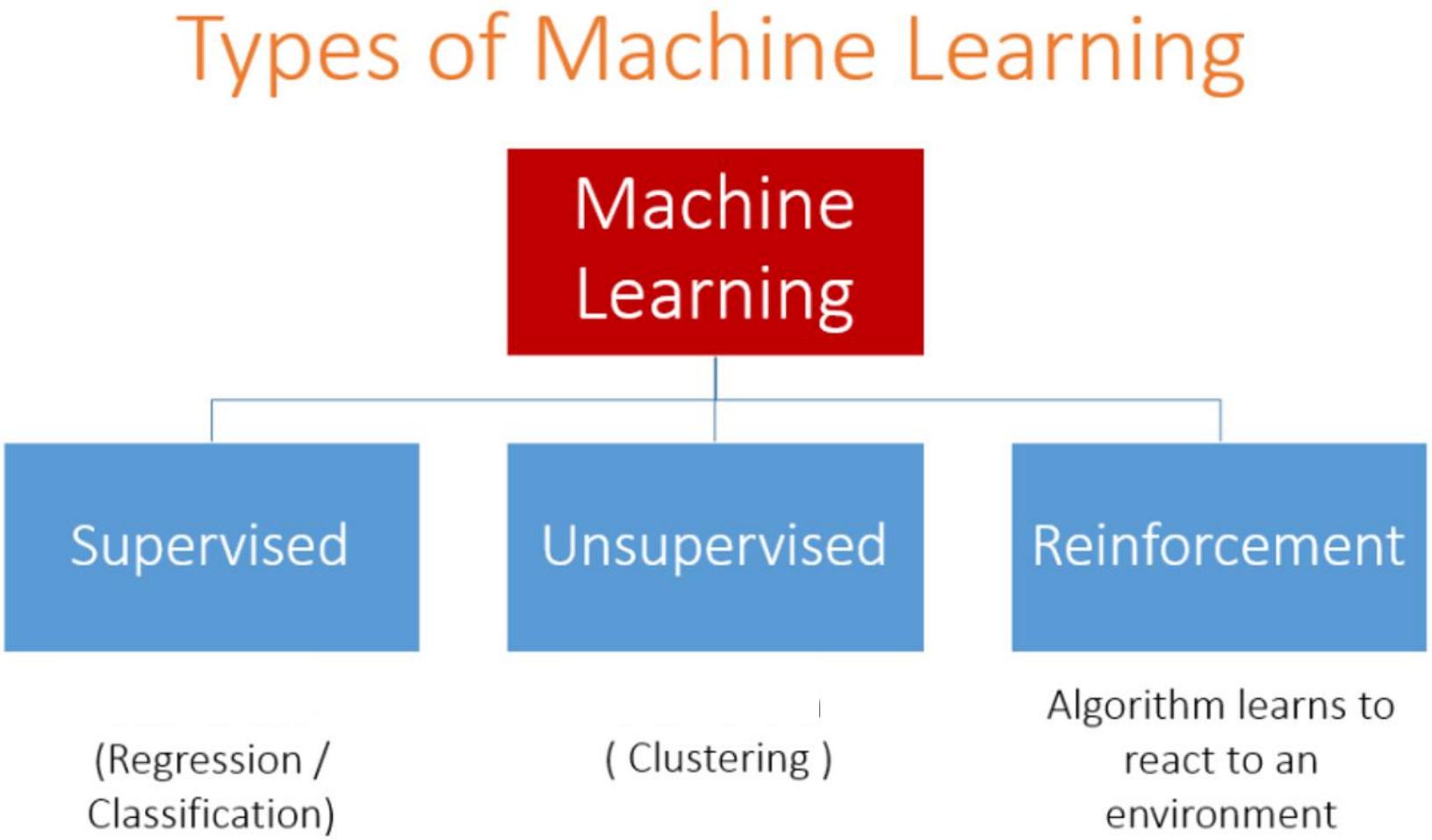
Course Schedule

1. Bayesian Theory (1 week)
2. LinearRegression/LinearClassifier (1 week)
3. Clustering/Neural Networks (1 week)
4. Dimension Reduction/Matrix Factorization/Topic Model (1 week)
5. Deep Learning Key Techniques (1 week)
6. CNN Network (2 weeks)
7. Transformer Network (2 weeks)
8. Speech (2 weeks)
9. Computer Vision (2 weeks)
10. Pytorch (3 weeks)
11. Project Presentation (1 weeks)

What is machine learning?

- Machine learning is the study of computer systems that improve their performance through experience.
  - Learn existing and known structures and rules.
    - Face recognition
  - Discover new findings and structures.
    - News summarization
- In machine learning, we study two types of problems

Types of Machine Learning



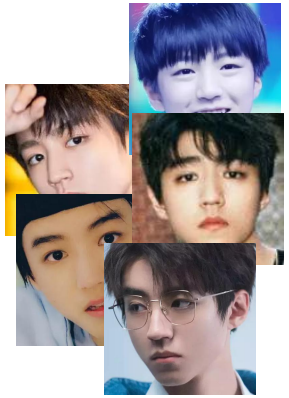
Supervised Learning

- Supervised learning
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y
  - Training data: a labeled set of input-output pairs
  - Classification (Categorization, Decision making…)
    - y is a categorical variable
  - Regression
    - y is real-valued

Supervised Learning (Classification)

刘德华

章子怡



王俊凯

……

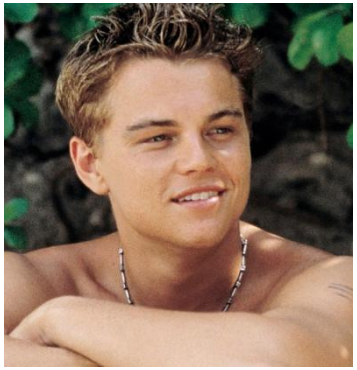
章子怡

Supervised Learning (Classification)

同一个人

不同人

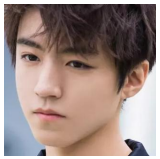
同一个人



Supervised Learning (Regression)

30岁

28岁



18岁

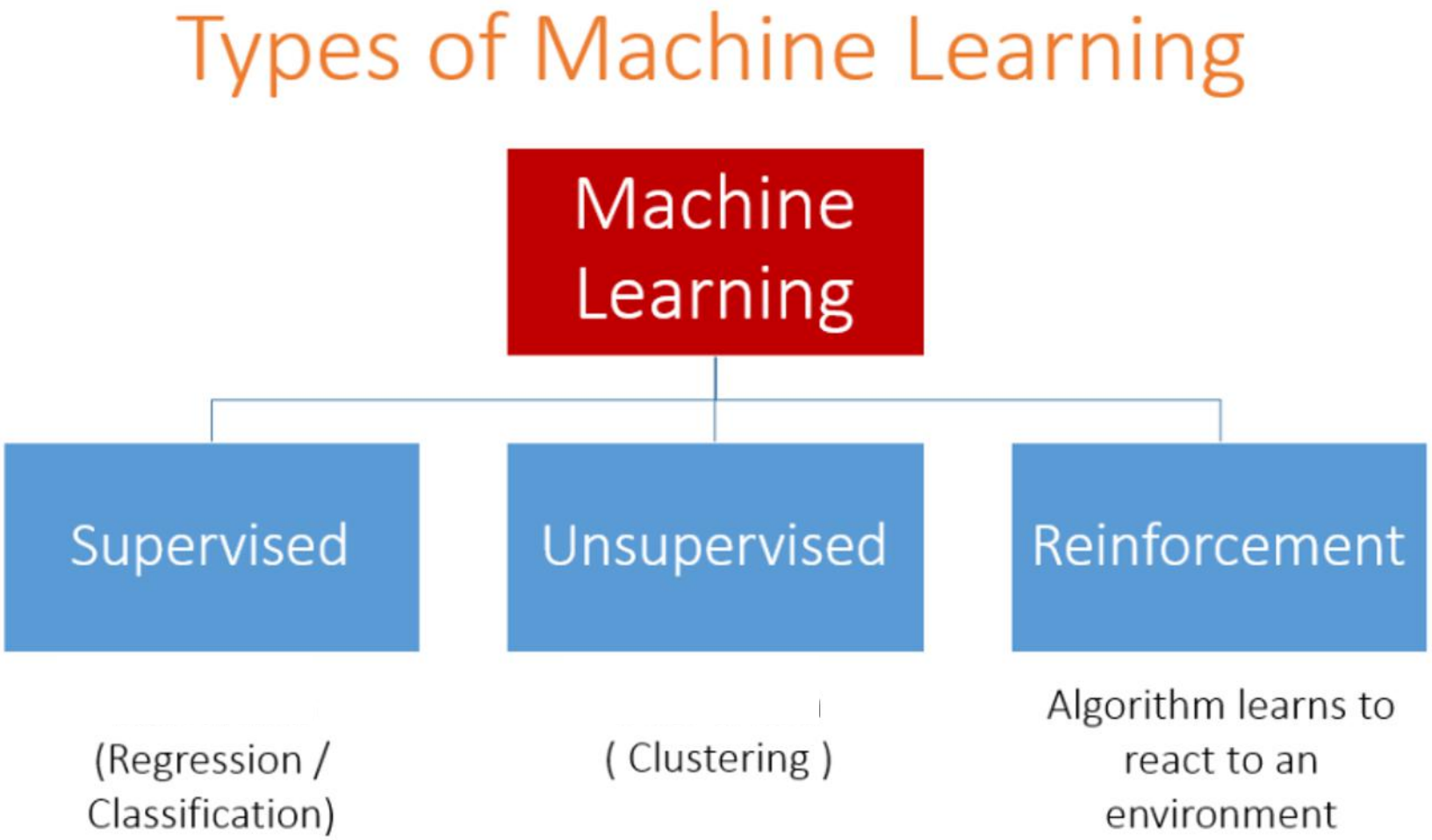
14岁

57岁

... ...

33岁

Types of Machine Learning



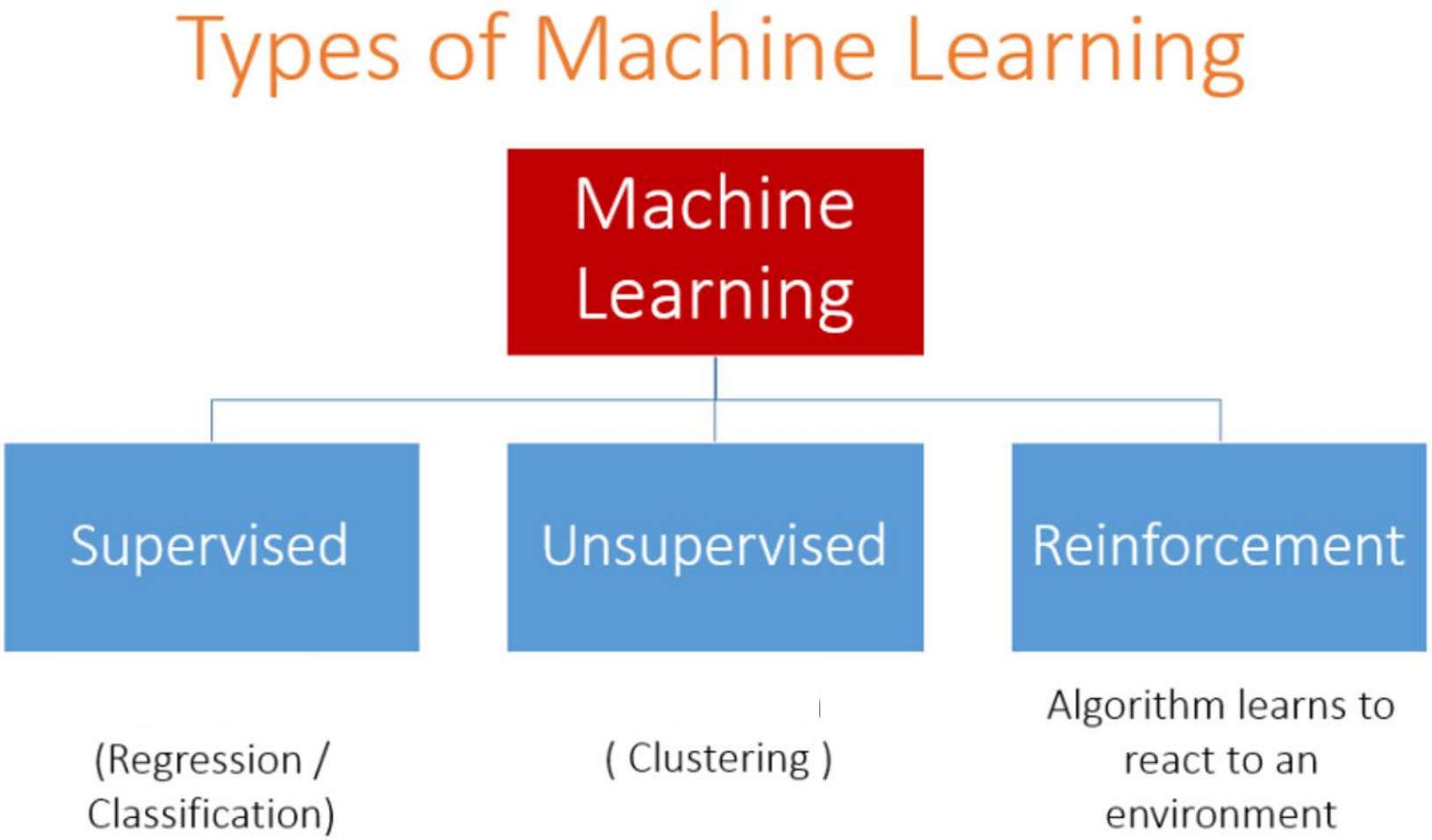
Unsupervised Learning

- Unsupervised learning
  - We are only given inputs
  - Goal: find "interesting patterns"
  - Much less well-defined problem

  - Discovering clusters, Clustering
  - Discovering latent factors
    - Dimensionality reduction, Matrix factorization, Topic modeling

Unsupervised Learning (Clustering)

Types of Machine Learning



Reinforcement Learning

- Reinforcement learning
  - It is a supervised learning scenario
  - No desired category signal is given
  - The only teaching feedback is that the tentative category is right or wrong.
  - This is useful for learning how to act or behave when given occasional reward or punishment signals.

Focus of This Course

- What are the typical machine learning **problems**?
  - Supervised Learning
    - Classification (decision making)
    - Regression
  - Unsupervised Learning
    - Cluster analysis
    - Latent factor analysis

- What are the basic machine learning **tools (methods, algorithms)**?

- Python programming

Basic Concepts of Supervised Learning

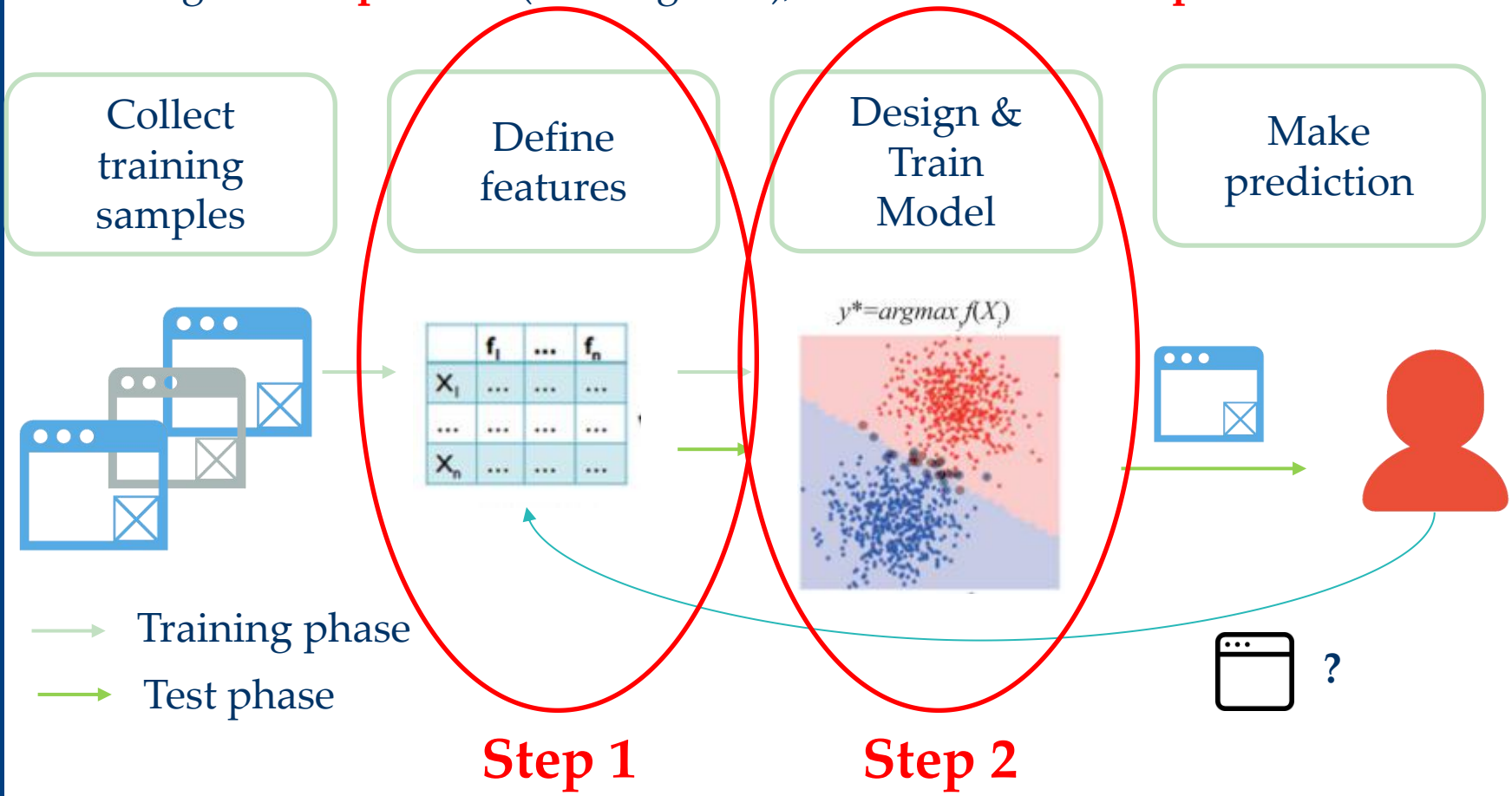
- Sample, example, pattern



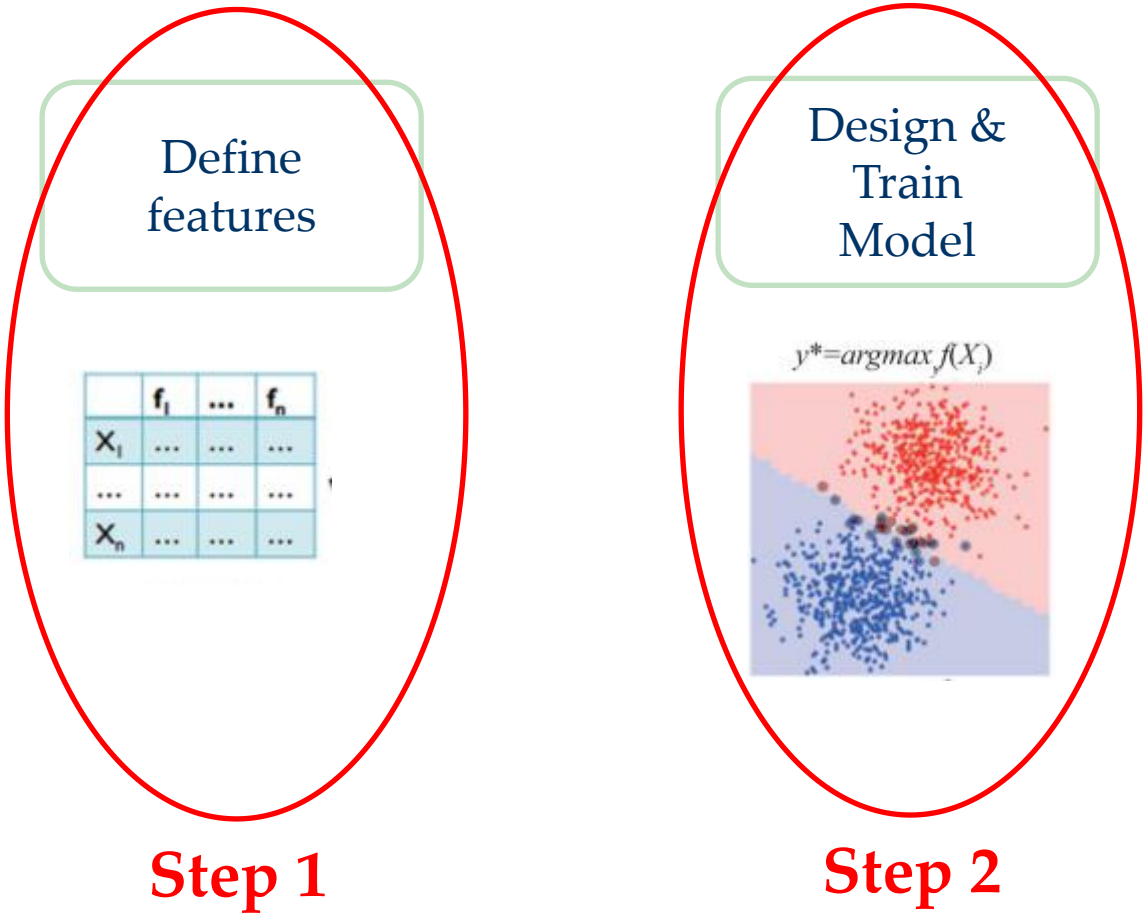
- Features, predictors, independent variables
  - $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n$
- State of the nature, labels, pattern class, class, responses, dependent variables
  - $\omega_1, \omega_2, \cdots \omega_c$ or $y_1, y_2, \cdots y_c$ or $z_1, z_2, \cdots z_c$
- Training data
  - $(\boldsymbol{x}_1, \omega_1), (\boldsymbol{x}_2, \omega_2), \cdots (\boldsymbol{x}_n, \omega_n)$
- Model, statistical model, pattern class model, classifier
  - f
- Test data
- Training error & test error

Supervised Learning

Learning from **experience**(training data), and build **model** to **predict** the future



Supervised Learning



- Which step is more important in building a successful system?
- Which one is the focus of this course?

Why general classification hard?

- Intra-class variability

The letter "T" in different typefaces

Define features

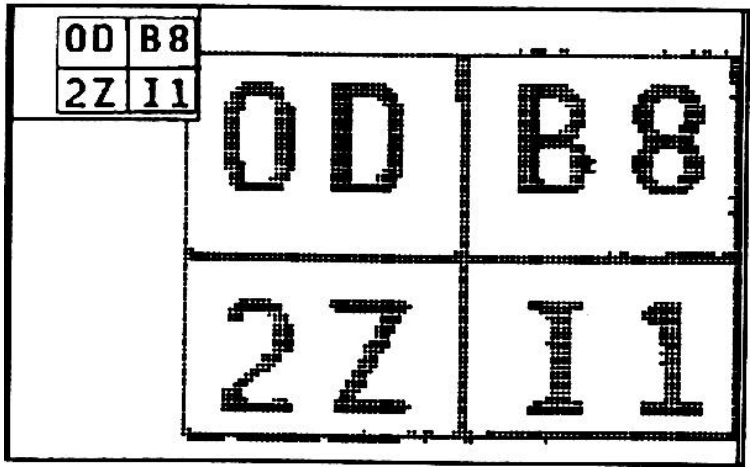
	f_1	...	f_n
x_1	...	...	...
...	...	...	...
x_n	...	...	...

Step 1 is not good enough

Same face under different expression, pose, illumination

Why general classification hard?

- Inter-class similarity



Define features

	f_1	...	f_n
x_1	...	...	...
...	...	...	...
x_n	...	...	...

Step 1 is not good enough

Representation: Features

- Extract features to represent the samples
- Feature vector

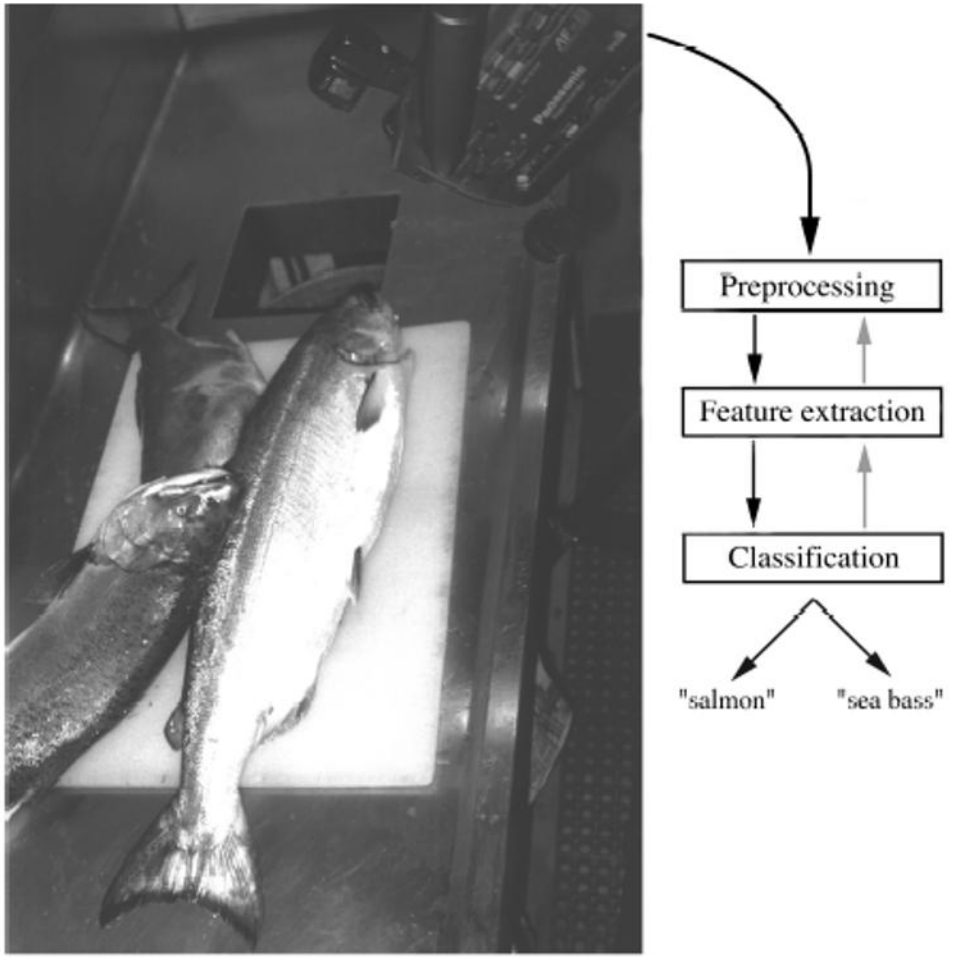
- Good representation:
  - Low intra-class variability
  - Low inter-class similarity

Fish Classification: Salmon v. Sea Bass

Preprocessing involves image enhancement and segmentation;

(i) separate touching or occluding fishes and

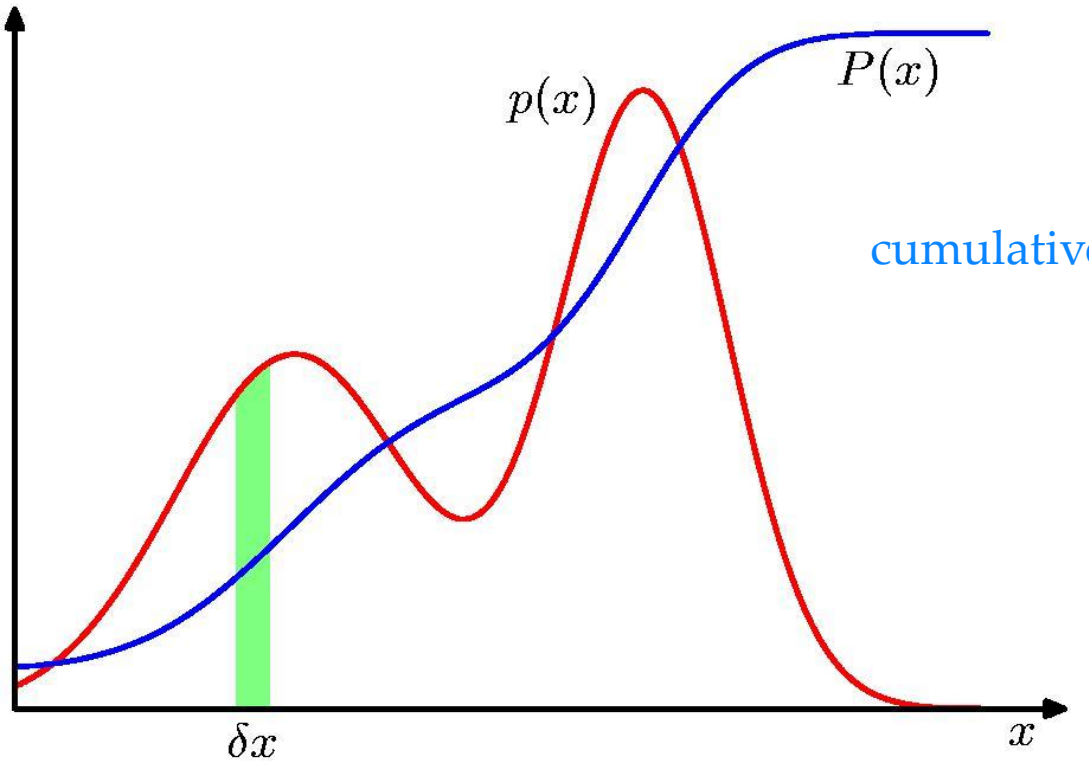
(ii) extract fish contour



How to design a classifier?

- Supervised learning
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y
    - Fish length as a feature
  - Training data: a labeled set of input-output pairs
    - (Salmon, 10cm)
    - (bass, 20cm)
    - …
  - Features of different class should be different.
    - Meaning what?

Probability Densities



$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

cumulative distribution function (CDF)

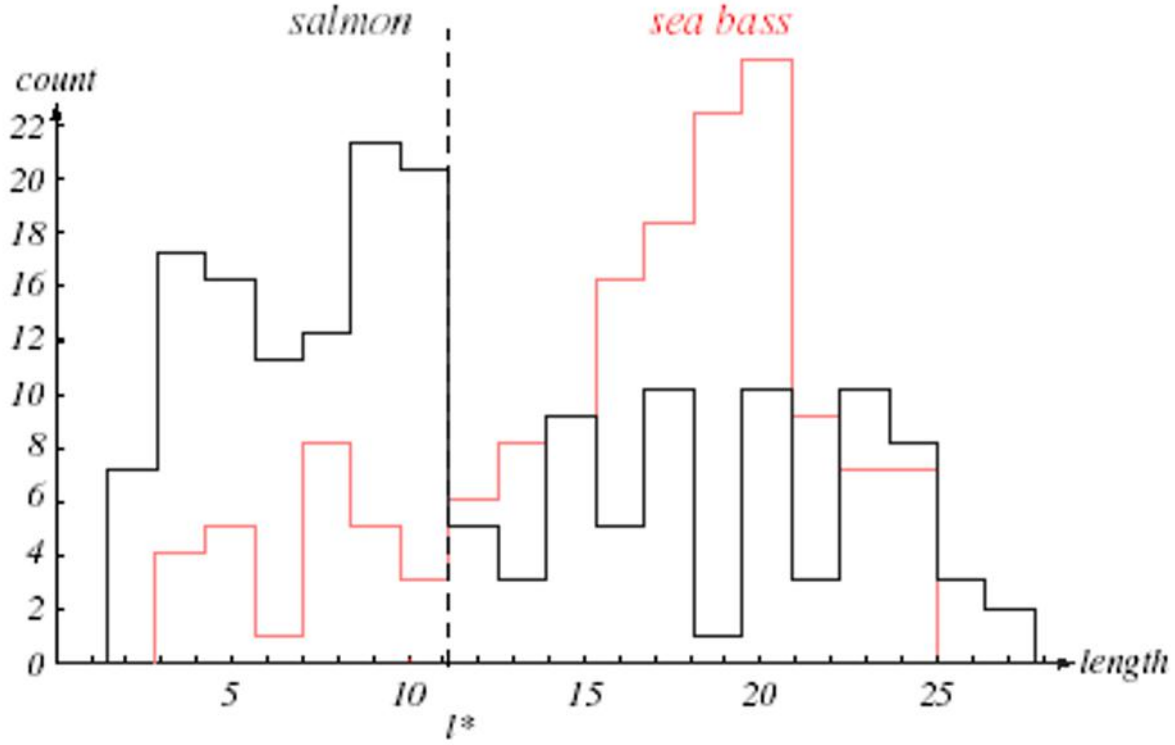
probability density function (PDF)

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \qquad p(x \in (a,b)) = \int_{a}^{b} p(x)\,\mathrm{d}x$$

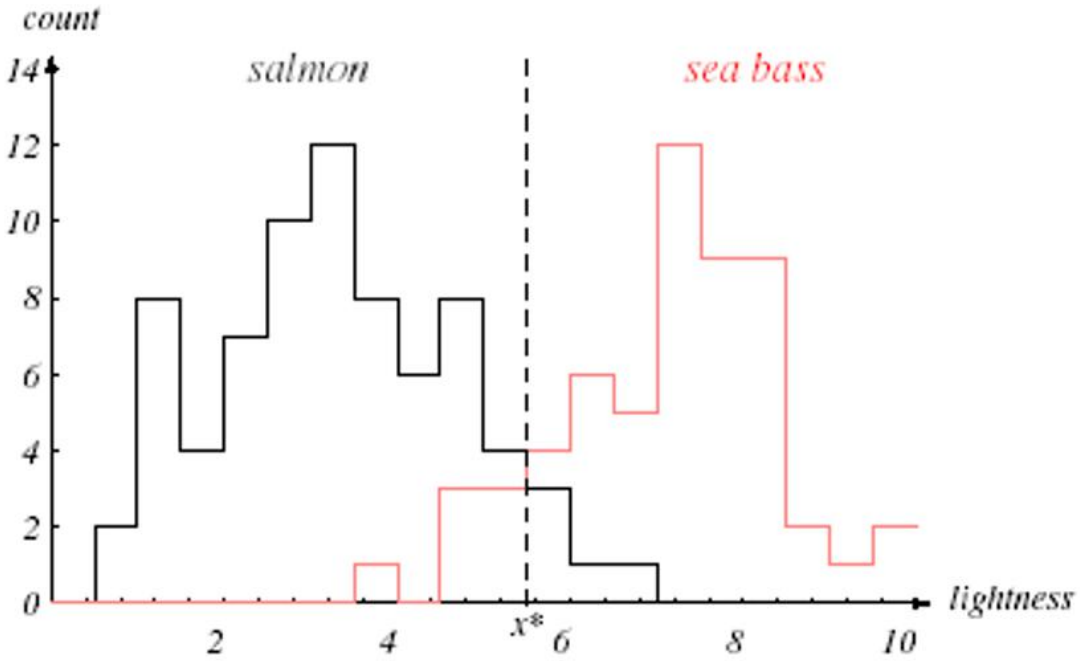
Representation: Fish Length As Feature

Training (design or learning) Samples

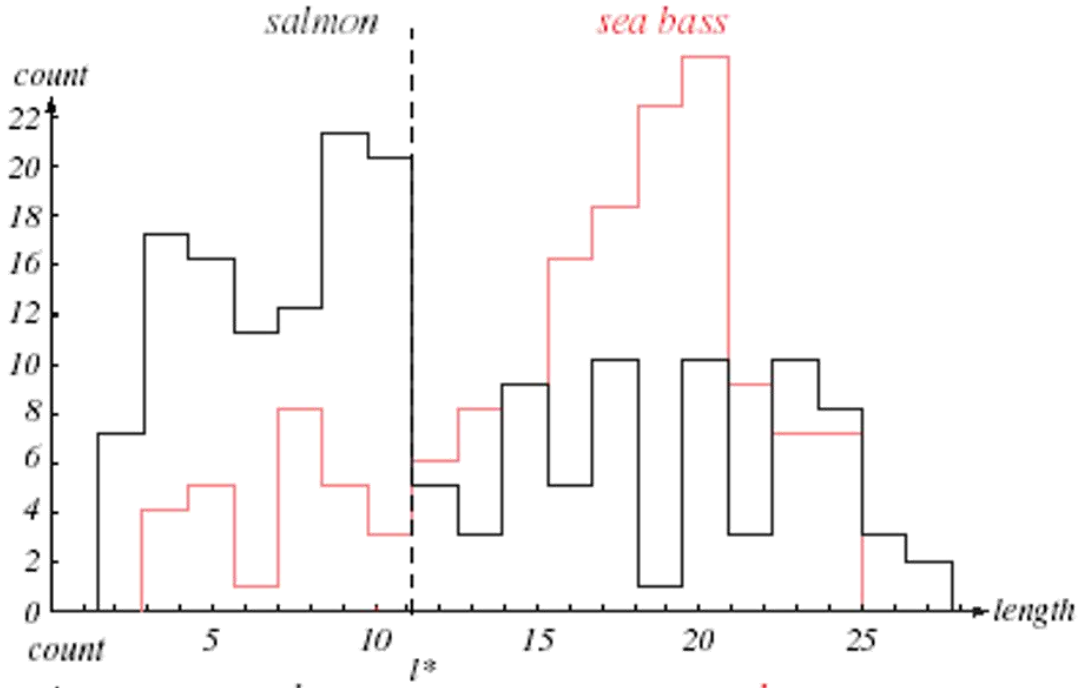
$p(x|\text{salmon})$ $p(x|\text{bass})$

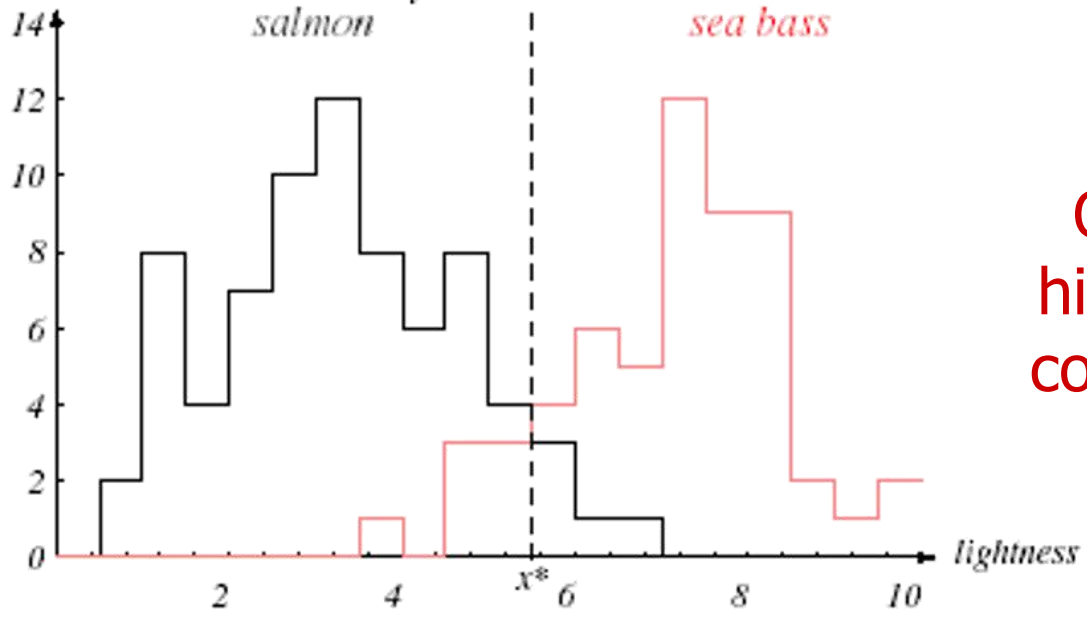


Fish Lightness As Feature



Which Feature is better

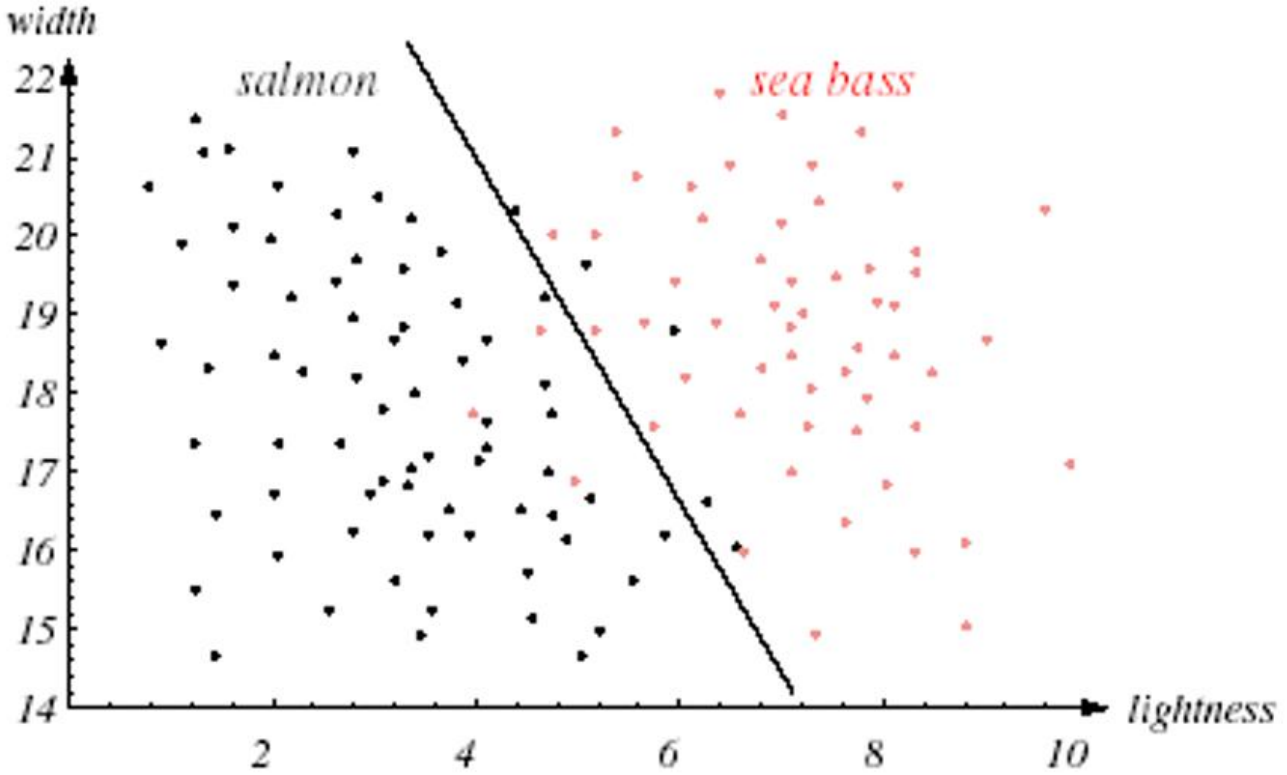




Overlap of these histograms is small compared to length feature

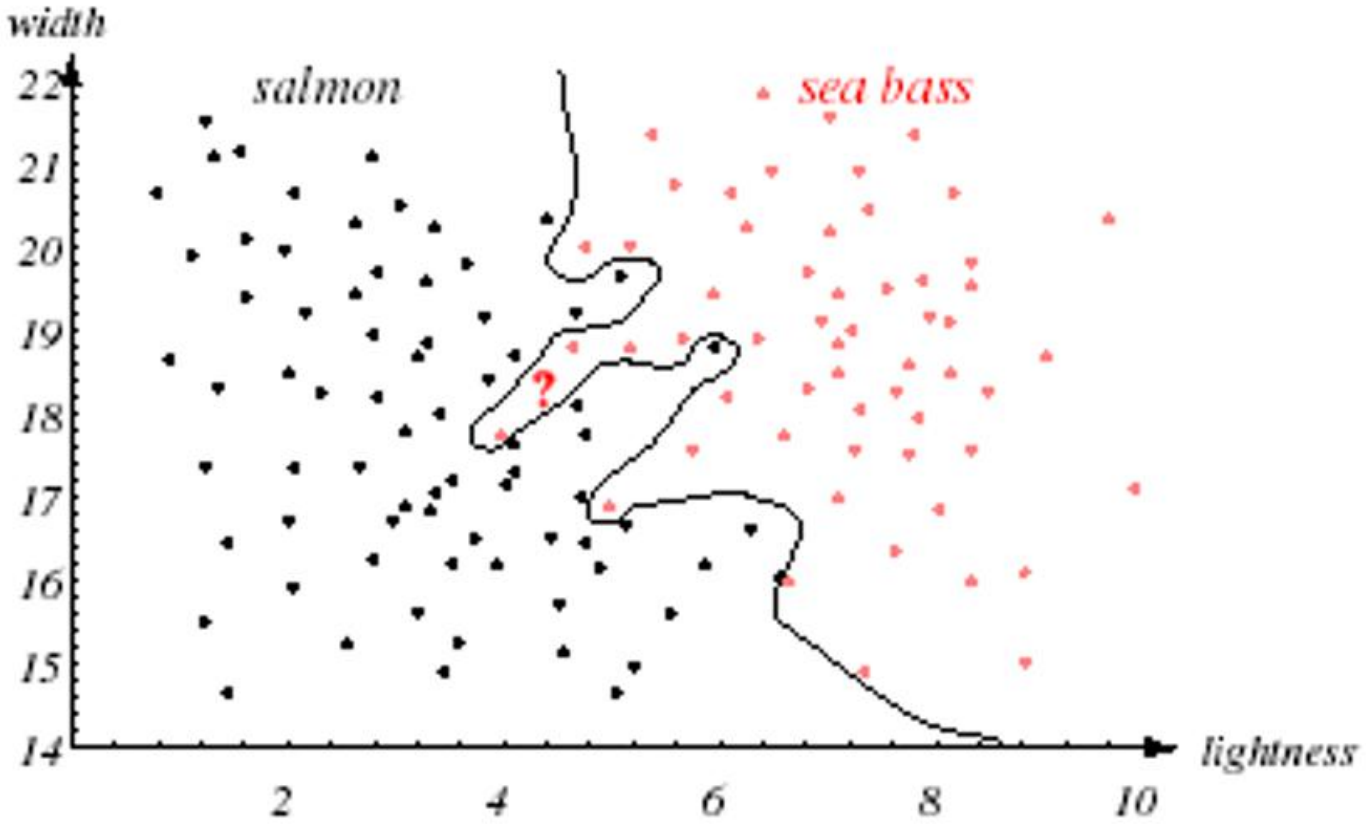
Two-dimensional Feature Space

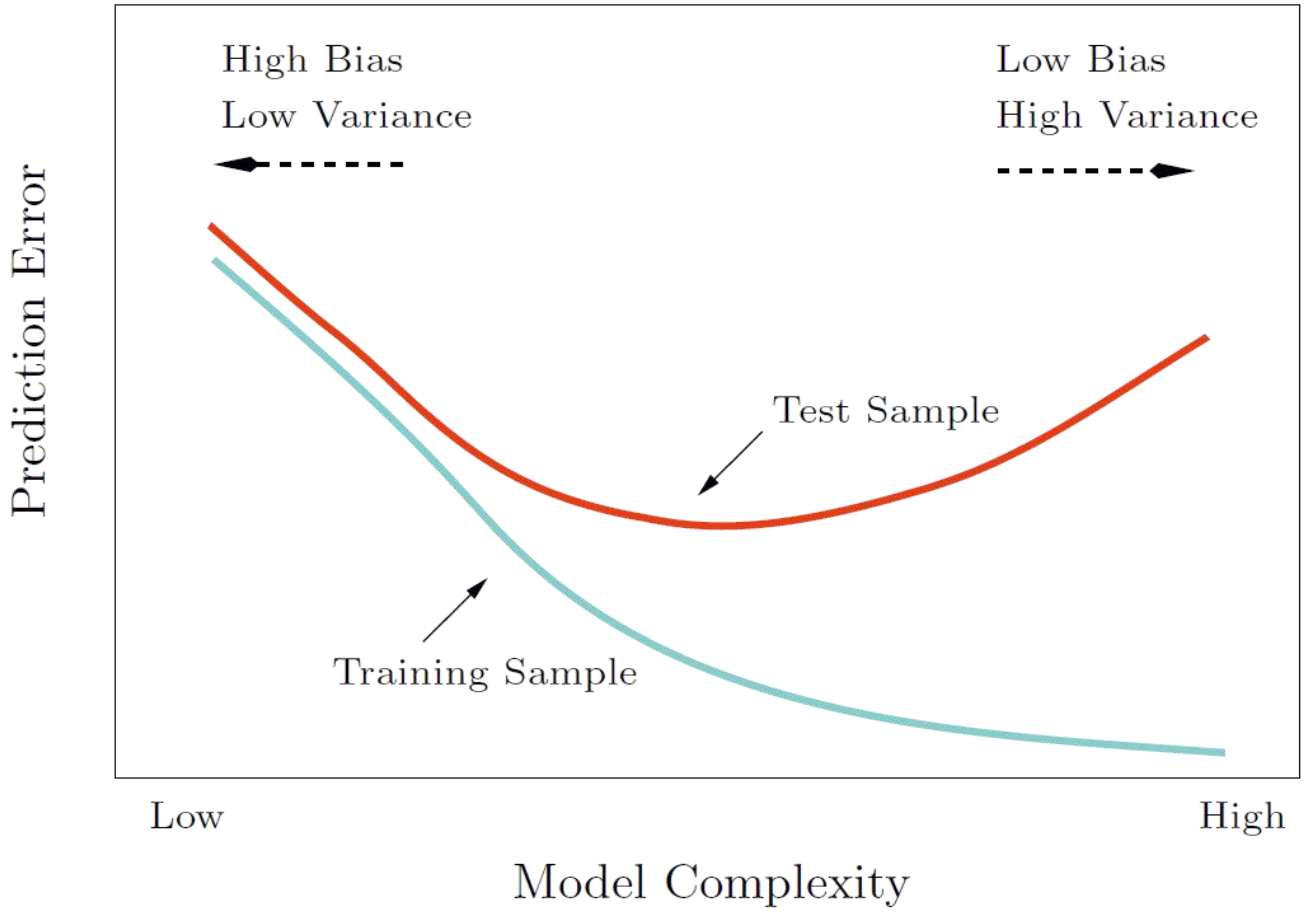
Linear (simple) decision boundary



Two features together are better than individual features

Complex Decision Boundary





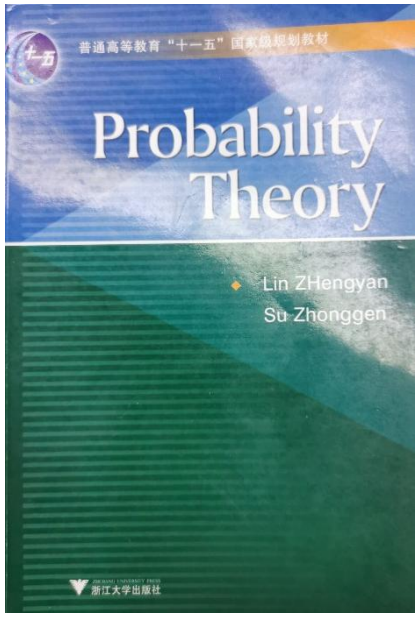
Generalization

- A generalization of a concept is an extension of the concept to less-specific criteria.
- Generalization of the classifier (model)
  - The performance of the classifier on test data.

- Training error:
- Simple model → large training error
- Complex model → less training error
- Test error:
- Simple model → ?
- Complex model → ?

Prerequisite Knowledge

- Probability:
  - 浙大出版社《概率论》
- Analysis:
  - 高教出版社《数学分析》上下
- Linear Algebra
  - 高教出版社《代数与几何》



Prerequisite Knowledge

- Probability: P p1-70
  - Bayes' rule, P p34
- Analysis:
  - Taylor series, A 上 p134
  - Constrained optimization, A 下 p176
    - Lagrangian multiplier, A 下 p343
- Linear Algebra
  - Linear space, L p58-82
  - Matrix , L p119-150
    - Rank, L p139
    - Positive definite matrix, L p263
    - Eigenvector, eigenvalue , L p234
    - Singular vector, singular value, wiki