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
- Associate 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)

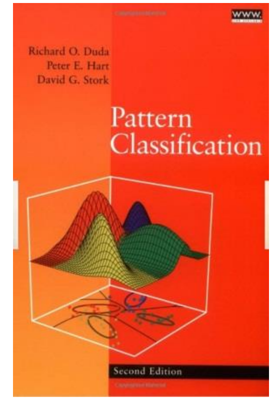
- Prerequisite:
  - Linear algebra, analysis, probability theory
  - Basic programming skills

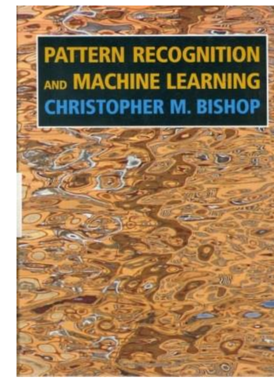
- Course textbook: No textbook is required. (Papers and other materials are available at the class web page)

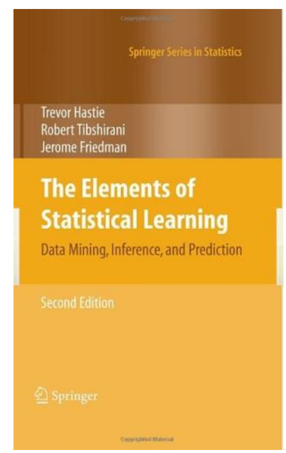
- Objective:
  - Basic understandings of some of the important machine learning methods.
  - Basic ability to use some machine learning techniques to solve real world problems.

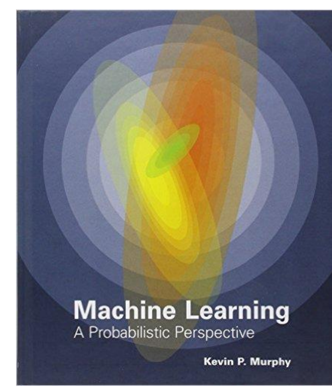
Reference Books

- 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









评测指标

- 大作业（图片分类）：70%
  - 思路PPT讲解：10%
  - 作业报告：30%
  - 编程代码：30%
  - 报告截止日期：第15周周五晚上12点整
- 小作业：
  - 阅读SVM开源算法报告：10%
    - 报告截止日期：第8周周五晚上12点整
  - 阅读Transformer开源算法报告：10%
    - 报告截止日期：第15周周五晚上12点整
- 课堂参与：10%
  - 签到10次，每次占1%

大作业

图片分类（http://yann.lecun.com/exdb/mnist/）

CLASSIFIER	PREPROCESSING	TEST ERROR RATE (%)	Reference
Linear Classifiers			
linear classifier (1-layer NN)	none	12.0	LeCun et al. 1998
linear classifier (1-layer NN)	deskewing	8.4	LeCun et al. 1998
pairwise linear classifier	deskewing	7.6	LeCun et al. 1998
K-Nearest Neighbors			
K-nearest-neighbors, Euclidean (L2)	none	5.0	LeCun et al. 1998
K-nearest-neighbors, Euclidean (L2)	none	3.09	Kenneth Wilder, U. Chicago
K-nearest-neighbors, L3	none	2.83	Kenneth Wilder, U. Chicago
K-nearest-neighbors, Euclidean (L2)	deskewing	2.4	LeCun et al. 1998
K-nearest-neighbors, Euclidean (L2)	deskewing, noise removal, blurring	1.80	Kenneth Wilder, U. Chicago
K-nearest-neighbors, L3	deskewing, noise removal, blurring	1.73	Kenneth Wilder, U. Chicago
K-nearest-neighbors, L3	deskewing, noise removal, blurring, 1 pixel shift	1.33	Kenneth Wilder, U. Chicago
K-nearest-neighbors, L3	deskewing, noise removal, blurring, 2 pixel shift	1.22	Kenneth Wilder, U. Chicago
K-NN with non-linear deformation (IDM)	shiftable edges	0.54	Keysers et al. IEEE PAMI 2007
K-NN with non-linear deformation (P2DHMDM)	shiftable edges	0.52	Keysers et al. IEEE PAMI 2007
K-NN, Tangent Distance	subsampling to 16x16 pixels	1.1	LeCun et al. 1998
K-NN, shape context matching	shape context feature extraction	0.63	Belongie et al. IEEE PAMI 2002
Boosted Stumps			
boosted stumps	none	7.7	Kegl et al., ICML 2009
products of boosted stumps (3 terms)	none	1.26	Kegl et al., ICML 2009
boosted trees (17 leaves)	none	1.53	Kegl et al., ICML 2009
stumps on Haar features	Haar features	1.02	Kegl et al., ICML 2009
product of stumps on Haar f.	Haar features	0.87	Kegl et al., ICML 2009
Non-Linear Classifiers			
40 PCA + quadratic classifier	none	3.3	LeCun et al. 1998
1000 RBF + linear classifier	none	3.6	LeCun et al. 1998
SVMs			
SVM, Gaussian Kernel	none	1.4	
SVM deg 4 polynomial	deskewing	1.1	LeCun et al. 1998
Reduced Set SVM deg 5 polynomial	deskewing	1.0	LeCun et al. 1998
Virtual SVM deg-9 poly [distortions]	none	0.8	LeCun et al. 1998
Virtual SVM, deg-9 poly, 1-pixel jittered	none	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 1-pixel jittered	deskewing	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 2-pixel jittered	deskewing	0.56	DeCoste and Scholkopf, MLJ 2002

Neural Nets			
2-layer NN, 300 hidden units, mean square error	none	4.7	LeCun et al. 1998
2-layer NN, 300 HU, MSE, [distortions]	none	3.6	LeCun et al. 1998
2-layer NN, 300 HU	deskewing	1.6	LeCun et al. 1998
2-layer NN, 1000 hidden units	none	4.5	LeCun et al. 1998
2-layer NN, 1000 HU, [distortions]	none	3.8	LeCun et al. 1998
3-layer NN, 300+100 hidden units	none	3.05	LeCun et al. 1998
3-layer NN, 300+100 HU [distortions]	none	2.5	LeCun et al. 1998
3-layer NN, 500+150 hidden units	none	2.95	LeCun et al. 1998
3-layer NN, 500+150 HU [distortions]	none	2.45	LeCun et al. 1998
3-layer NN, 500+300 HU, softmax, cross entropy, weight decay	none	1.53	Hinton, unpublished, 2005
2-layer NN, 800 HU, Cross-Entropy Loss	none	1.6	Simard et al., ICDAR 2003
2-layer NN, 800 HU, cross-entropy [affine distortions]	none	1.1	Simard et al., ICDAR 2003
2-layer NN, 800 HU, MSE [elastic distortions]	none	0.9	Simard et al., ICDAR 2003
2-layer NN, 800 HU, cross-entropy [elastic distortions]	none	0.7	Simard et al., ICDAR 2003
NN, 784-500-500-2000-30 + nearest neighbor, RBM + NCA training [no distortions]	none	1.0	Salakhutdinov and Hinton, AI-Stats 2007
6-layer NN 784-2500-2000-1500-1000-500-10 (on GPU) [elastic distortions]	none	0.35	Ciresan et al. Neural Computation 10, 2010 and arXiv 1003.0358, 2010
committee of 25 NN 784-800-10 [elastic distortions]	width normalization, deslanting	0.39	Meier et al. ICDAR 2011
deep convex net, unsup pre-training [no distortions]	none	0.83	Deng et al. Interspeech 2010
Convolutional nets			
Convolutional net LeNet-1	subsampling to 16x16 pixels	1.7	LeCun et al. 1998
Convolutional net LeNet-4	none	1.1	LeCun et al. 1998
Convolutional net LeNet-4 with K-NN instead of last layer	none	1.1	LeCun et al. 1998
Convolutional net LeNet-4 with local learning instead of last layer	none	1.1	LeCun et al. 1998
Convolutional net LeNet-5, [no distortions]	none	0.95	LeCun et al. 1998
Convolutional net LeNet-5, [huge distortions]	none	0.85	LeCun et al. 1998
Convolutional net LeNet-5, [distortions]	none	0.8	LeCun et al. 1998
Convolutional net Boosted LeNet-4, [distortions]	none	0.7	LeCun et al. 1998
Trainable feature extractor + SVMs [no distortions]	none	0.83	Lauer et al., Pattern Recognition 40-6, 2007
Trainable feature extractor + SVMs [elastic distortions]	none	0.56	Lauer et al., Pattern Recognition 40-6, 2007
Trainable feature extractor + SVMs [affine distortions]	none	0.54	Lauer et al., Pattern Recognition 40-6, 2007
unsupervised sparse features + SVM, [no distortions]	none	0.59	Labusch et al., IEEE TNN 2008
Convolutional net, cross-entropy [affine distortions]	none	0.6	Simard et al., ICDAR 2003
Convolutional net, cross-entropy [elastic distortions]	none	0.4	Simard et al., ICDAR 2003
large conv. net, random features [no distortions]	none	0.89	Ranzato et al., CVPR 2007
large conv. net, unsup features [no distortions]	none	0.62	Ranzato et al., CVPR 2007
large conv. net, unsup pretraining [no distortions]	none	0.60	Ranzato et al., NIPS 2006
large conv. net, unsup pretraining [elastic distortions]	none	0.39	Ranzato et al., NIPS 2006
large conv. net, unsup pretraining [no distortions]	none	0.53	Jarrett et al., ICCV 2009
large/deep conv. net, 1-20-40-60-80-100-120-120-10 [elastic distortions]	none	0.35	Ciresan et al. IJCAI 2011
committee of 7 conv. net, 1-20-P-40-P-150-10 [elastic distortions]	width normalization	0.27 +-0.02	Ciresan et al. ICDAR 2011
committee of 35 conv. net, 1-20-P-40-P-150-10 [elastic distortions]	width normalization	0.23	Ciresan et al. CVPR 2012

大作业要求

- 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)**

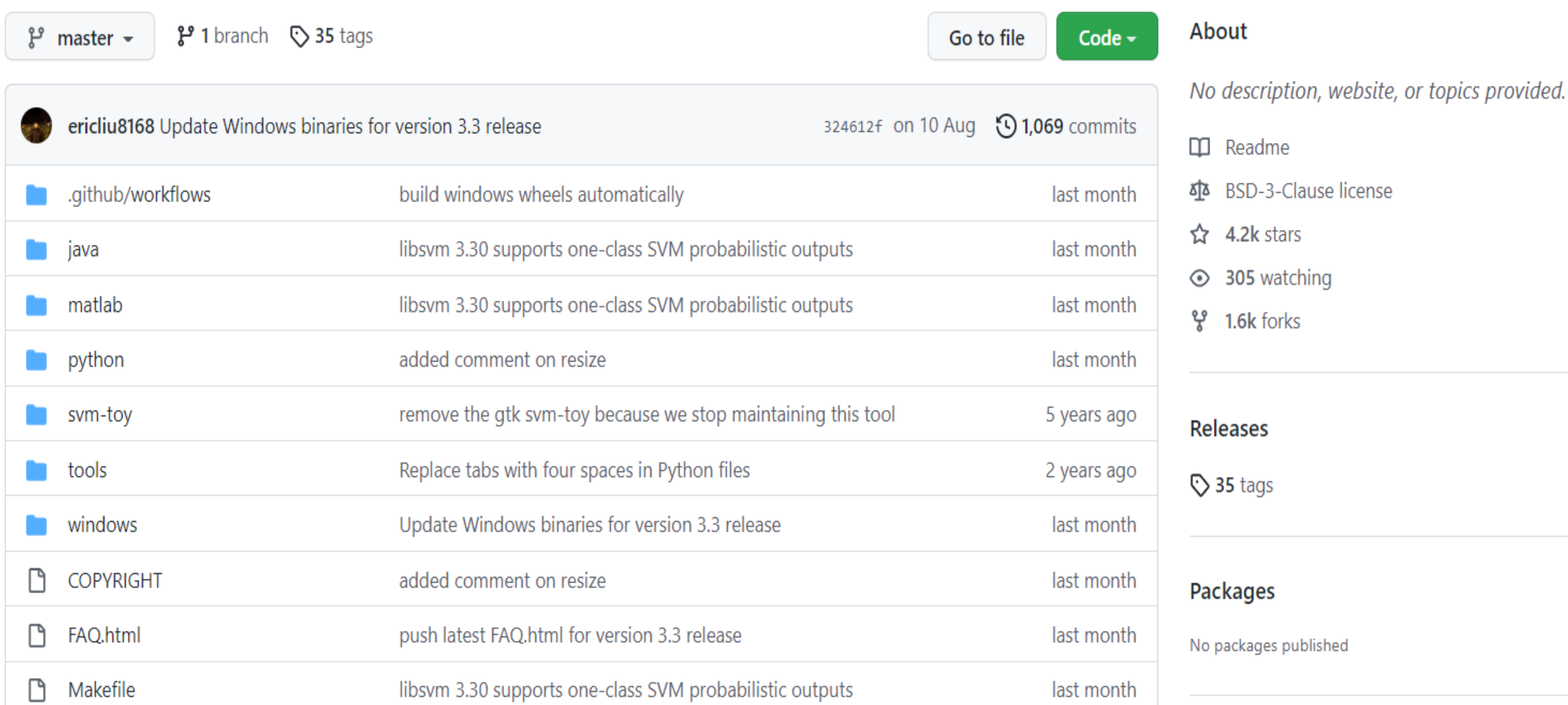
Presentation Slot

- Send the email to RA (jiangqingyun@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，including
  - 15-th week Monday
  - 15-th week Tuesday
  - 16-th week Monday
  - 16-th week Tuesday

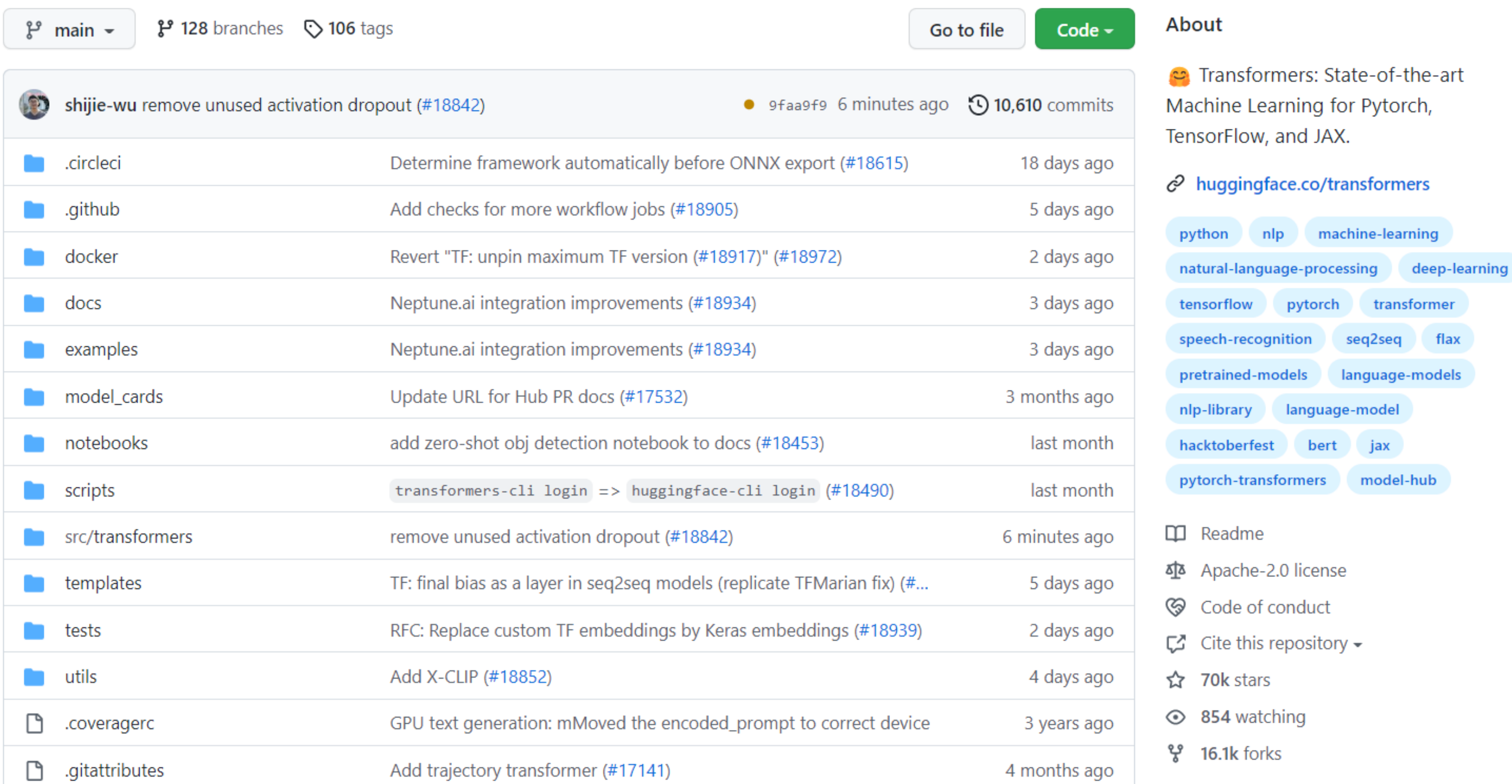
小作业1

- **LIBSVM (https://github.com/cjlin1/libsvm)**
- 6页报告（包括SVM原理，代码理解以及数据集上实验结果）



小作业2

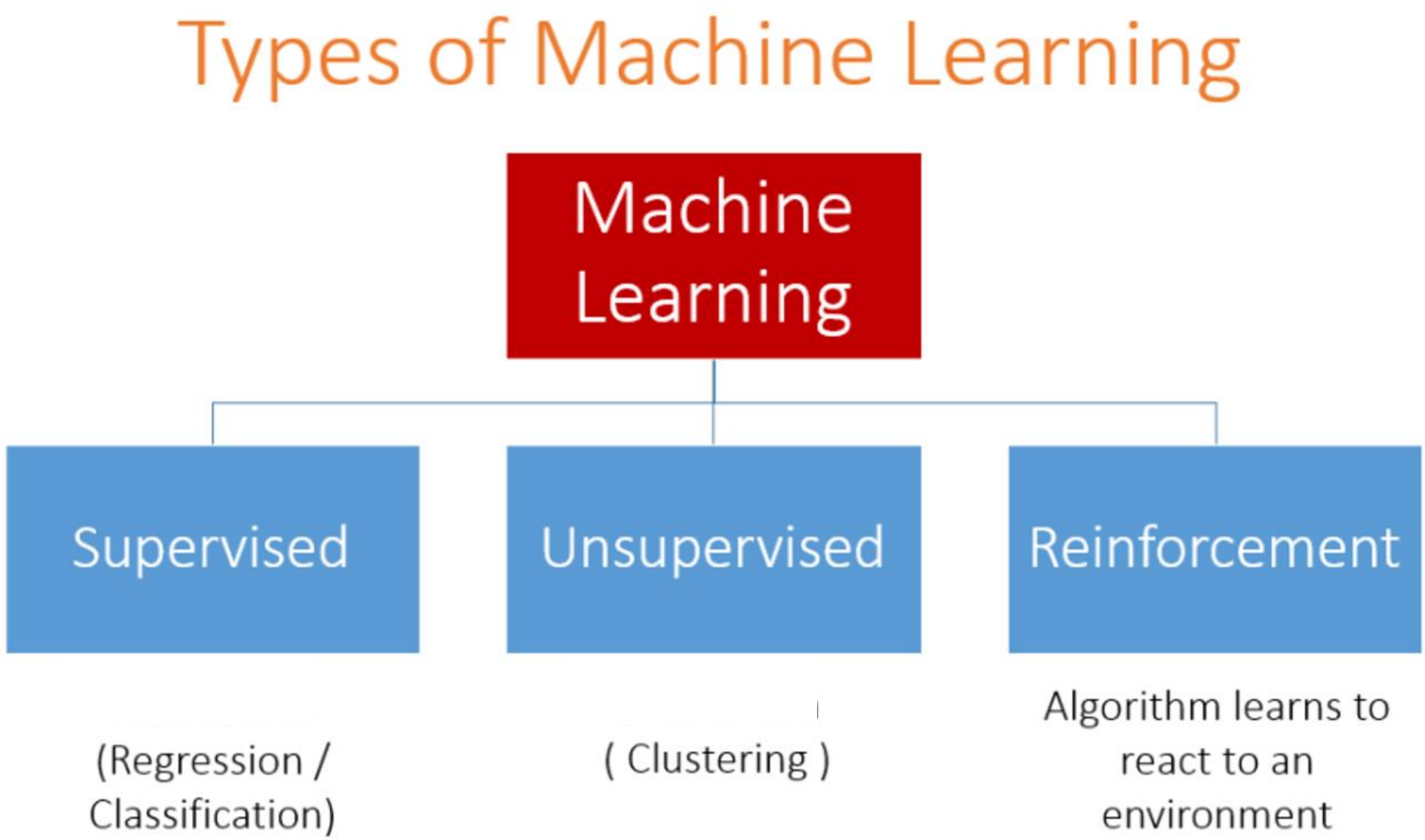
- Transformer(https://github.com/huggingface/transformers)
- 8页报告（包括Transformer原理和代码理解）



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

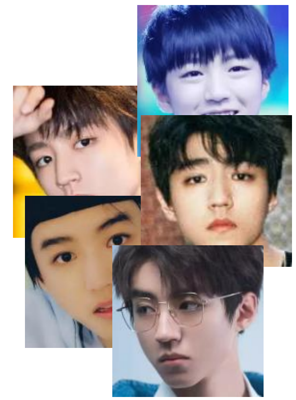


Supervised Learning

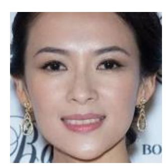
- Supervised learning
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y
  - Training data: a labeled set of input-output pairs

  - 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 ……



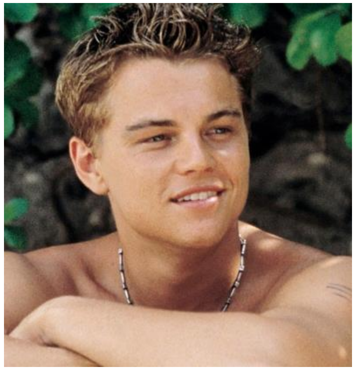
章子怡

Supervised Learning (Classification)

同一个人

不同人

同一个人



Supervised Learning (Regression)

30岁

28岁

18岁

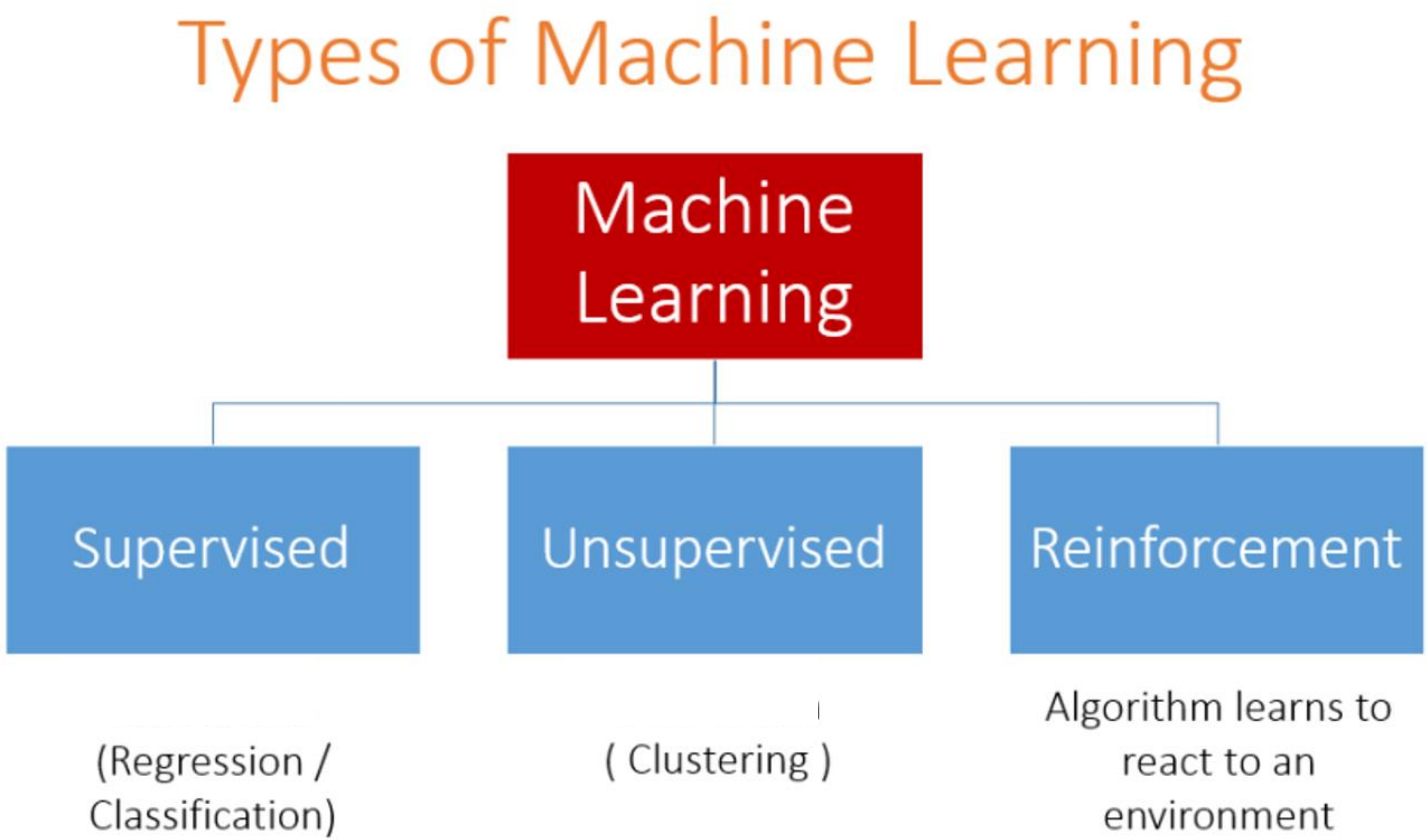
14岁

57岁

… …

33岁

Types of Machine Learning



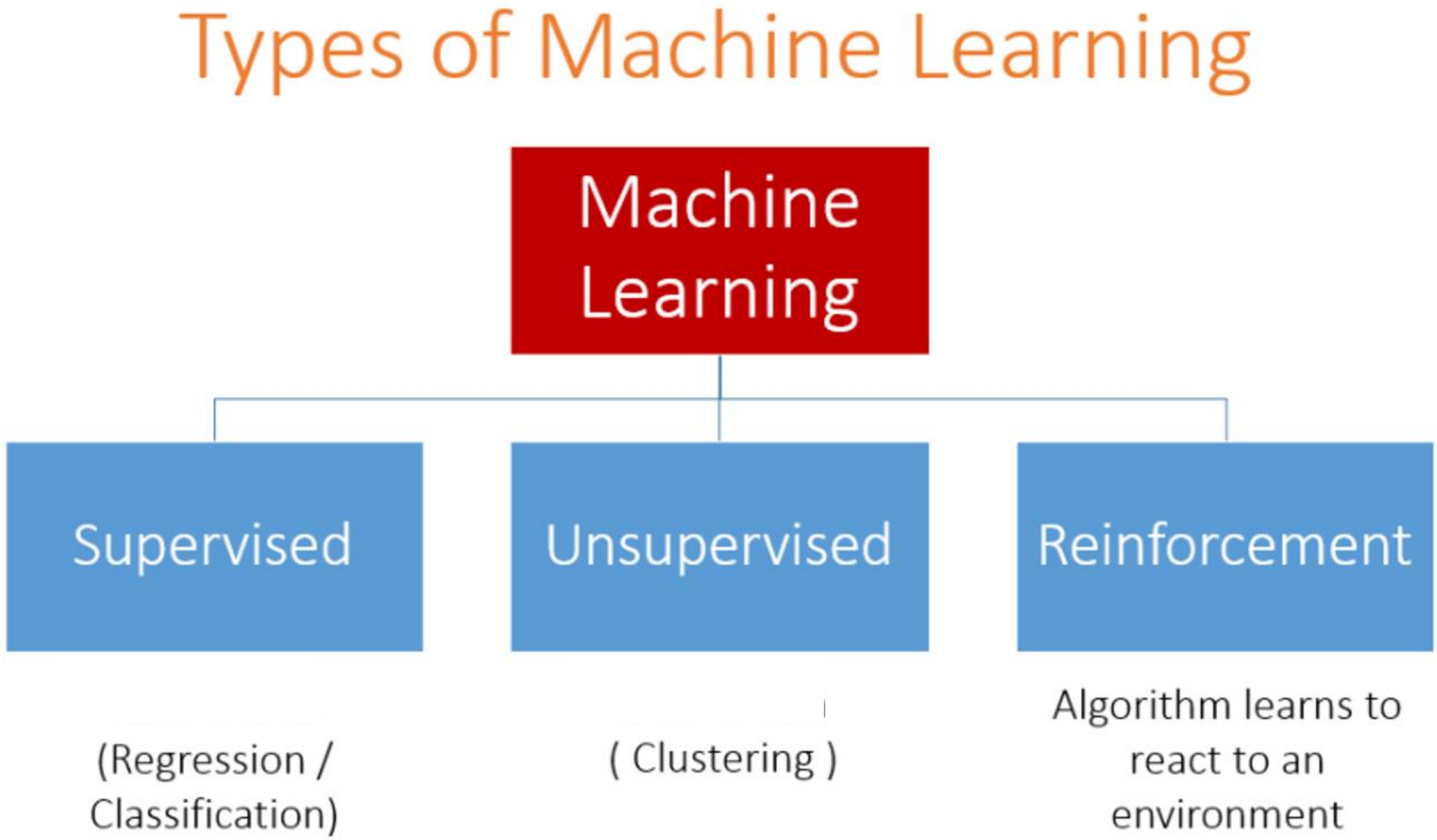
Unsupervised Learning

- 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

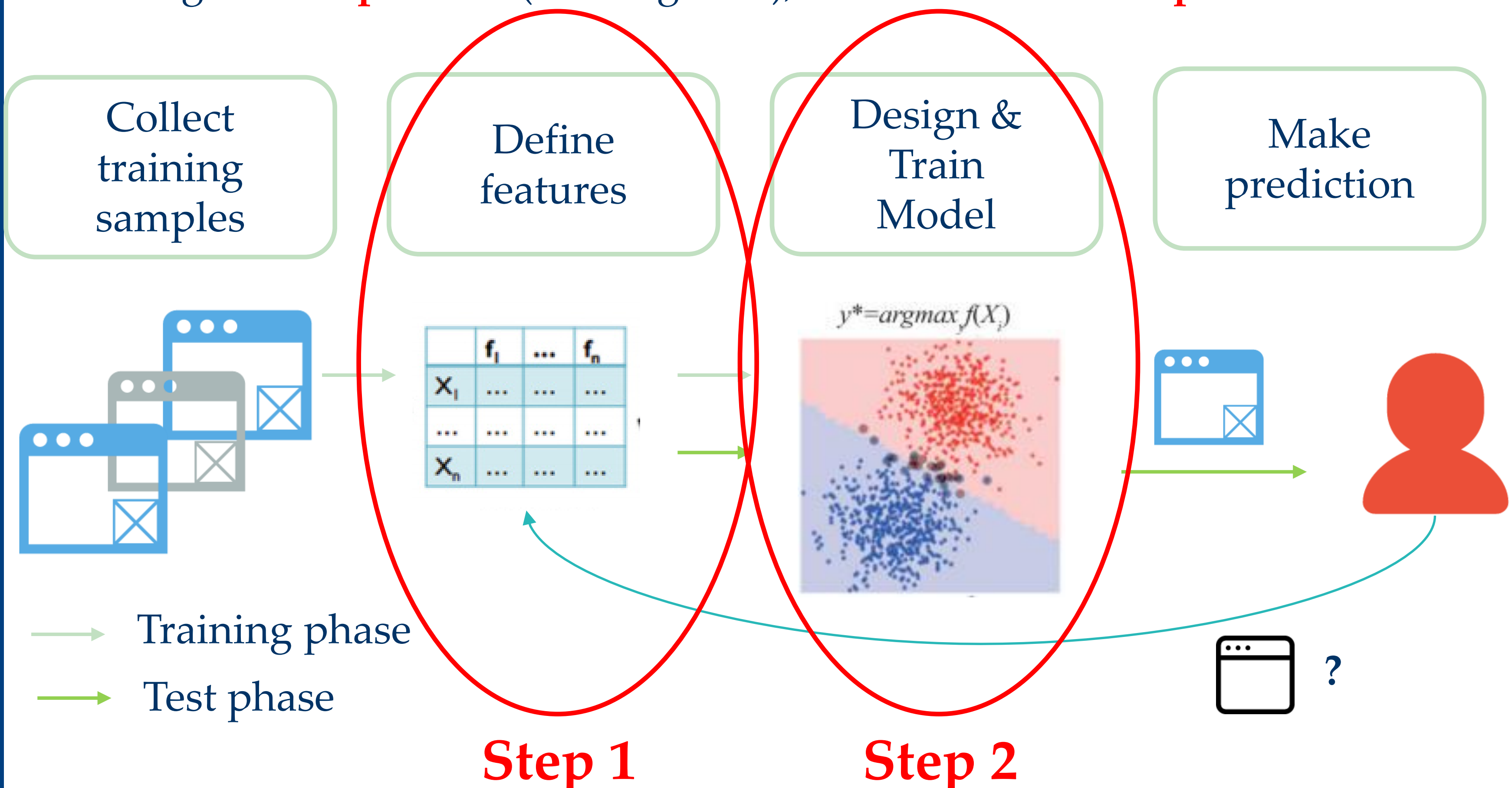
- Sample, example, pattern

- 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



Representation Learning

Supervised Learning

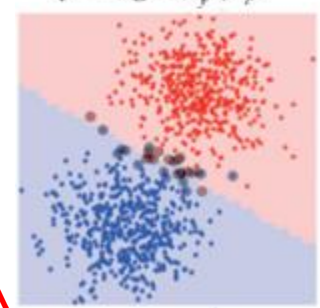
Define features

	f_1	...	f_n
X_1	...	...	...
...	...	...	...
X_n	...	...	...

Step 1

Design & Train Model

$y^* = \operatorname{argmax}_j f(X_i)$



Step 2

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

Why general classification hard?

Define features

	f_1	...	f_n
X_1	...	...	...
...	...	...	...
X_n	...	...	...

Step 1 is not good enough

- Intra-class variability

The letter "T" in different typefaces

Same face under different expression, pose, illumination

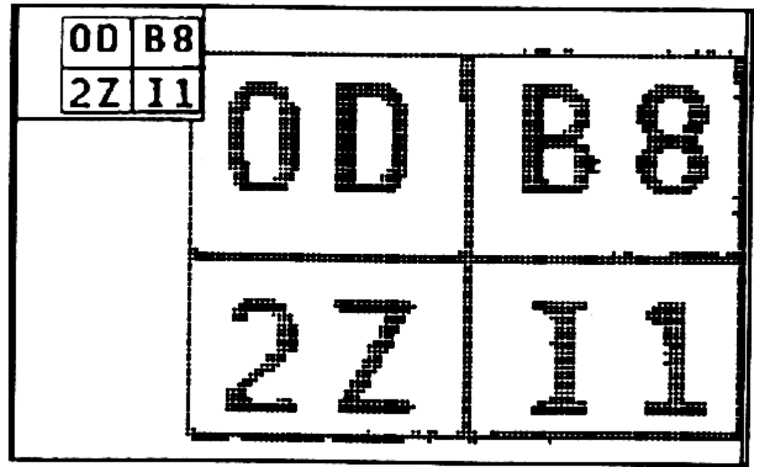
Why general classification hard?

Define features

	f_1	...	f_n
X_1	...	...	...
...	...	...	...
X_n	...	...	...

Step 1 is not good enough

- Inter-class similarity



Representation: Features

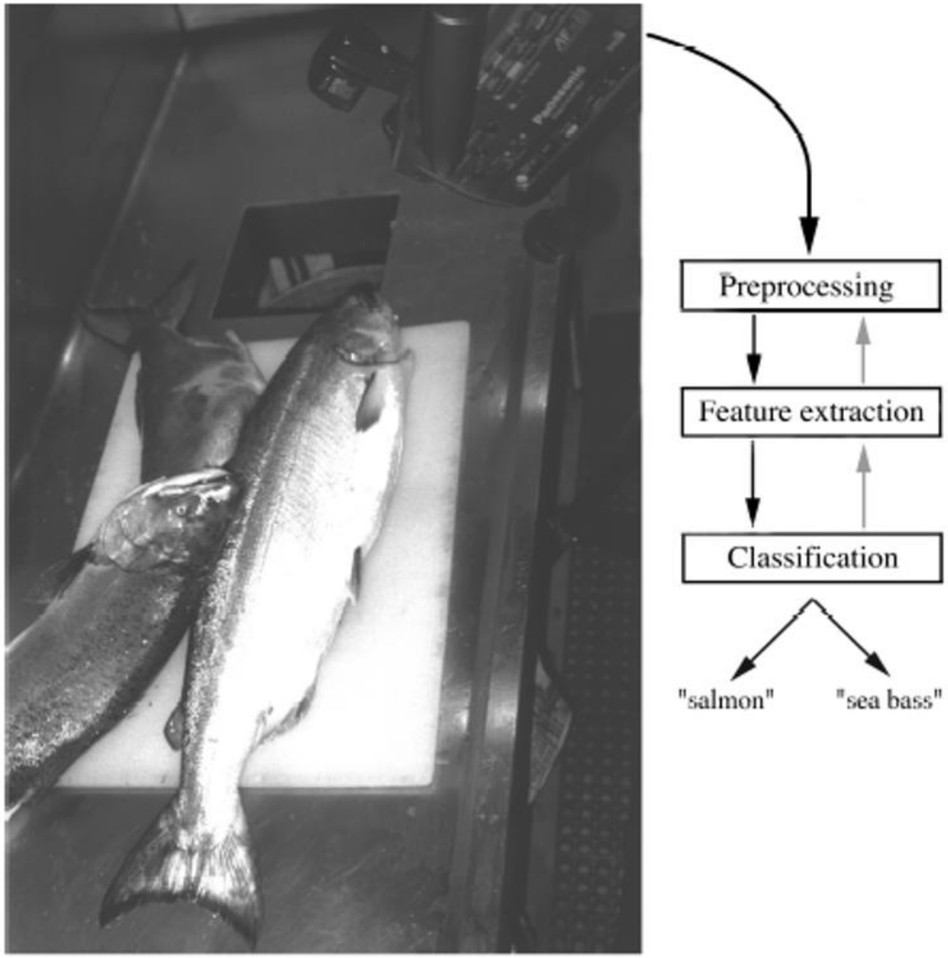
- Extract features to represent the samples
- Feature vector
- Good representation:
  - Low intra-class variability
  - Low inter-class similarity

Fish Classification: Salmon v. Sea Bass

Preprocessing involves image enhancement and segmentation;

(i) separate touching or occluding fishes and

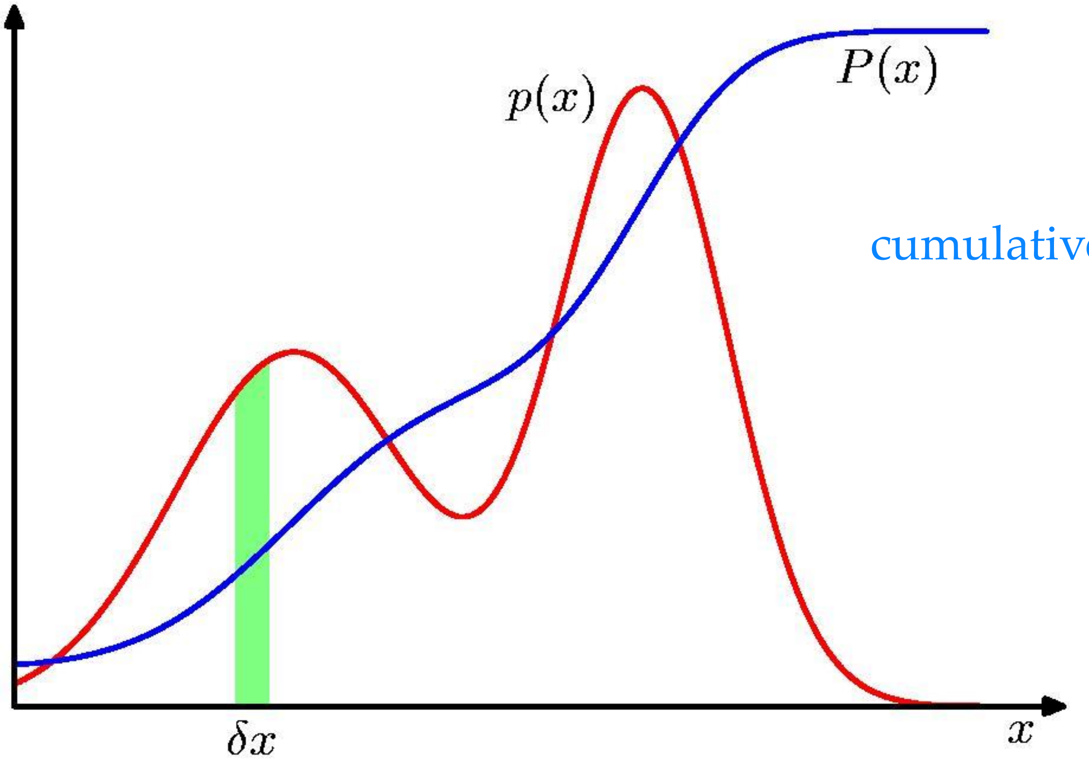
(ii) extract fish contour



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)

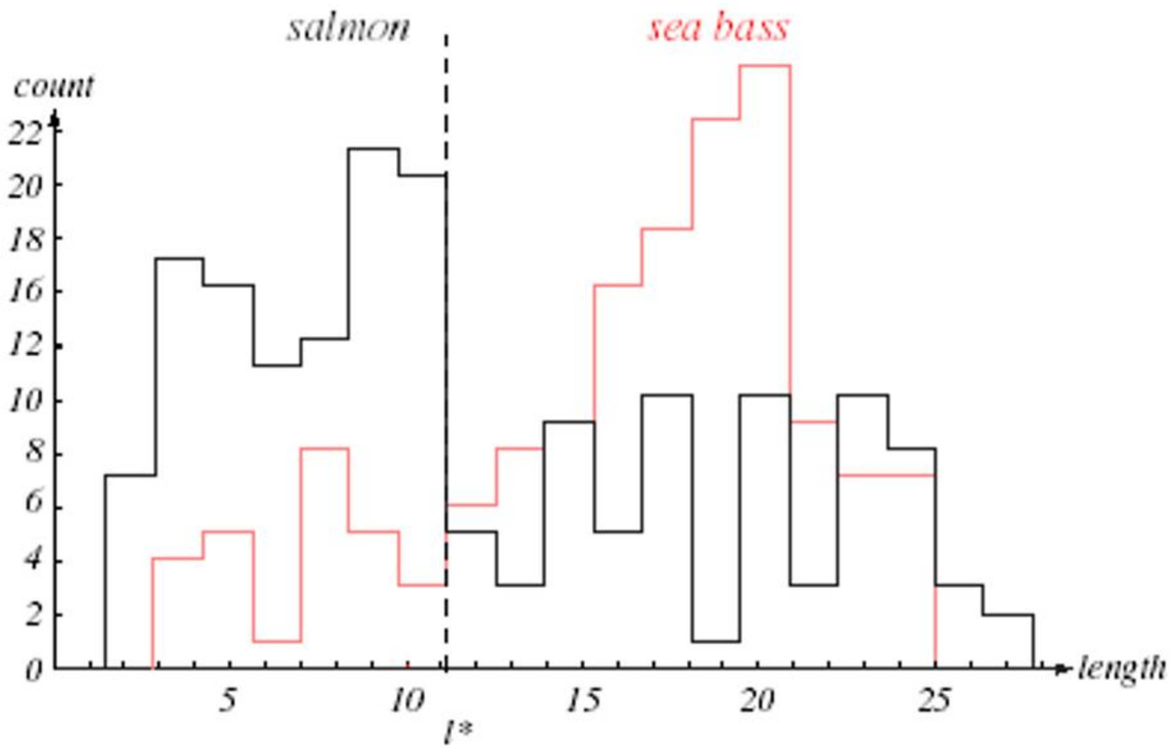
probability density function (PDF)

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

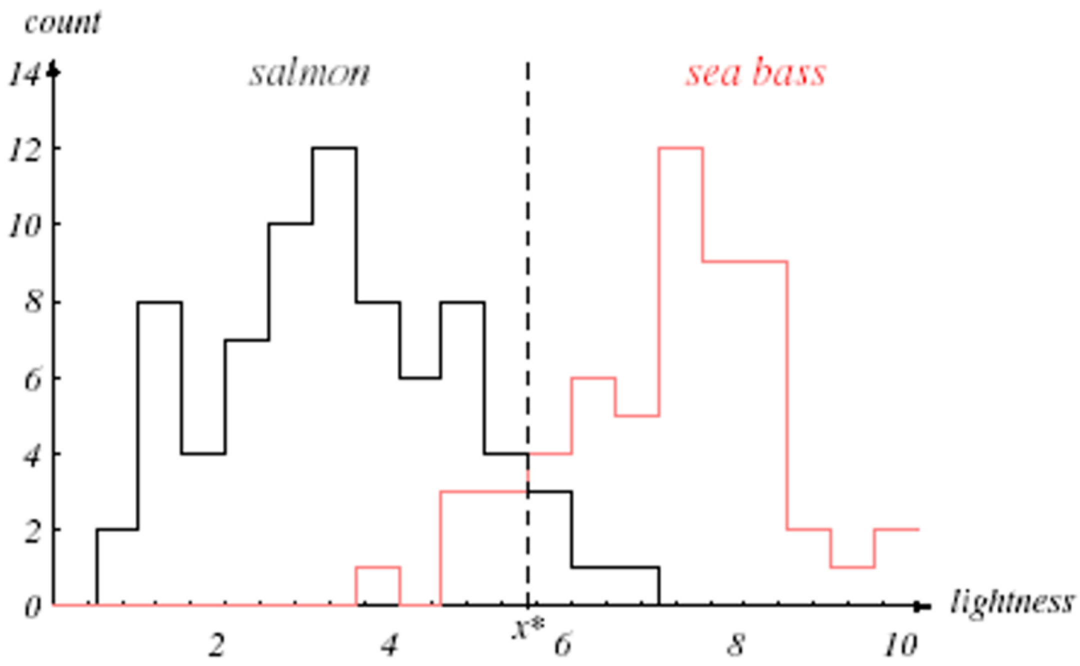
Representation: Fish Length As Feature

Training (design or learning) Samples

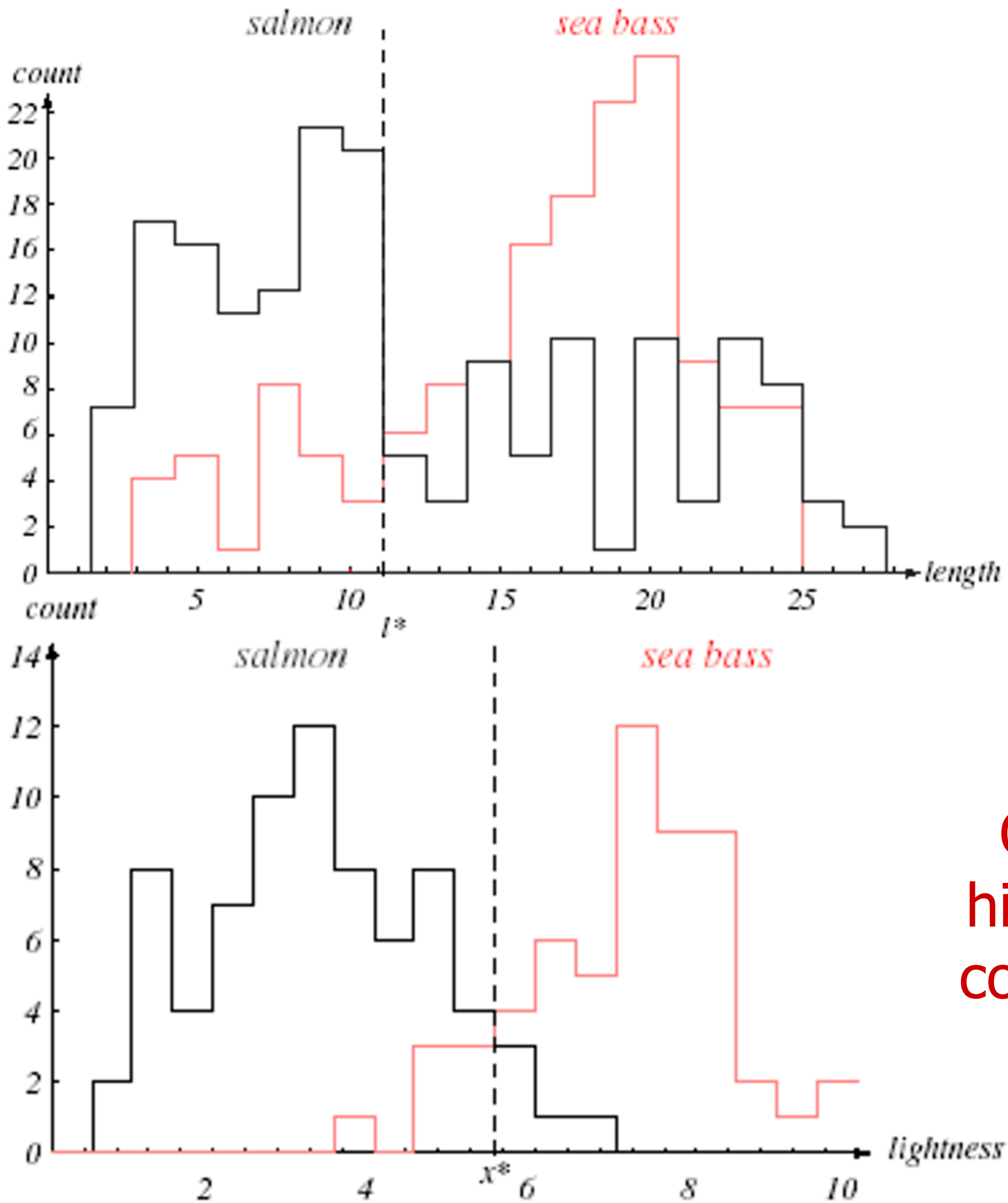
$p(x|\text{salmon})$ $p(x|\text{bass})$



Fish Lightness As Feature



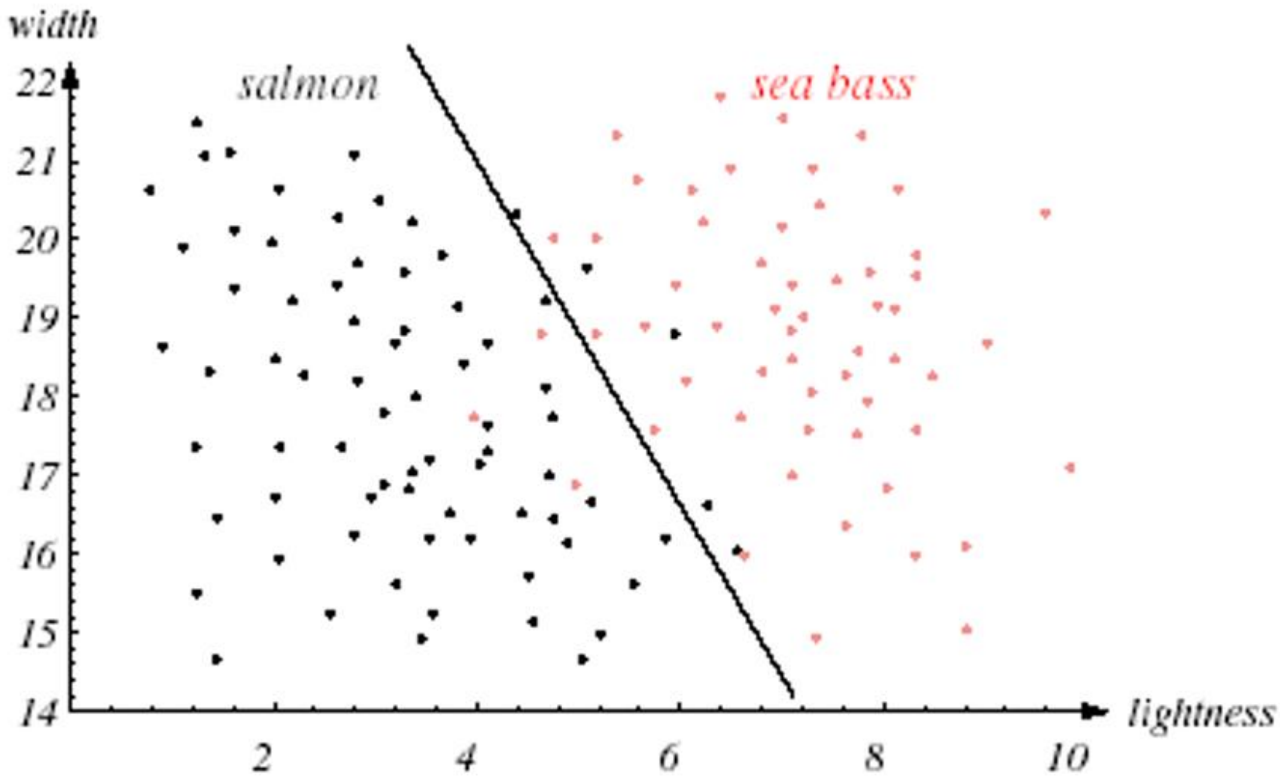
Which Feature is better



Overlap of these histograms is small compared to length feature

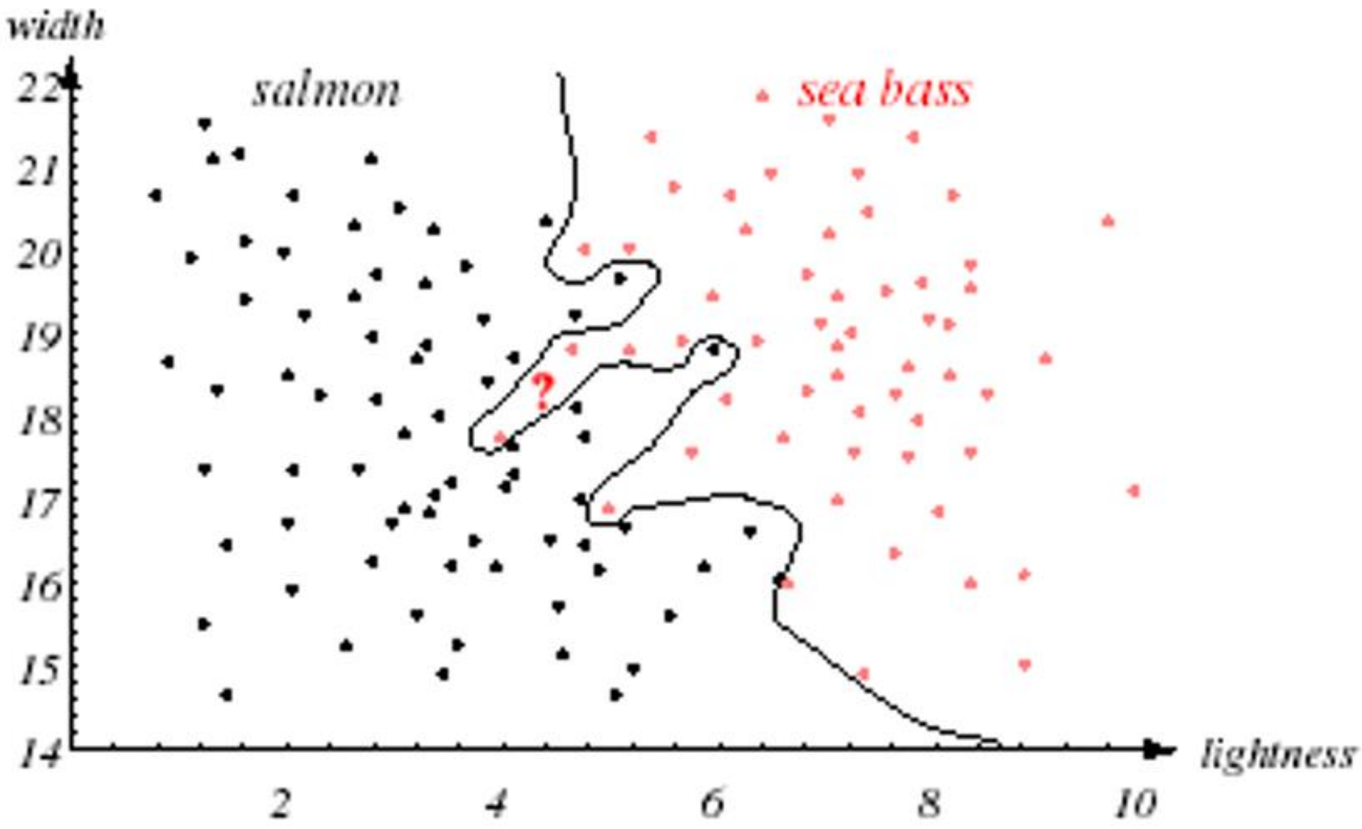
Two-dimensional Feature Space

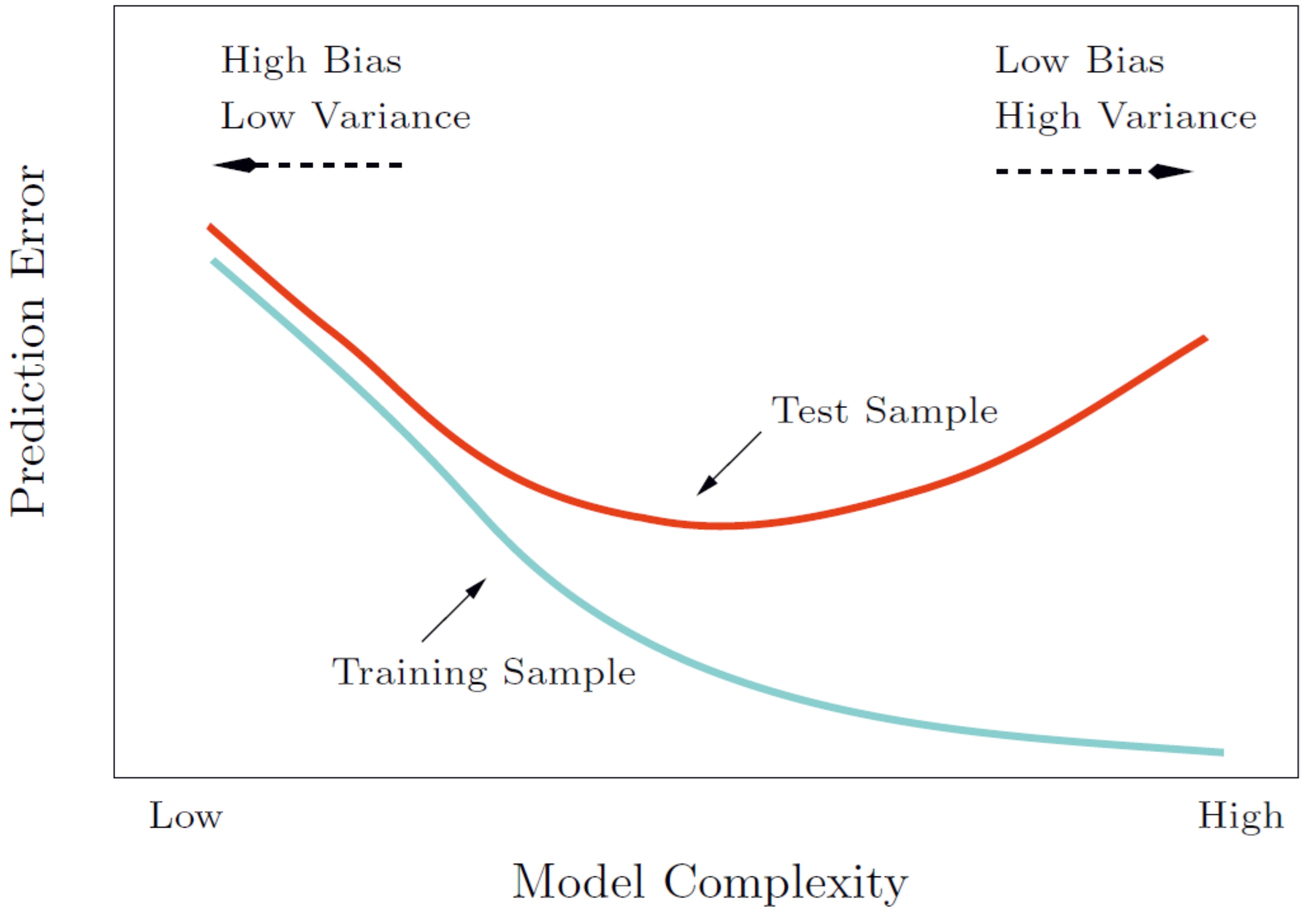
Linear (simple) decision boundary



Two features together are better than individual features

Complex Decision Boundary




High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity

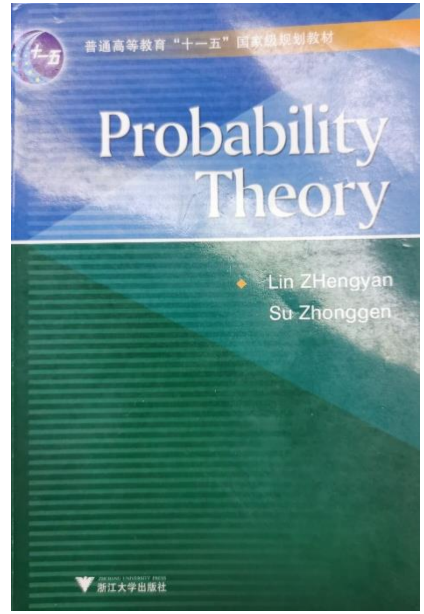
Generalization

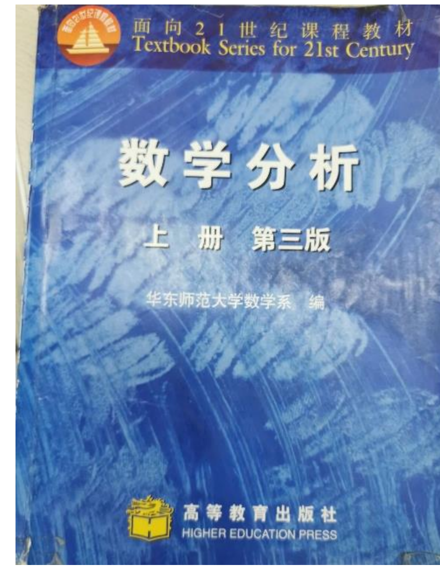
- A generalization of a concept is an extension of the concept to less-specific criteria.
- Generalization of the classifier (model)
  - The performance of the classifier on test data.

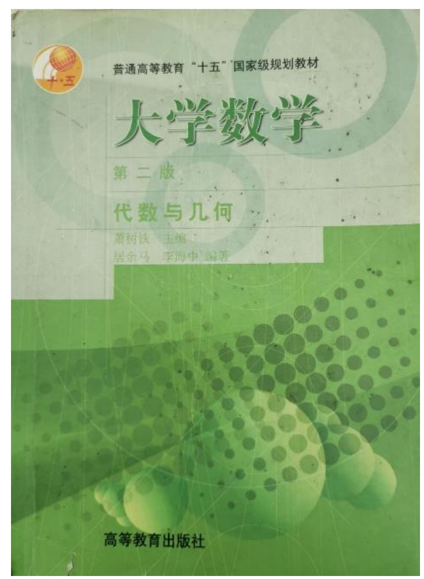
- Training error:
- Simple model → large training error
- Complex model → less training error
- Test error:
- Simple model → ?
- Complex model → ?

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