统计学习方法概论

Peng Li

https://simplelp.github.io/

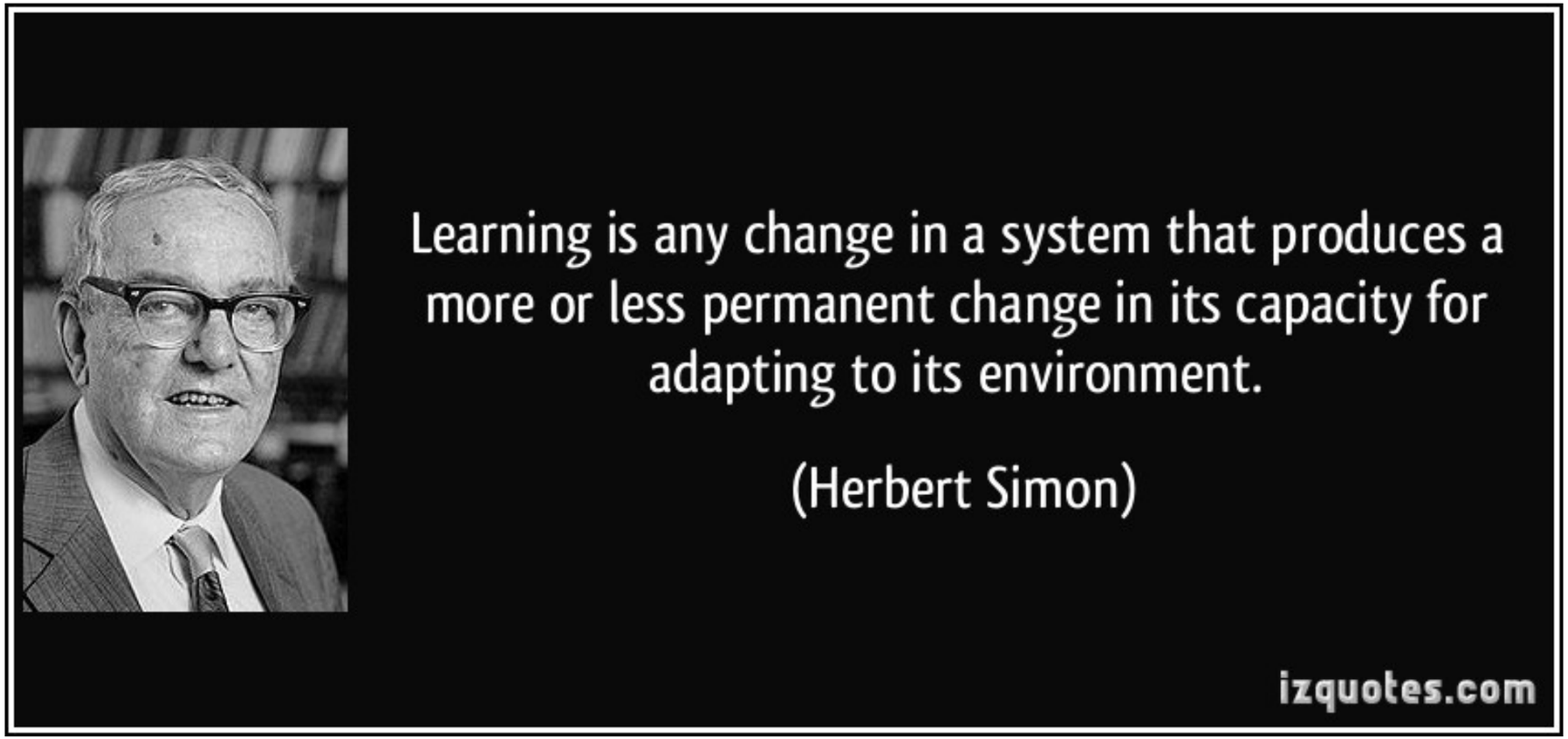
2019/05/30

一、统计学习概述

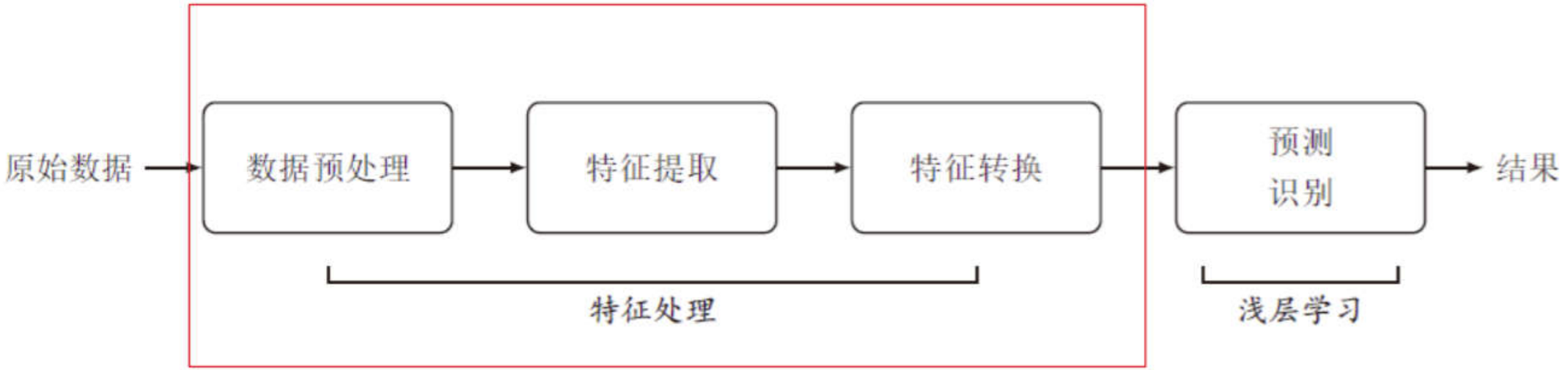
统计机器学习概念

- 统计（机器）学习是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测和分析的一门学科

统计机器学习

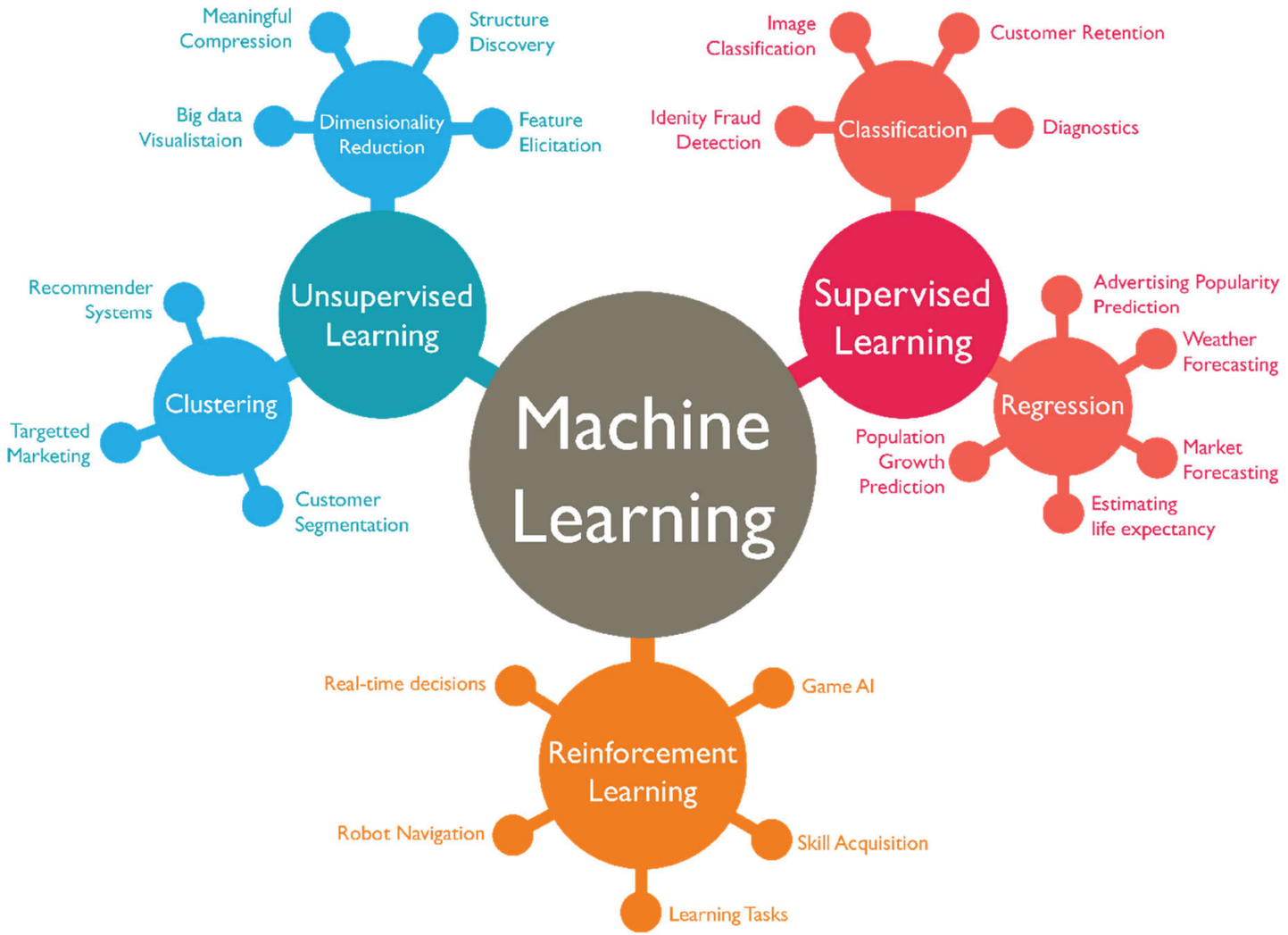

Learning is any change in a system that produces a more or less permanent change in its capacity for adapting to its environment.
(Herbert Simon)
izquotes.com

经典机器学习流程



https://nndl.github.io/

机器学习方法类别


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

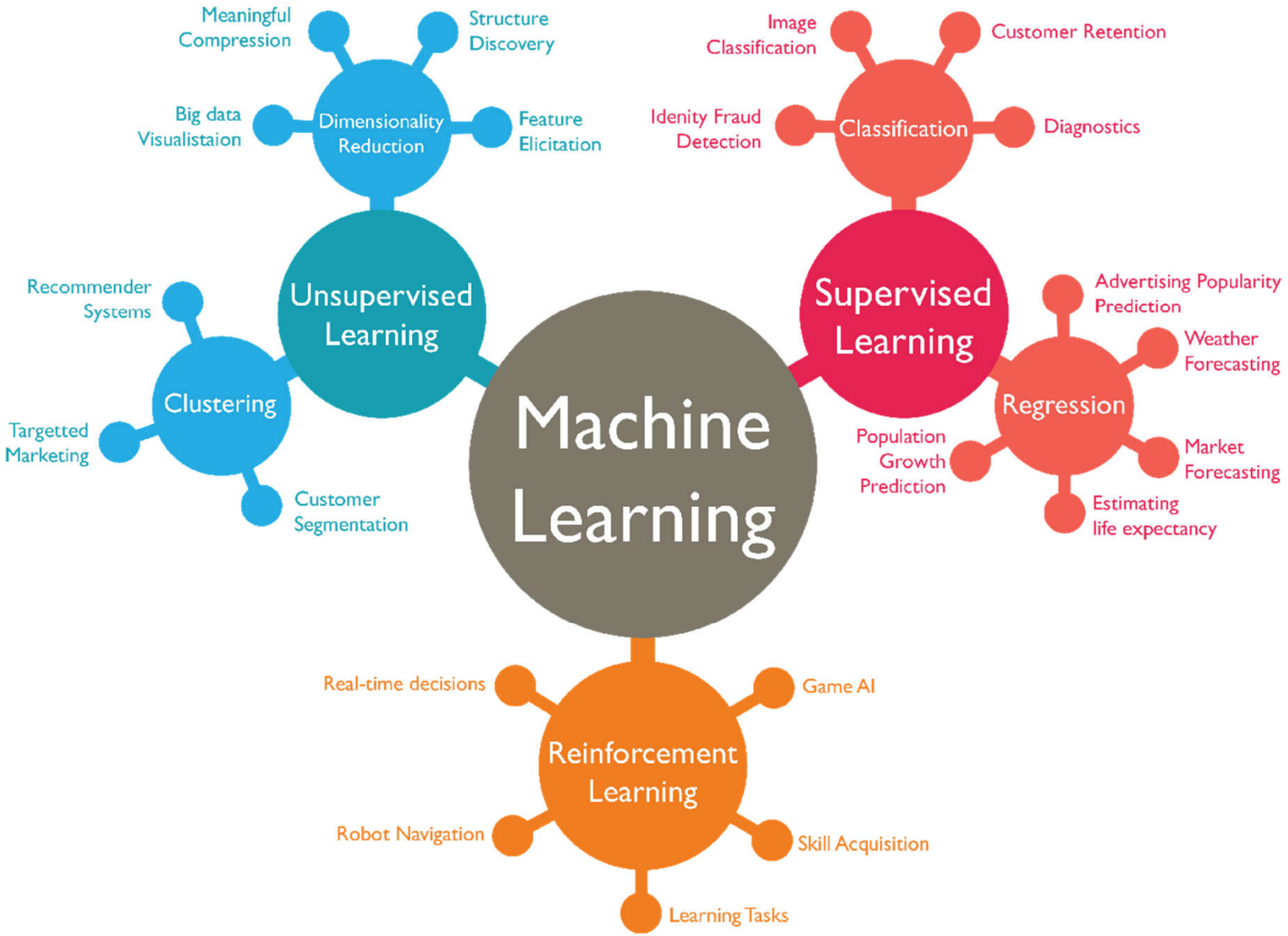
统计学习方法三要素

统计学习的研究

- 统计学习方法
- 统计学习理论
- 统计学习应用

二、统计学习分类

基本分类


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

监督学习

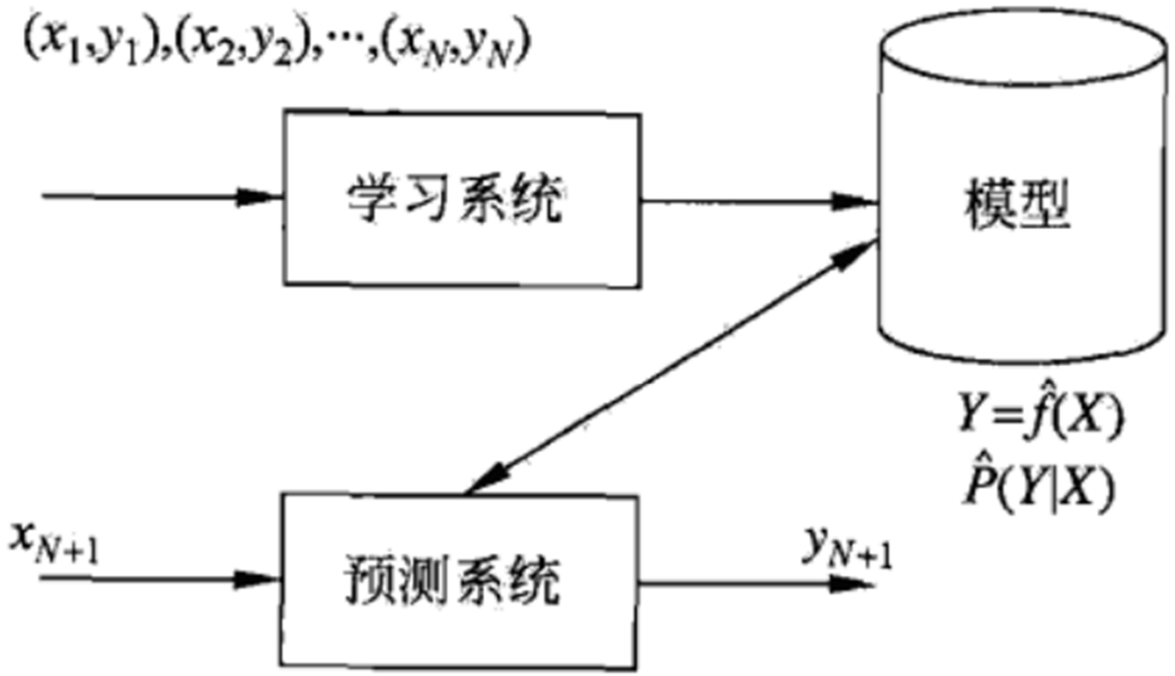


图 1.1　监督学习问题

非监督学习

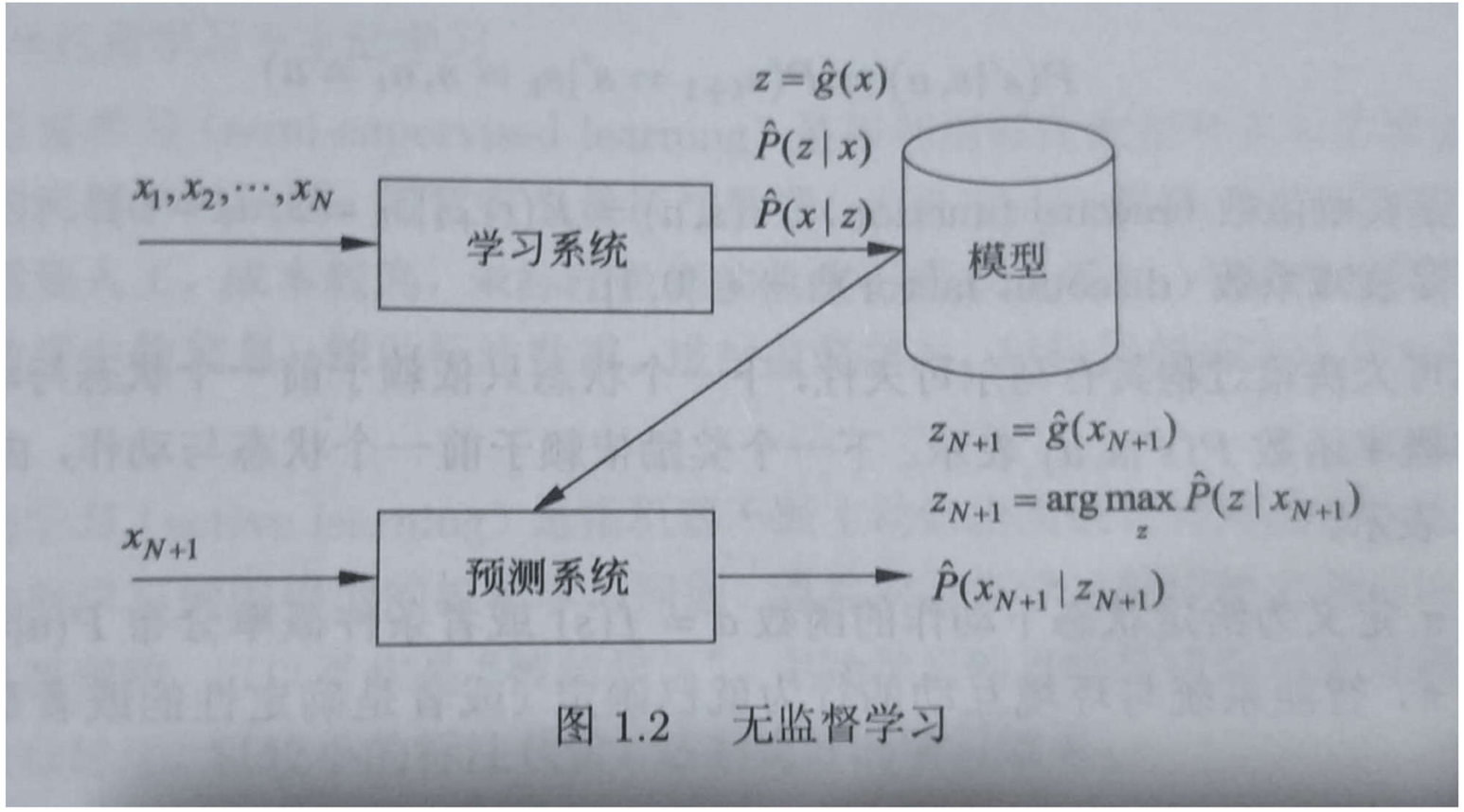
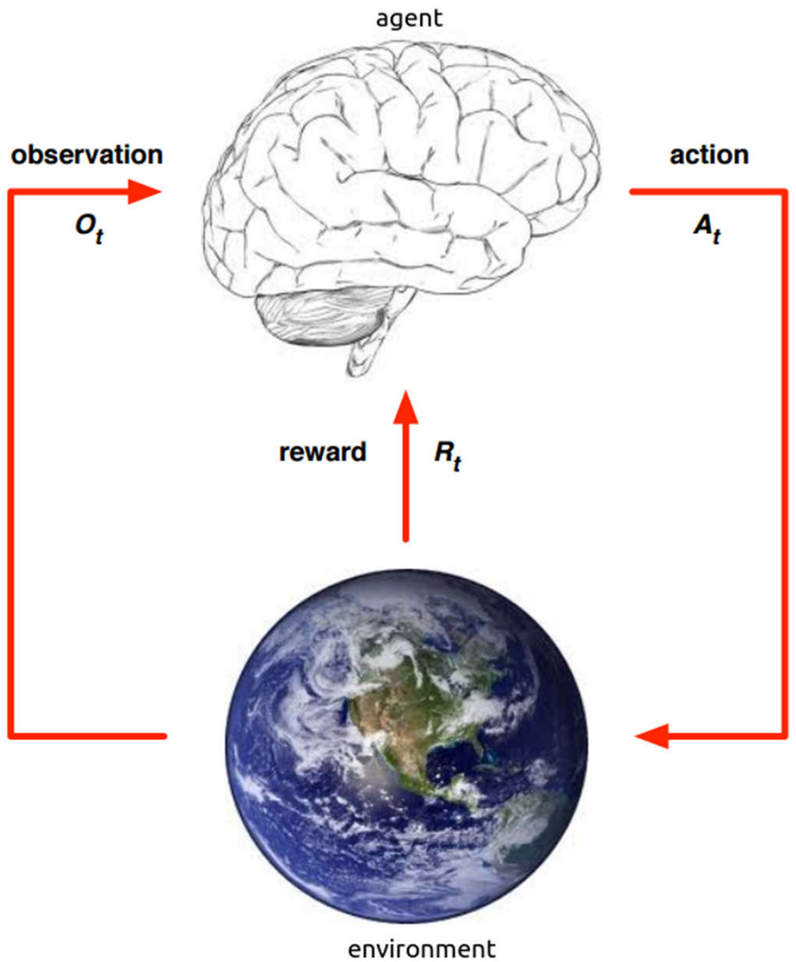


图 1.2　无监督学习

强化学习



http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Teaching.html

其他分类方法

- 按模型分（概率/非概率、线性/非线性、参数/非参数等）

其他分类方法

- 按模型分（概率/非概率、线性/非线性、参数/非参数等）

其他分类方法

- 按模型分（概率/非概率、线性/非线性、参数/非参数等）
- 按算法分（在线/批量）

其他分类方法

- 按模型分（概率/非概率、线性/非线性、参数/非参数等）
- 按算法分（在线/批量）
- 按技巧分（贝叶斯、核方法等）

三、统计学习方法三要素

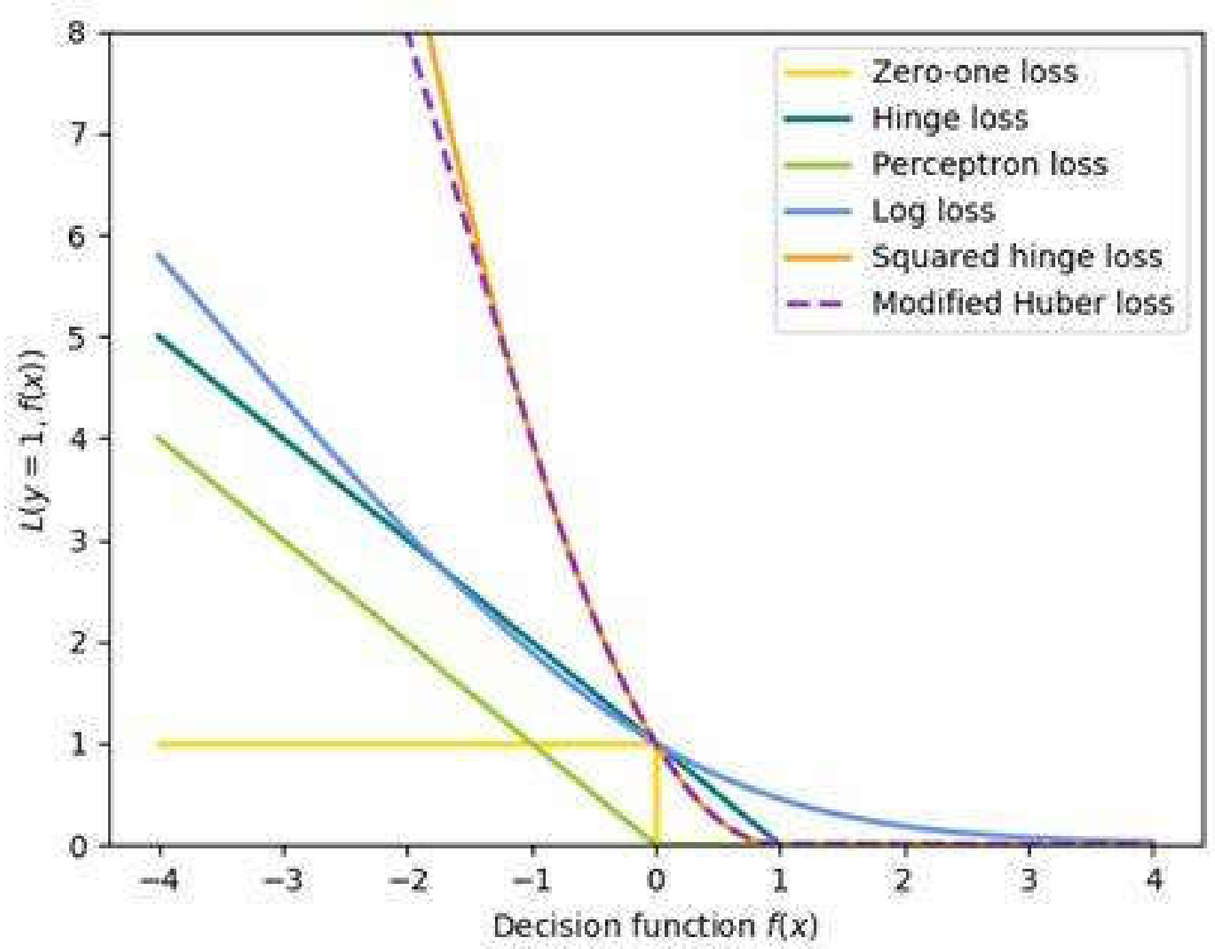
模型

$$\mathcal{F} = \{f \mid Y = f_\theta(X), \theta \in \mathbf{R}^n\}$$

$$\mathcal{F} = \{P \mid P_\theta(Y \mid X), \theta \in \mathbf{R}^n\}$$

策略

- 损失函数


Zero-one loss
Hinge loss
Perceptron loss
Log loss
Squared hinge loss
Modified Huber loss
$L(y=1, f(x))$
Decision function $f(x)$
8
7
6
5
4
3
2
1
0
−4
−3
−2
−1
0
1
2
3
4

策略

- 风险函数（期望损失）

$$R_{\exp}(f) = E_P[L(Y, f(X))] = \int_{x \times y} L(y, f(x)) P(x, y) \mathrm{d}x\mathrm{d}y$$

策略

- 经验风险（经验损失）

$$R_{\text{emp}}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$$

策略

- 经验风险最小化（ERM）

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

策略

- 结构风险

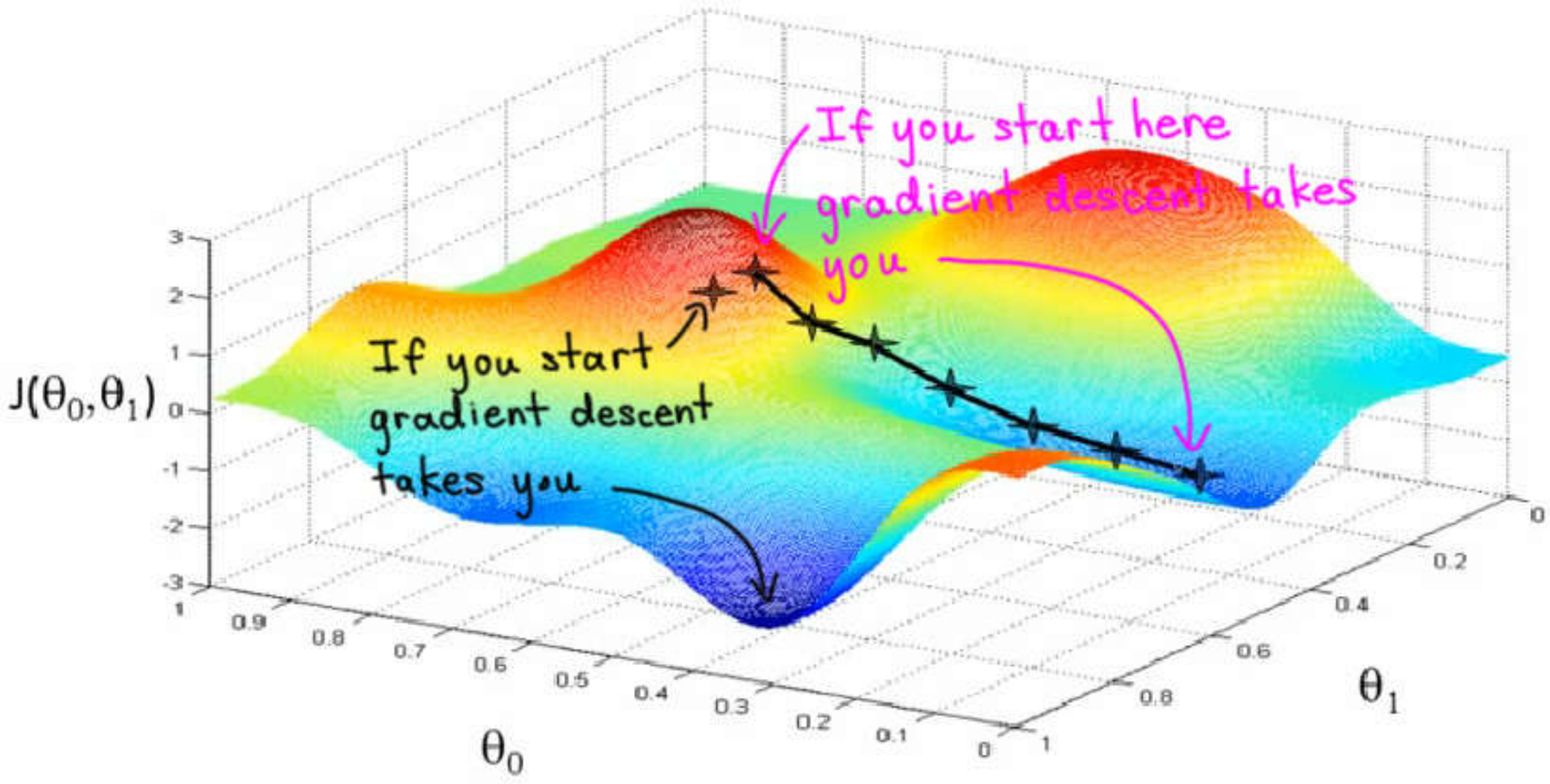
$$R_{\text{srm}}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

策略

- 结构风险最小化

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

算法


If you start here gradient descent takes you
If you start gradient descent takes you
$J(\theta_0,\theta_1)$
θ_0
θ_1

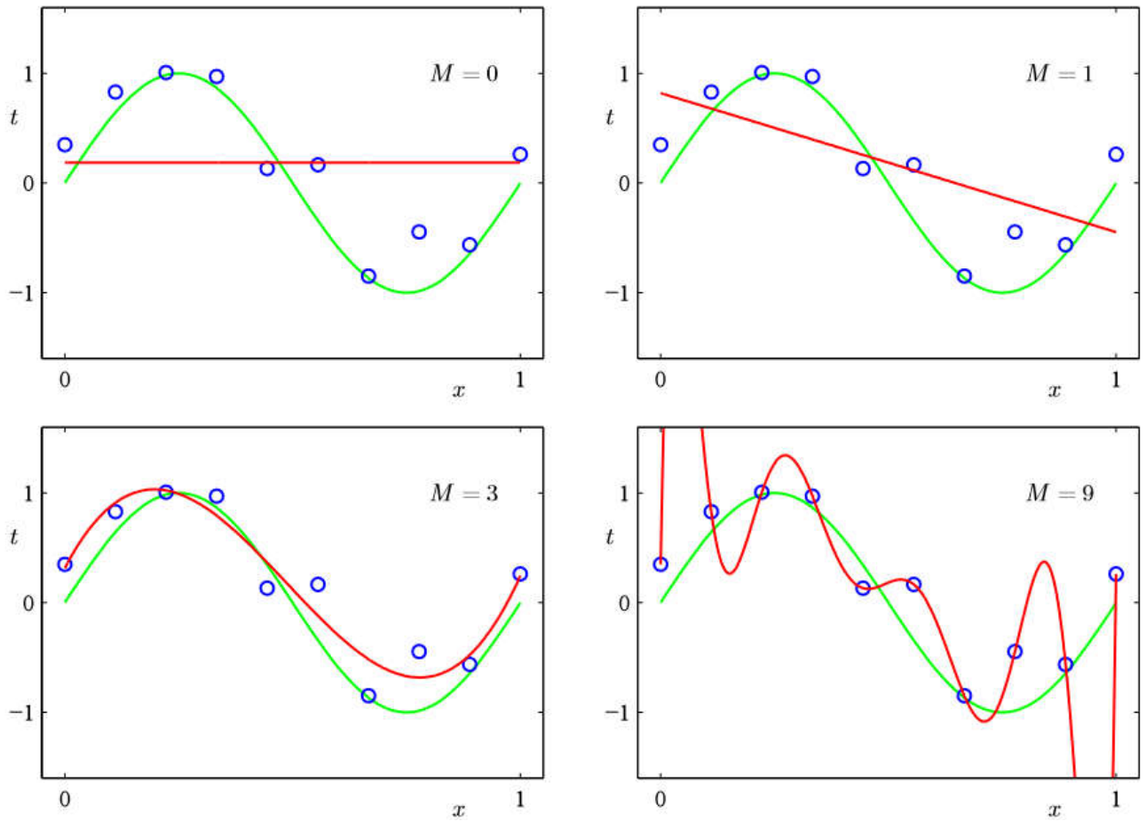
四、模型评估与模型选择

训练误差与测试误差

$$R_{\text{emp}}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

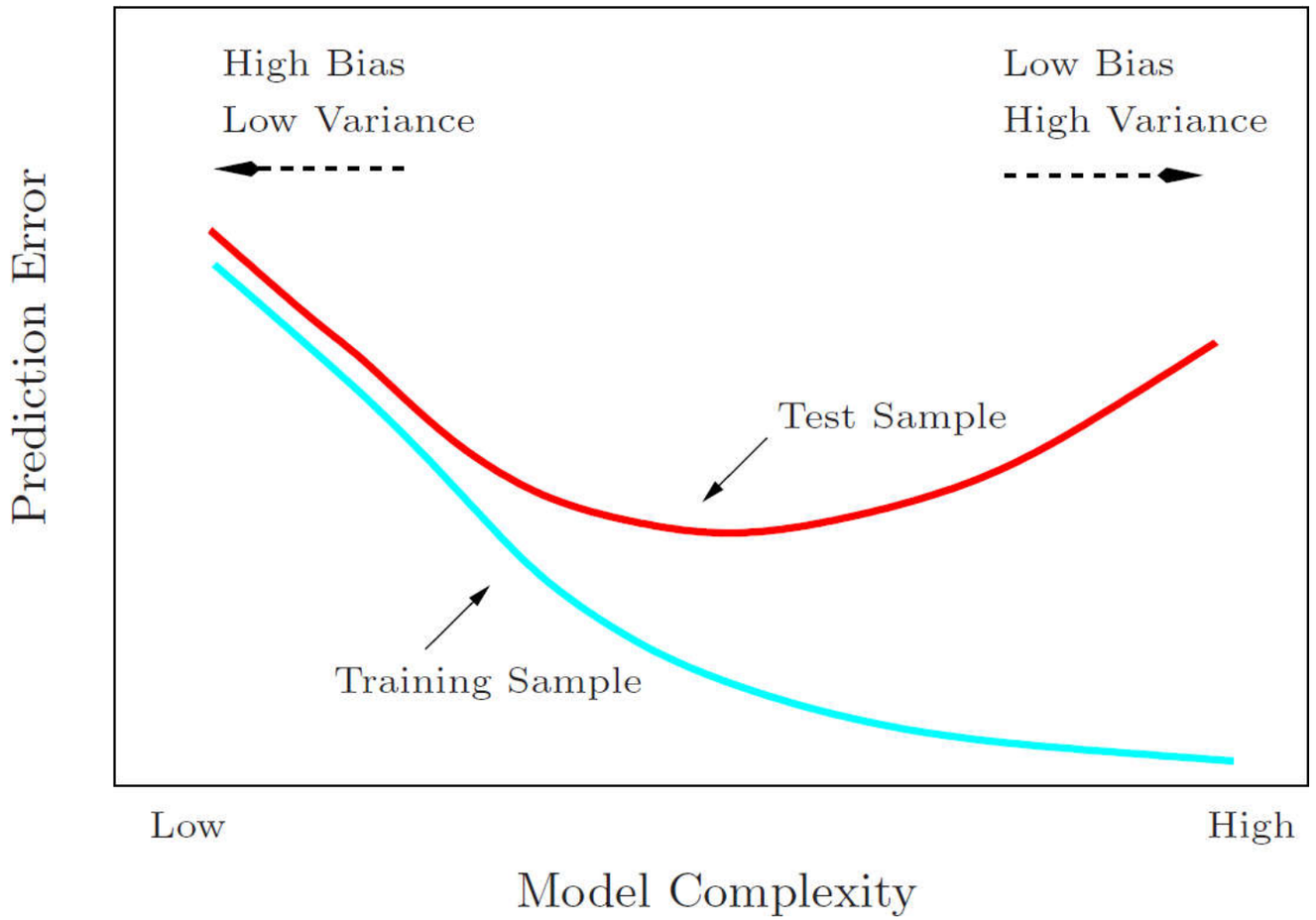
$$e_{\text{test}} = \frac{1}{N'}\sum_{i=1}^{N'} L(y_i, \hat{f}(x_i))$$

过拟合



$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

过拟合


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

模型选择方法

- 正则化
- 交叉验证

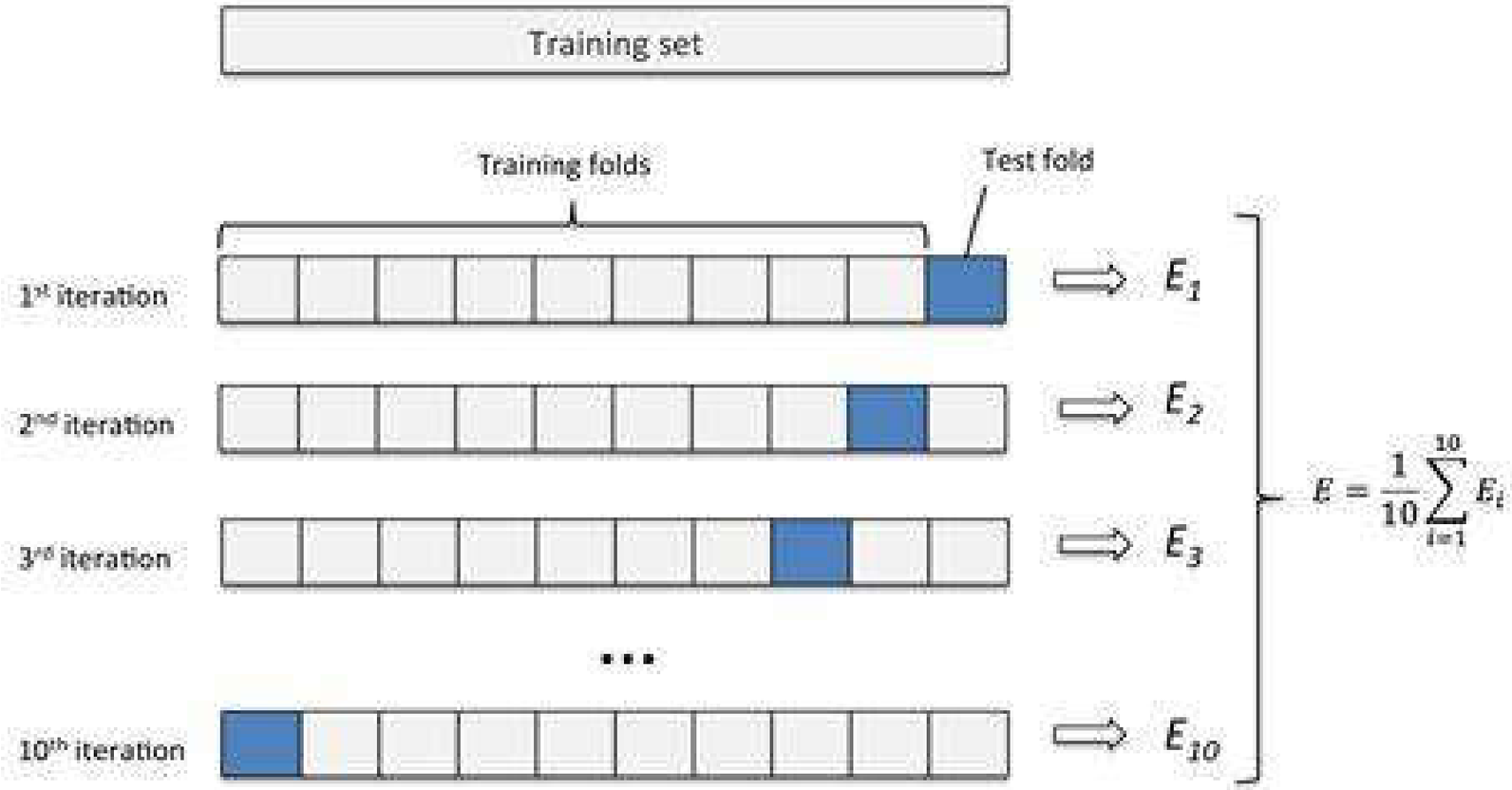
正则化

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \frac{\lambda}{2} \| w \|^2$$

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \lambda \| w \|_1$$

交叉验证


Training set
Training folds
Test fold
1st iteration
2nd iteration
3rd iteration
10th iteration
E_1
E_2
E_3
E_{10}
$E = \frac{1}{10}\sum_{i=1}^{10} E_i$

五、泛化能力

泛化误差

$$R_{exp}(\hat{f}) = E_P[L(Y, \hat{f}(X))] = \int_{x \times y} L(y, \hat{f}(x)) P(x, y) \mathrm{d}x\mathrm{d}y$$

泛化误差上界

定理 1.1（泛化误差上界） 对二类分类问题，当假设空间是有限个函数的集合 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$ 时，对任意一个函数 $f\in\mathcal{F}$，至少以概率 $1-\delta$，以下不等式成立：

$$R(f)\leqslant\hat{R}(f)+\varepsilon(d,N,\delta) \tag{1.25}$$

其中，

$$\varepsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)} \tag{1.26}$$

六、监督学习

生成模型与判别模型

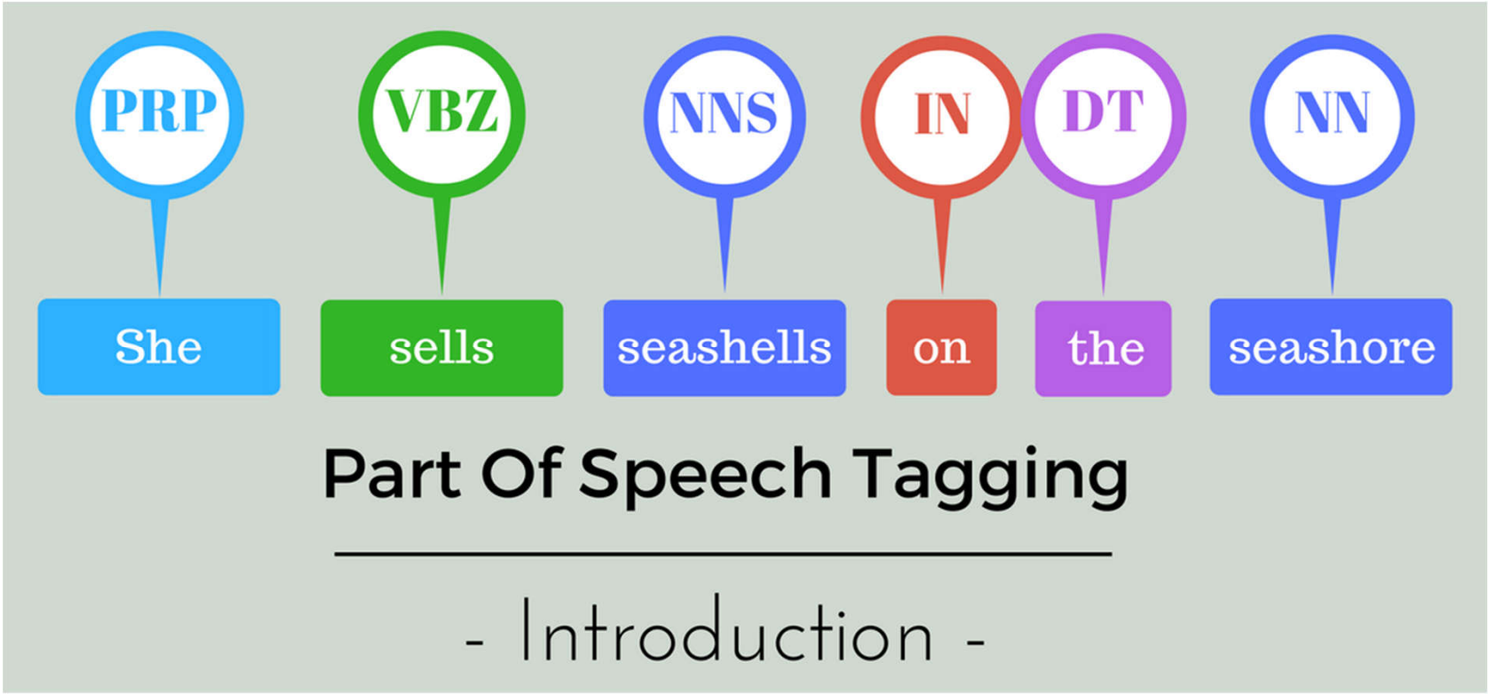
- 生成模型：学习P(X,Y)
  - 朴素贝叶斯、隐马尔科夫模型等
  - 收敛速度快，样本容量增加时，更快地收敛到真实模型
  - 存在隐变量时只能用生成模型
- 判别模型：学习 f(X) 或者 P(Y|X)
  - K近邻、感知机、决策树、逻辑回归、最大熵等等
  - 直接面对预测，往往准确率更高
  - 对数据进行抽象、定义并使用特征，可以简化学习问题

监督学习应用

- 分类

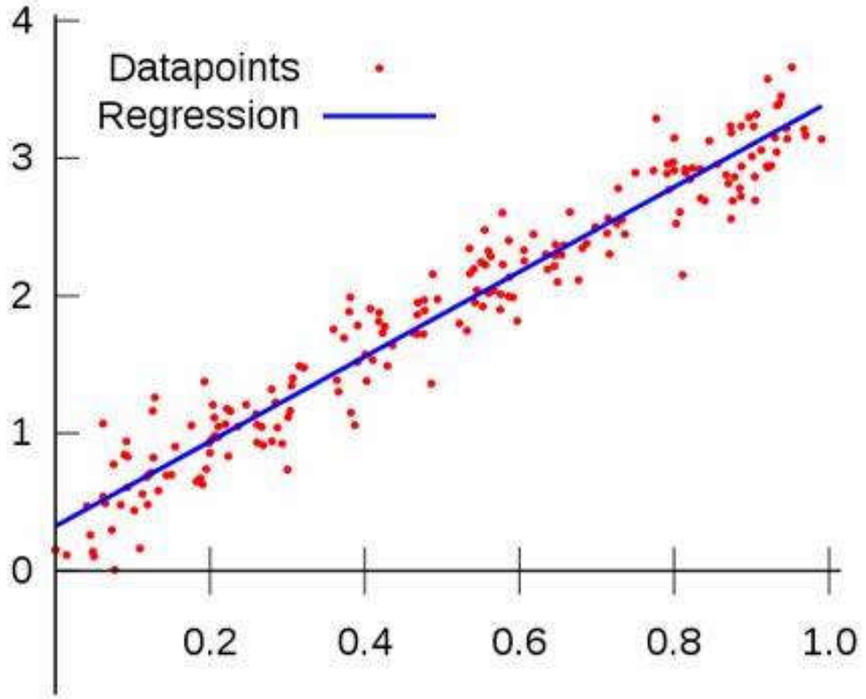
监督学习应用

- 标注


PRP
VBZ
NNS
IN
DT
NN
She
sells
seashells
on
the
seashore
Part Of Speech Tagging
- Introduction -

监督学习应用

- 回归

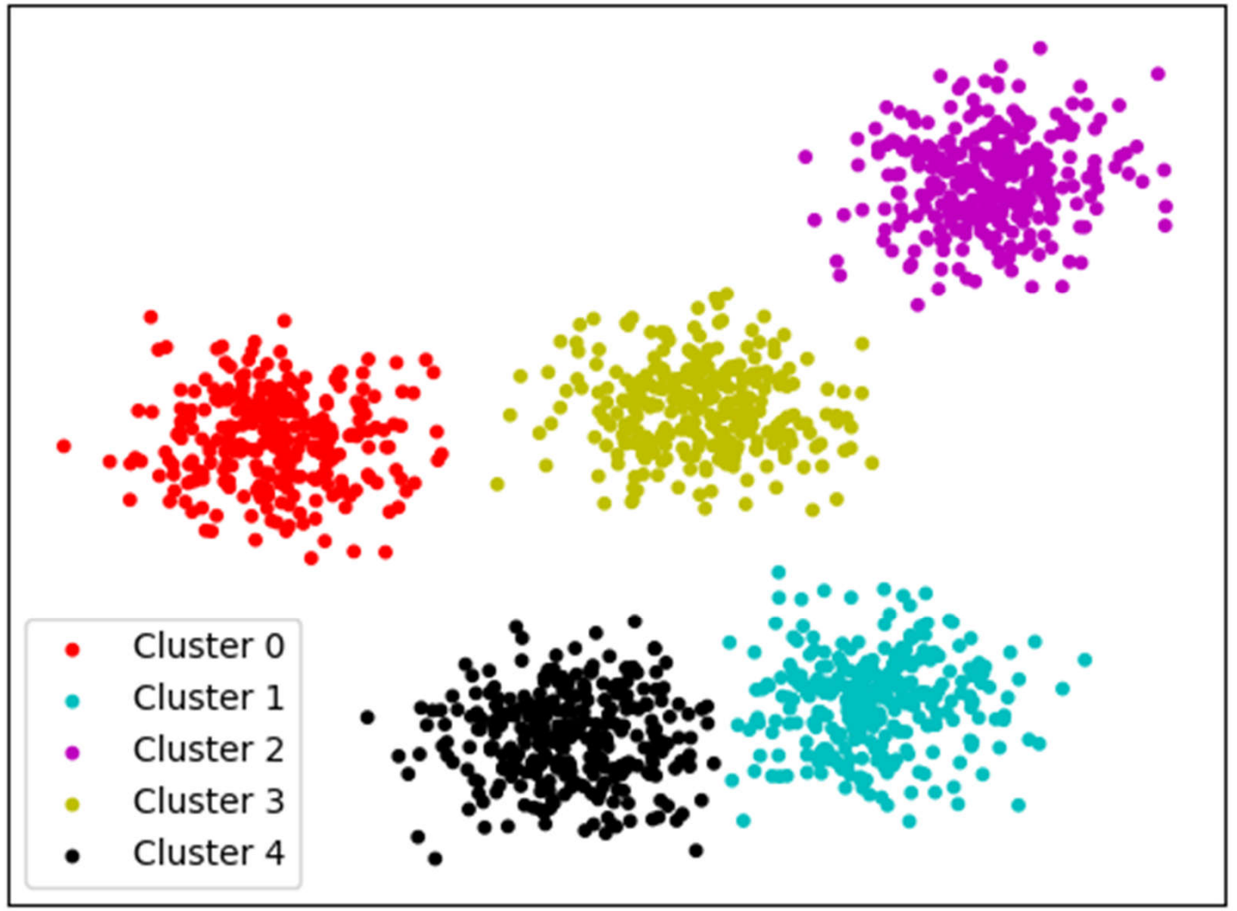


七、非监督模型

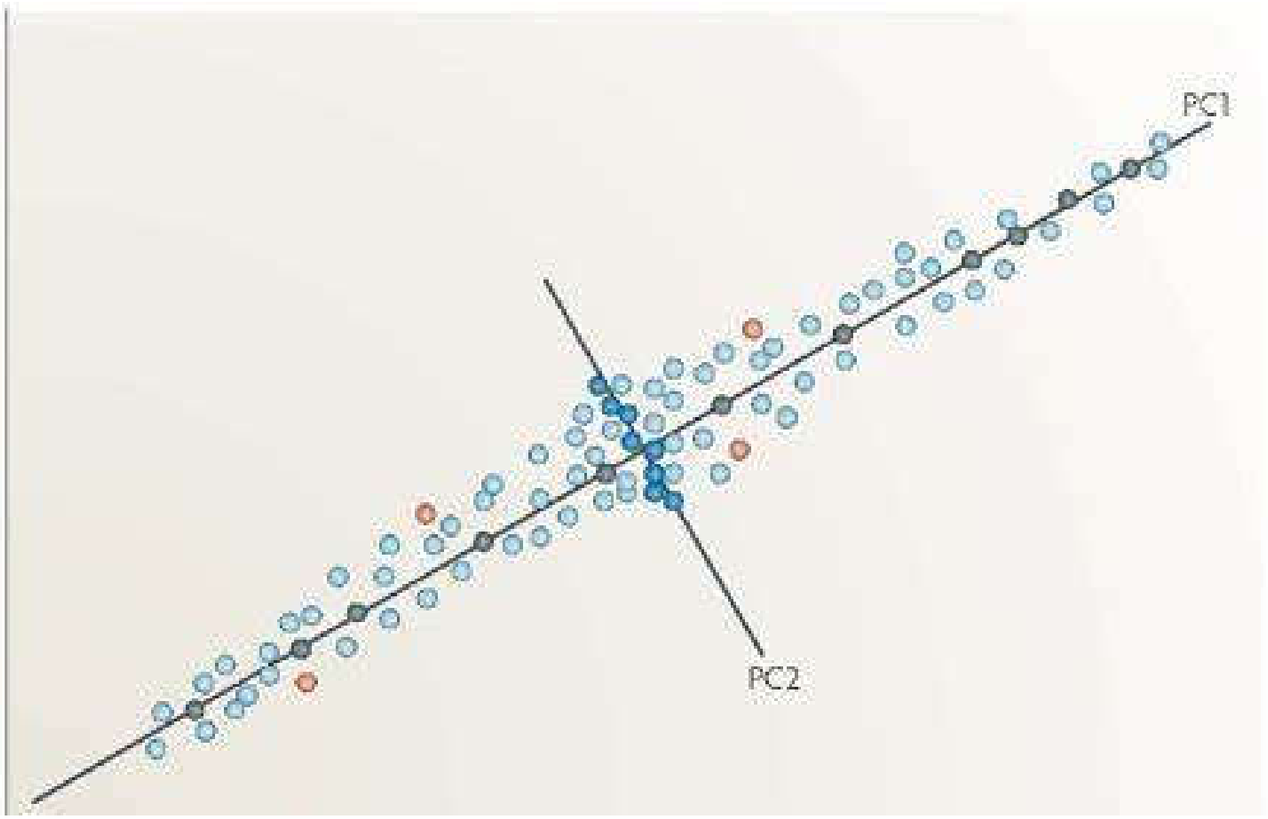
基本原理

- 从无标签数据中学习数据的统计规律或者内在结构
- 主要包括聚类、降维、概率估计
- 用于数据分析或者监督学习的预处理

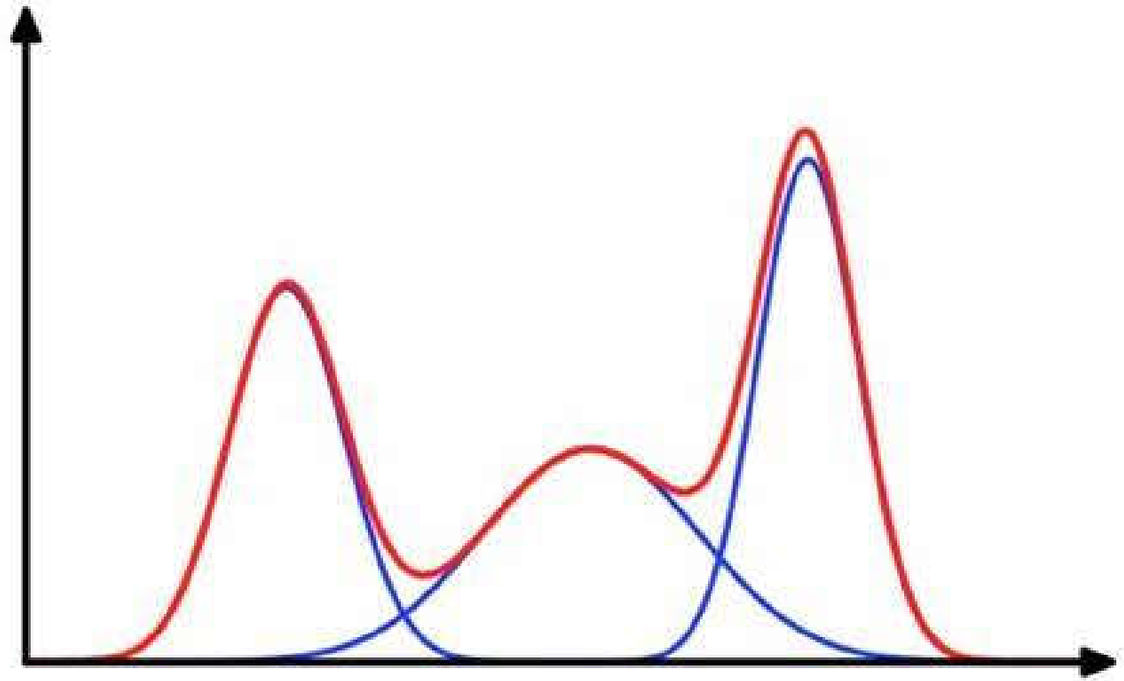
聚类



降维



概率模型估计



思考

思考题

- 损失函数的先验分布？为什么选择某一种损失函数?
- MLE/MAP与经验风险/结构风险最小化的关系?
- 最小二乘法的数学推导与矩阵表示?
- 噪声的数学含义是什么?
- L1正则化与L2正则化的对比?
- 怎么理解正则化对应着模型的先验概率?
- 交叉验证的数学原理?

思考题

- 交叉验证的损失与测试集上的损失的关系？测试集上的损失与期望损失的关系？
- 泛化误差上界的证明？
- 生成模型与判别模型特征的理解？
- 概率模型估计学习的是数据的横向纵向结构，怎么理解？
- 第一章的课后习题

参考资料

- 《统计学习方法（第二版）》，李航著
- 本笔记引用的图片多来自互联网，侵删

Thanks !