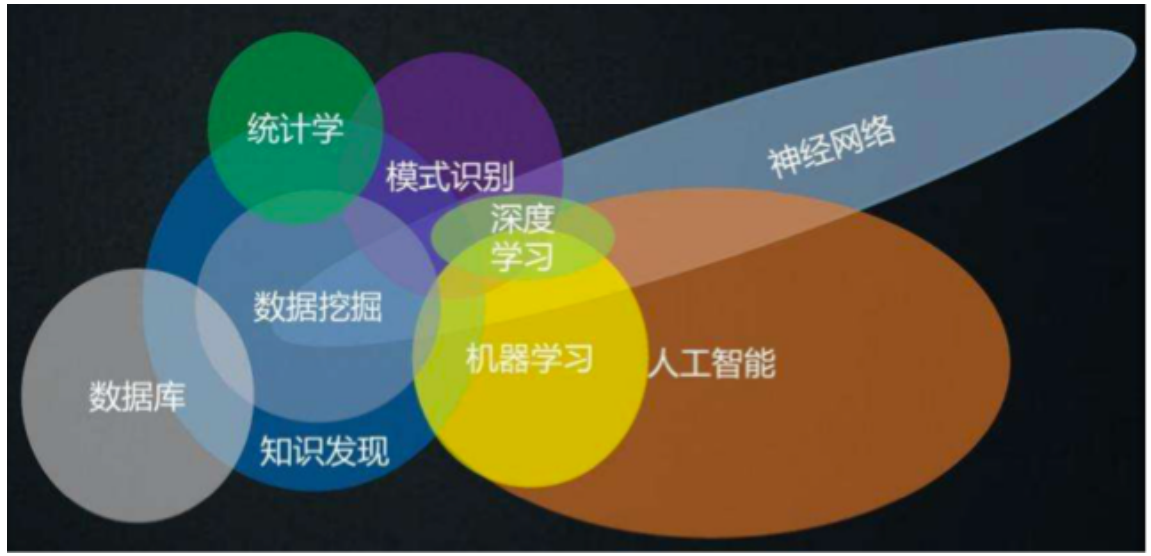
机器学习和基本算法介绍

Guangrui Qian

机器学习

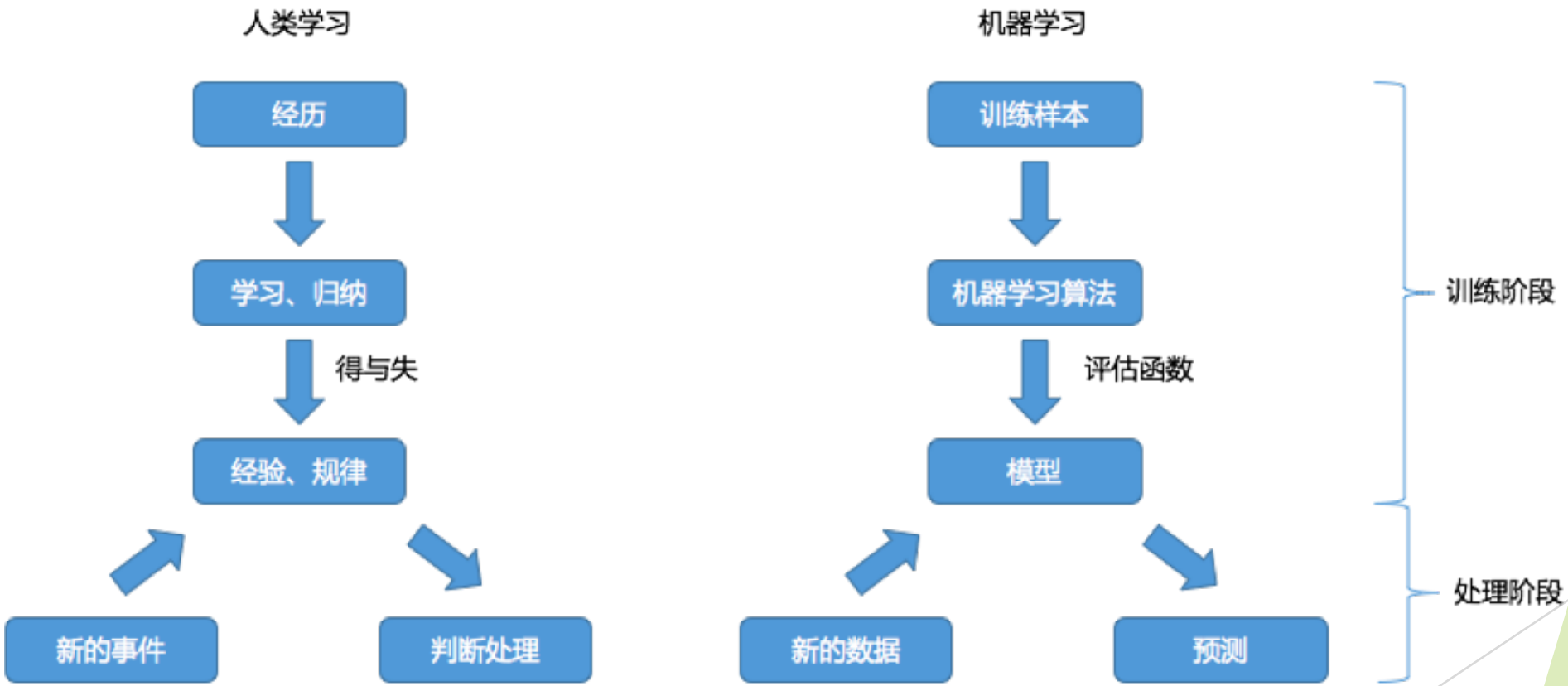
- 机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。
- 在信息时代中，我们周围处处都充满着机器学习，很多生活被机器学习影响。
- 机器学习有广泛和狭义的两种概念理解
  - 泛指所有数据分析的学习算法
  - 针对数值分析的分类、回归、聚类的数据分析算法



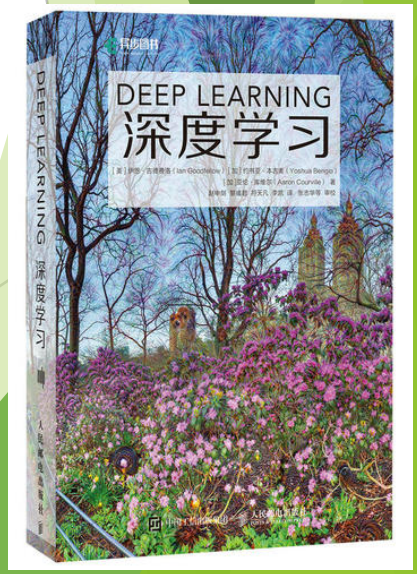
机器学习的内容和目标

- **数据堆积** 使用观察、记忆和联想的方法来为进一步的推理提供事实依据
- **抽象思维** 将存储的数据转换为更广泛的表示和概念
- **理论概括** 使用抽象化的数据来创建在新环境下采取进一步动作的知识和推论
- **评估过程** 为学习过程提供反馈机制以衡量所学知识的实用性并带来潜在的效果提升

何为机器学习？

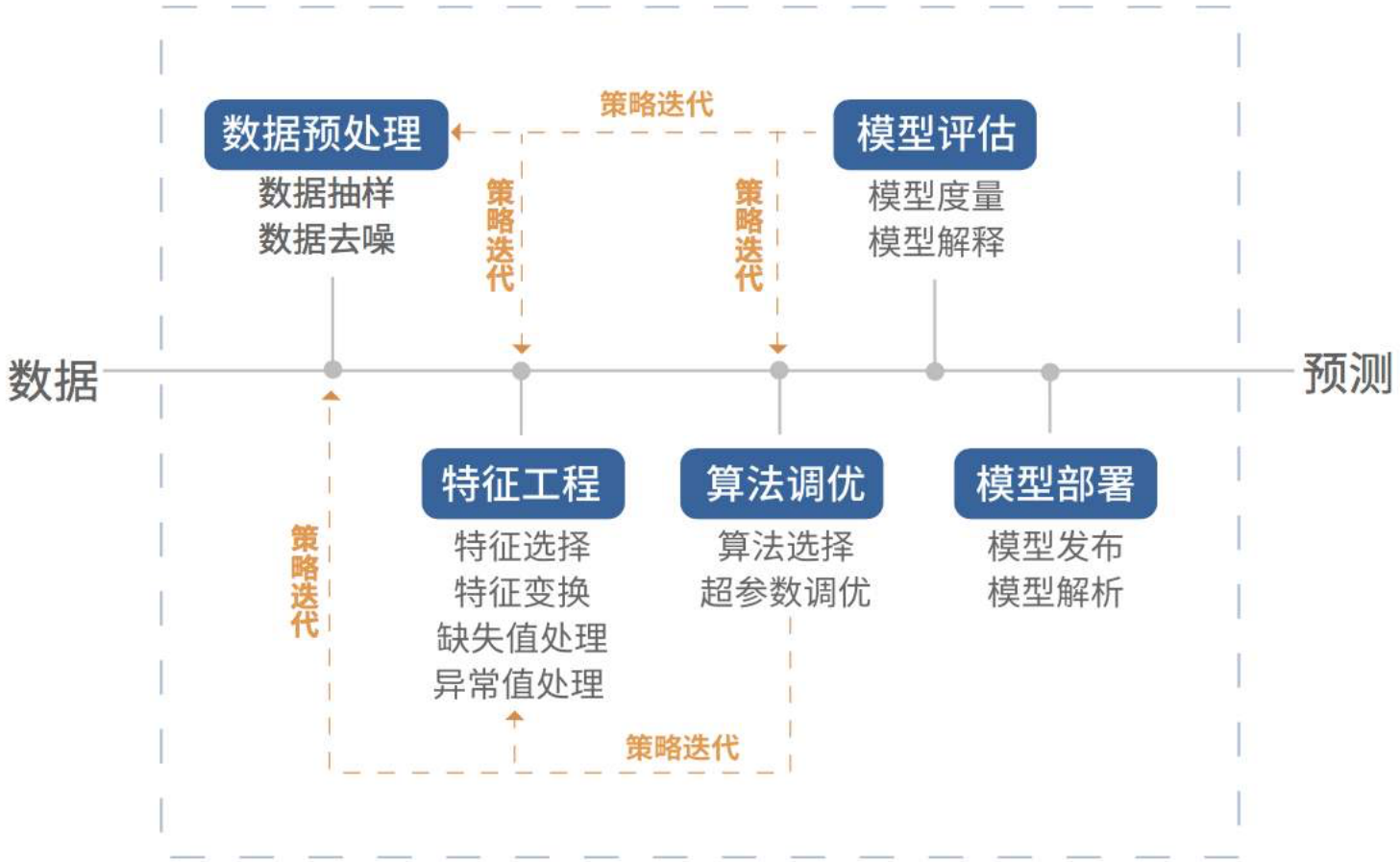

人类学习
经历
学习、归纳
得与失
经验、规律
新的事件
判断处理
机器学习
训练样本
机器学习算法
评估函数
模型
新的数据
预测
训练阶段
处理阶段


机器学习
周志华
MACHINE LEARNING
机器学习


DEEP LEARNING
深度学习

机器学习流程

- 数据
  - 数据收集与探索性分析
  - 数据预处理与特征工程
  - 数据集分割
- 算法/模型
  - 模型的选择
  - 模型训练
- 评估和优化
  - 模型评估
  - 模型优化
- 应用
  - 模型部署
  - 模型上线
  - 模型更新



机器学习模型设计五要素

- **{x,y}　- 数据中有多少东西可以学？**
  - 数据可能没什么用，但是数据中包含的信息有用，能够减少不确定性，数据中信息量决定了算法能达到的上限。起步阶段，先搞“量”再搞“率”应该是出效果最快的方式。
- **f(x)　- 模型有多聪明？能够学多少？**
  - f(x)的设计主要围绕参数量和结构两个方向做创新，这两个参数决定了算法的学习能力，从数据里面挖掘信息的能力（信息利用率）
- **objective　- 定个学习目标**

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

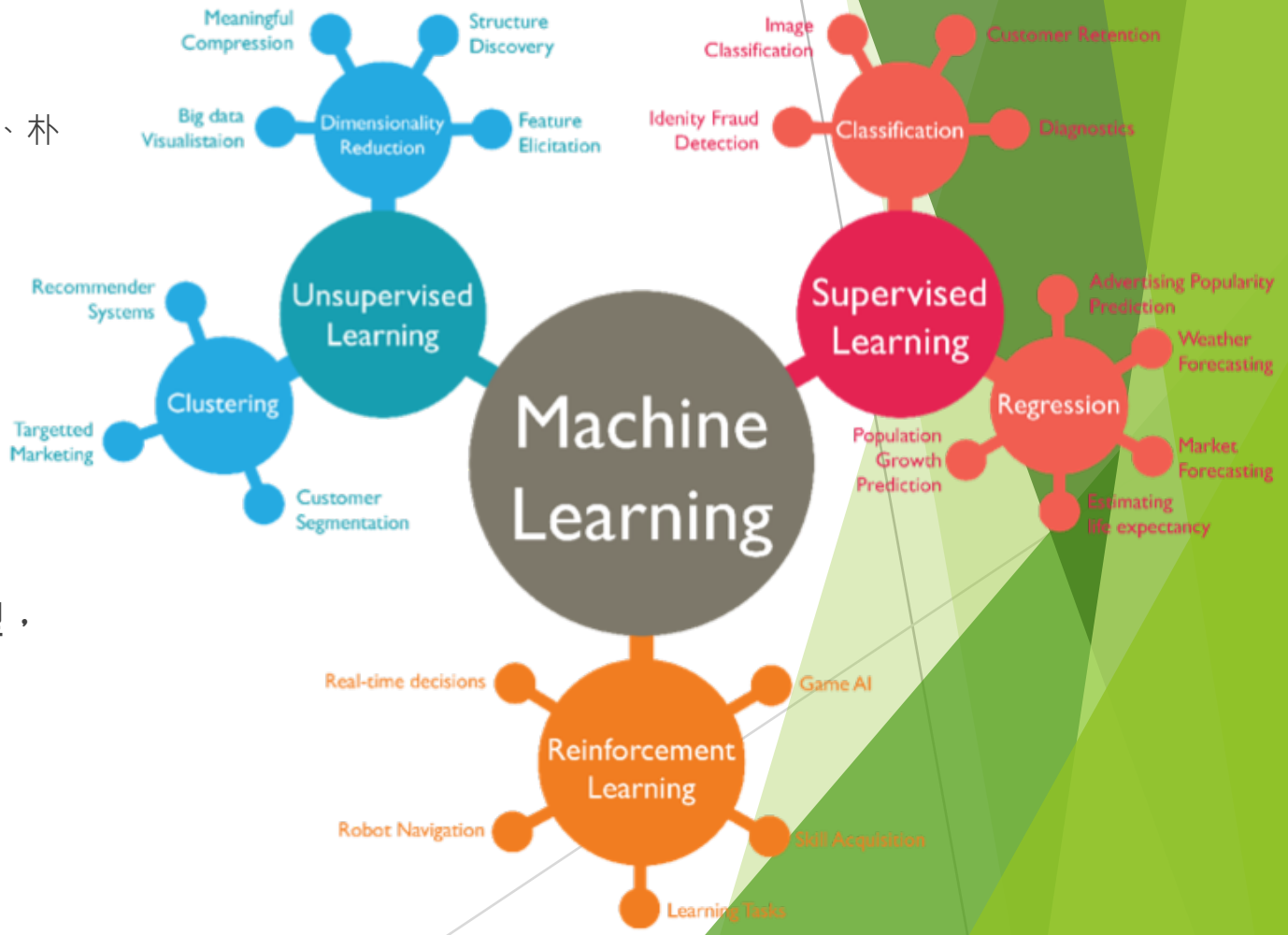
误差函数：我们的模型有多拟合数据

正则化项：惩罚复杂模型

- **optimization - 学习方法很重要**
- **evaluation　- 全面发展的“三好模型”**
  - 算法层面：准确率，覆盖率，auc，logloss...
  - 公司层面：revenue，ctr，cvr...
  - 用户层面：用户体验，满意度，惊喜度...

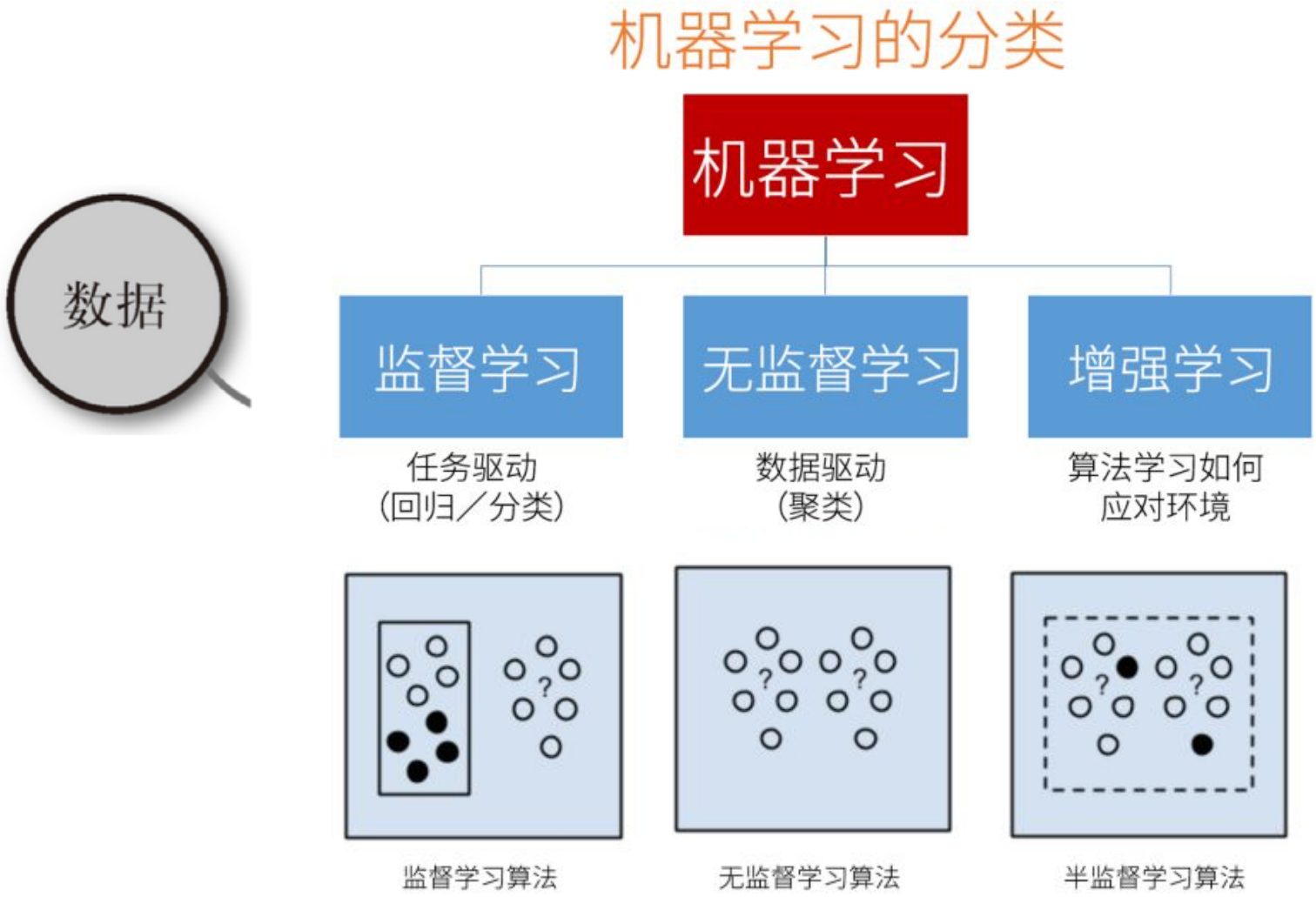
机器学习算法/模型分类

- **监督学习：构建预测模型，使用「标签化」数据**
  - 分类（逻辑回归、决策树、KNN、随机森林、支持向量机、朴素贝叶斯等）
  - 回归（线性回归、KNN、Gradient Boosting & AdaBoost等）
- **无监督学习：构建描述模型，使用「无标签」数据**
  - 聚类（K-Means）
  - 模式挖掘
  - 数据降维
- **增强/强化学习：构建智能体agent在与环境的模型，通过交互过程学习**



其他：半监督学习和迁移学习

机器学习算法/模型分类



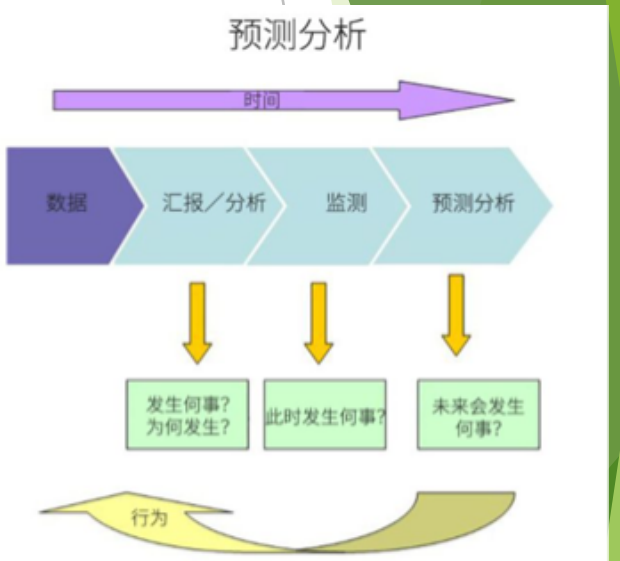
数据

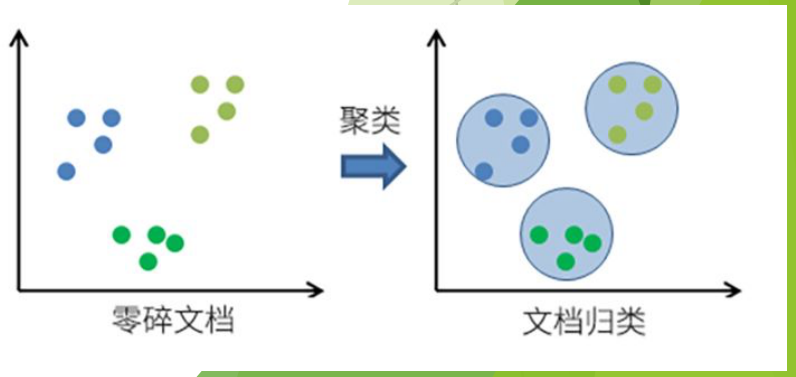
答案

机器学习算法分类和应用

模型类型

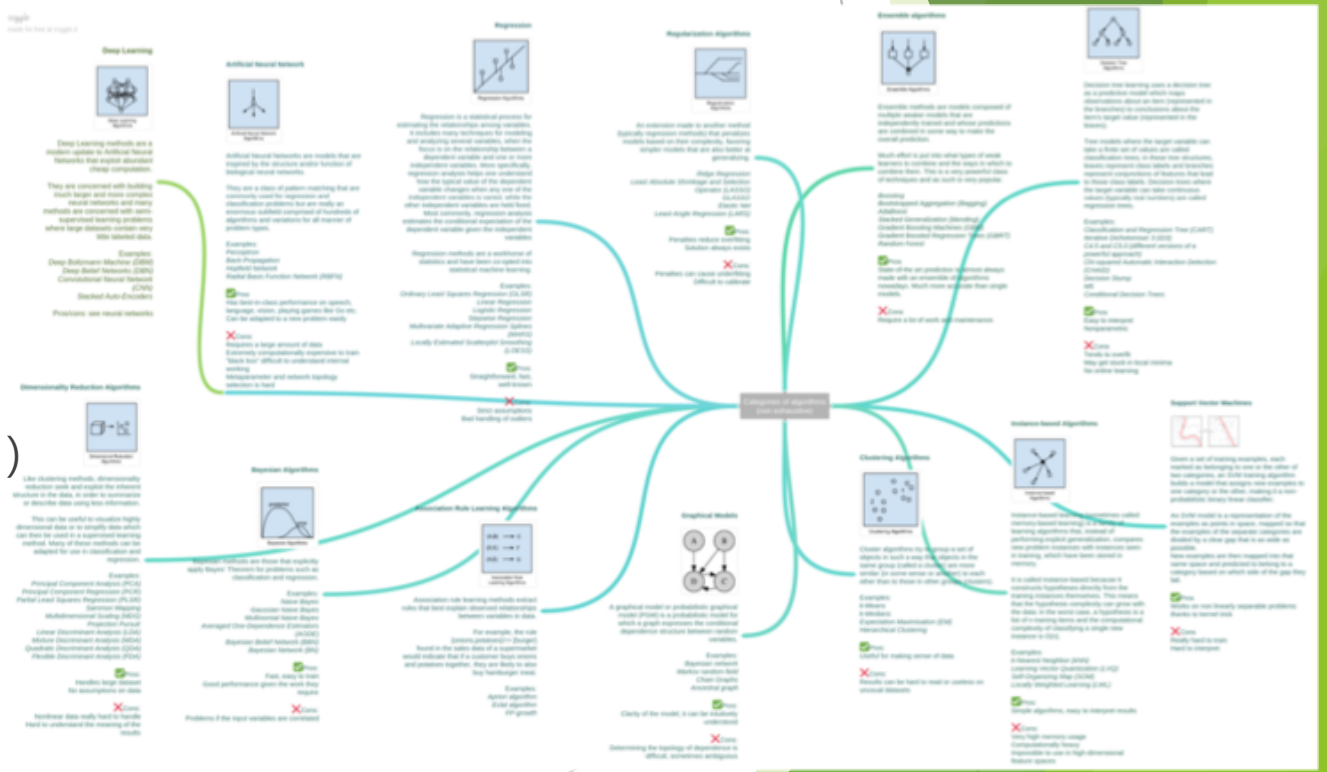
- 预测性模型
  - 预测模型是为了解决这样的任务而诞生的：即我们需要使用一个数据集中的很多参数来预测另外的一个参数。在这一过程中，学习算法会试着发现和建模目标特征（也就是需要预测的特征）与其他特征之间的联系。
- 描述性模型
  - 描述模型可以应用在这样的任务中：即我们需要从归纳性的数据中以新颖且多样化的方式来获取概念特征。与预测模型相比，描述模型倾向于预测一种「喜好」，而对于这种喜好而言，所有与其相关的特征都是平等的。也就是说，由于描述模型中并不存在用于训练的目标参数，因此对于描述模型的训练过程我们可以称之为无监督学习。





机器学习模型和算法概述

- 回归（Regression）
- 聚类算法（Clustering Algorithms）
- 正则化算法（Regularization Algorithms）
- 集成算法（Ensemble Algorithms）
- 决策树算法（Decision Tree Algorithm）
- 人工神经网络（Artificial Neural Network）
- 深度学习（Deep Learning）
- 支持向量机（Support Vector Machine）
- 降维算法（Dimensionality Reduction Algorithms）
- 基于实例的算法（Instance-based Algorithms）
- 贝叶斯算法（Bayesian Algorithms）
- 关联规则学习算法（Association Rule Learning Algorithms）
- 图模型（Graphical Models）



机器学习算法概述 - 回归

回归是用于估计两种变量之间关系的统计过程。当用于分析因变量和一个多个自变量之间的关系时，该算法能提供很多建模和分析多个变量的技巧。具体一点说，回归分析可以帮助我们理解当任意一个自变量变化，另一个自变量不变时，因变量变化的典型值。最常见的是，回归分析能在给定自变量的条件下估计出因变量的条件期望。

回归算法是统计学中的主要算法，它已被纳入统计机器学习。

例子：

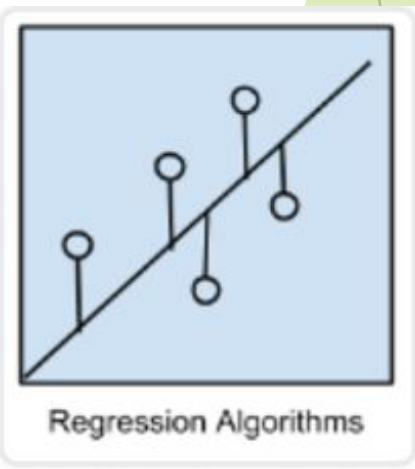
- 普通最小二乘回归（**Ordinary Least Squares Regression**，**OLSR**）
- 线性回归（**Linear Regression**）
- 逻辑回归（**Logistic Regression**）
- 逐步回归（**Stepwise Regression**）
- 多元自适应回归样条（**Multivariate Adaptive Regression Splines**，**MARS**）
- 本地散点平滑估计（**Locally Estimated Scatterplot Smoothing**，**LOESS**）

优点：

- 直接、快速
- 知名度高

缺点：

- 要求严格的假设
- 需要处理异常值



机器学习算法概述 – 支持向量机

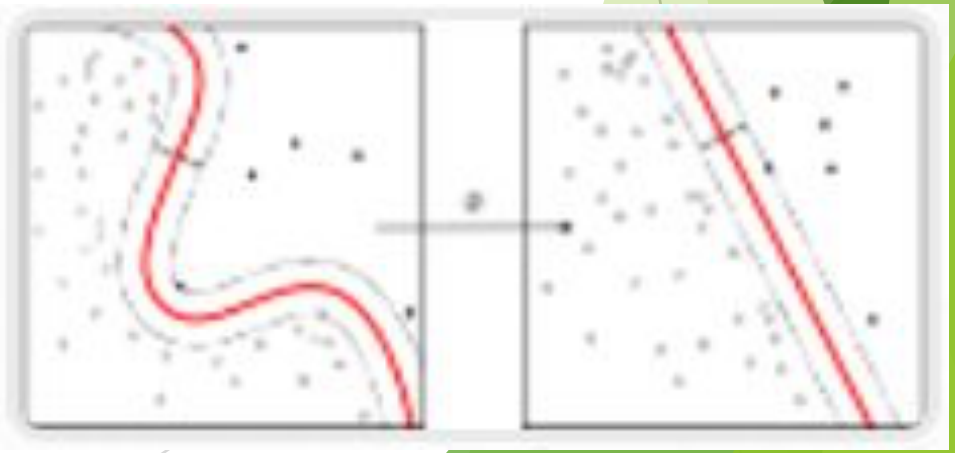
- 给定一组训练事例，其中每个事例都属于两个类别中的一个，支持向量机（**SVM**）训练算法可以在被输入新的事例后将其分类到两个类别中的一个，使自身成为非概率二进制线性分类器。
- **SVM** 模型将训练事例表示为空间中的点，它们被映射到一幅图中，由一条明确的、尽可能宽的间隔分开以区分两个类别。
- 随后，新的示例会被映射到同一空间中，并基于它们落在间隔的哪一侧来预测它属于的类别。

优点：

- 在非线性可分问题上表现优秀

缺点：

- 非常难以训练
- 很难解释



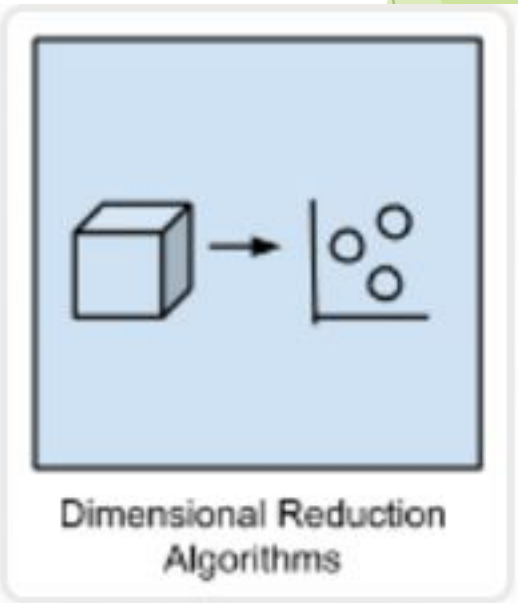
机器学习算法概述 – 降维算法

- 和集簇方法类似，降维追求并利用数据的内在结构，目的在于使用较少的信息总结或描述数据。
- 这一算法可用于可视化高维数据或简化接下来可用于监督学习中的数据。许多这样的方法可针对分类和回归的使用进行调整。

优点：
- 可处理大规模数据集
- 无需在数据上进行假设

缺点：
- 难以搞定非线性数据
- 难以理解结果的意义

例子：
- 主成分分析（**Principal Component Analysis (PCA)**）
- 主成分回归（**Principal Component Regression (PCR)**）
- 偏最小二乘回归（**Partial Least Squares Regression (PLSR)**）
- **Sammon** 映射（**Sammon Mapping**）
- 多维尺度变换（**Multidimensional Scaling (MDS)**）
- 投影寻踪（**Projection Pursuit**）
- 线性判别分析（**Linear Discriminant Analysis (LDA)**）
- 混合判别分析（**Mixture Discriminant Analysis (MDA)**）
- 二次判别分析（**Quadratic Discriminant Analysis (QDA)**）
- 灵活判别分析（**Flexible Discriminant Analysis (FDA)**）

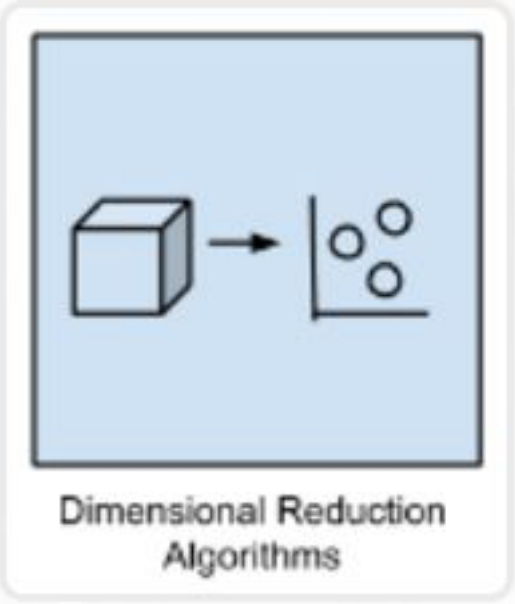


机器学习算法概述 – 聚类算法

- 聚类算法是指对一组目标进行分类，属于同一组（亦即一个类，**cluster**）的目标被划分在一组中，与其他组目标相比，同一组目标更加彼此相似（在某种意义上）。

例子：

- **K-均值（k-Means）**
- **k-Medians 算法**
- **Expectation Maximi 封层 ation (EM)**
- **最大期望算法（EM）**
- **分层集群（Hierarchical Clstering）**



优点：

- 让数据变得有意义

缺点：

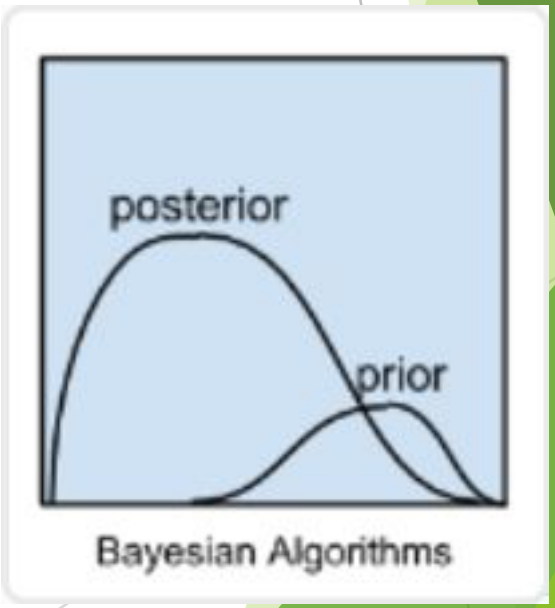
- 结果难以解读，针对不寻常的数据组，结果可能无用。

机器学习算法概述 – 贝叶斯算法

- 贝叶斯方法是指明确应用了贝叶斯定理来解决如分类和回归等问题的方法。

例子：

- 朴素贝叶斯（**Naive Bayes**）
- 高斯朴素贝叶斯（**Gaussian Naive Bayes**）
- 多项式朴素贝叶斯（**Multinomial Naive Bayes**）
- 平均一致依赖估计器（**Averaged One-Dependence Estimators (AODE)**）
- 贝叶斯信念网络（**Bayesian Belief Network (BBN)**）
- 贝叶斯网络（**Bayesian Network (BN)**）



优点：

- 快速、易于训练、给出了它们所需的资源能带来良好的表现

缺点：

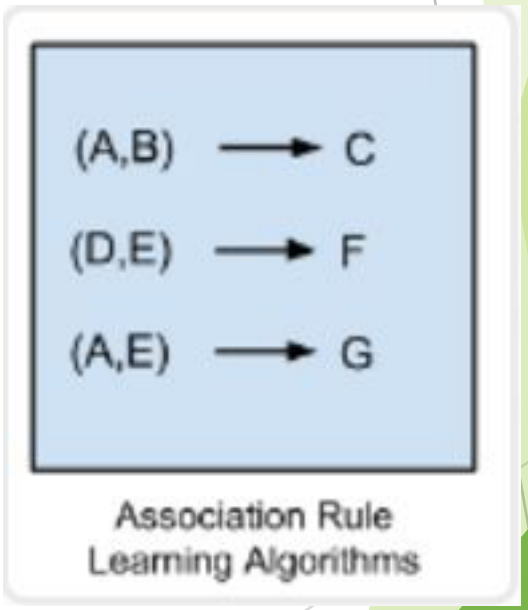
- 如果输入变量是相关的，则会出现问题

机器学习算法概述 - 关联规则学习算法

- 关联规则学习方法能够提取出对数据中的变量之间的关系的最佳解释。比如说一家超市的销售数据中存在规则 {洋葱，土豆}=> {汉堡}，那说明当一位客户同时购买了洋葱和土豆的时候，他很有可能还会购买汉堡肉。

例子：

- **Apriori 算法（Apriori algorithm）**
- **Eclat 算法（Eclat algorithm）**
- **FP-growth**



优点：

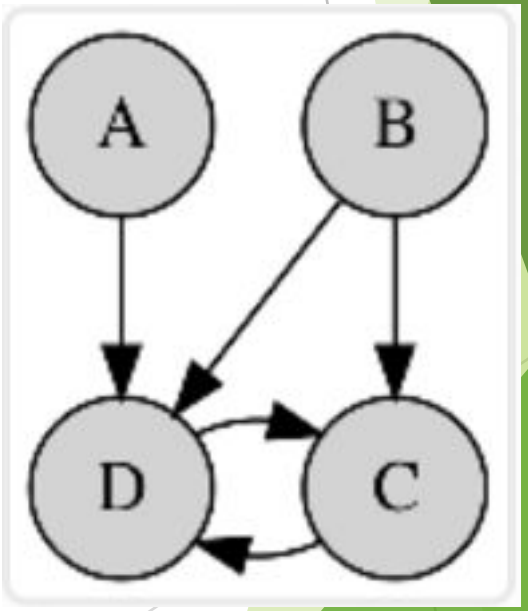
- 易编码实现

缺点：

- 大数据上速度较慢，候选集每次产生过多，未排除不应该参与计算支持度的点。每次都需要计算支持度，需对全部记录扫描，需要很大I/O负载

机器学习算法概述 - 图模型

- 图模型或概率图模型（**PGM/probabilistic graphical model**）是一种概率模型，一个图（**graph**）可以通过其表示随机变量之间的条件依赖结构（**conditional dependence structure**）。



例子：

- 贝叶斯网络（**Bayesian network**）
- 马尔可夫随机域（**Markov random field**）
- 链图（**Chain Graphs**）
- 祖先图（**Ancestral graph**）

优点：

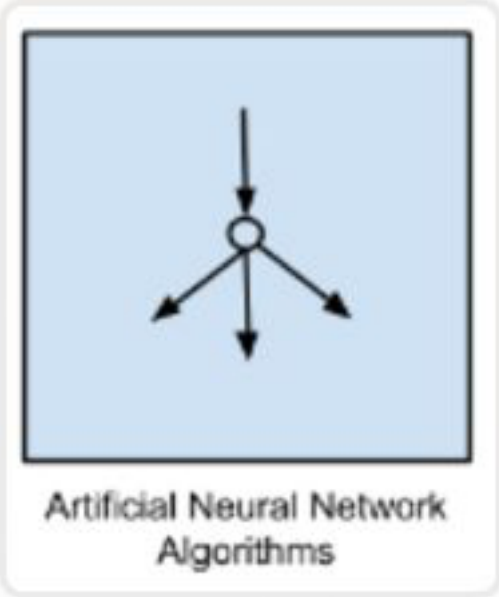
- 模型清晰，能被直观地理解

缺点：

- 确定其依赖的拓扑很困难，有时候也很模糊

机器学习算法概述 – 人工神经网络

- 人工神经网络是受生物神经网络启发而构建的算法模型。
- 它是一种模式匹配，常被用于回归和分类问题，但拥有庞大的子域，由数百种算法和各类问题的变体组成。



例子：

- 感知器
- 反向传播
- **Hopfield** 网络
- 径向基函数网络（**Radial Basis Function Network**，**RBFN**）

优点：

- 在语音、语义、视觉、各类游戏（如围棋）的任务中表现极好。
- 算法可以快速调整，适应新的问题。

缺点：

- 需要大量数据进行训练
- 训练要求很高的硬件配置
- 模型处于「黑箱状态」，难以理解内部机制
- 元参数（**Metaparameter**）与网络拓扑选择困难。

机器学习算法概述 – 深度学习

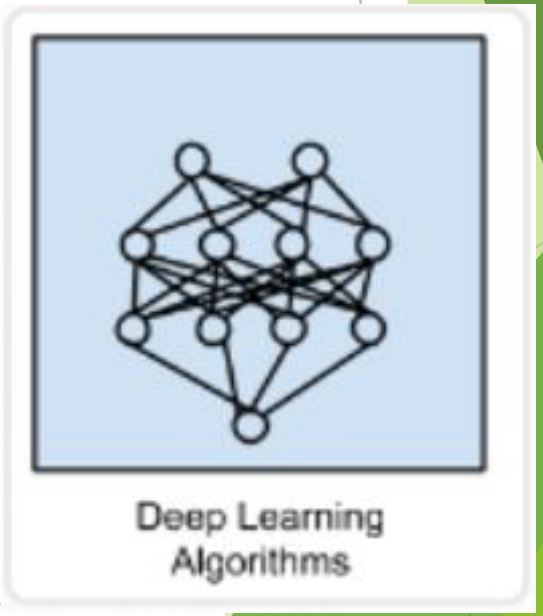
- 深度学习是人工神经网络的最新分支，它受益于当代硬件的快速发展。
- 众多研究者目前的方向主要集中于构建更大、更复杂的神经网络，目前有许多方法正在聚焦半监督学习问题，其中用于训练的大数据集只包含很少的标记。

例子：

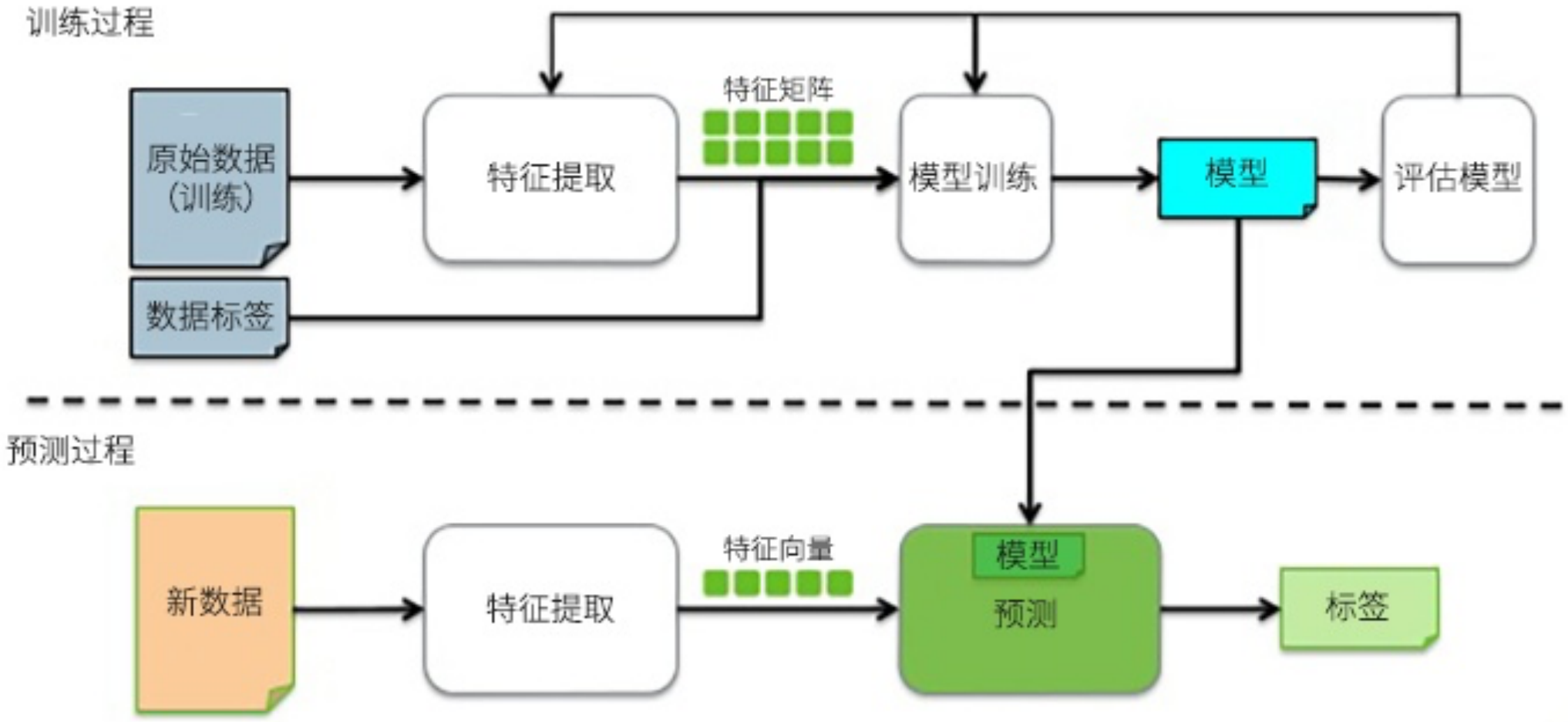
- 深玻耳兹曼机（**Deep Boltzmann Machine，DBM**）
- **Deep Belief Networks**（**DBN**）
- 卷积神经网络**(CNN)**
- 循环神经网络**(RNN)**
- 对抗神经网络**(GAN)**
- **Stacked Auto-Encoders**

优点/缺点：

- 优缺点和神经网络相同



监督学习流程

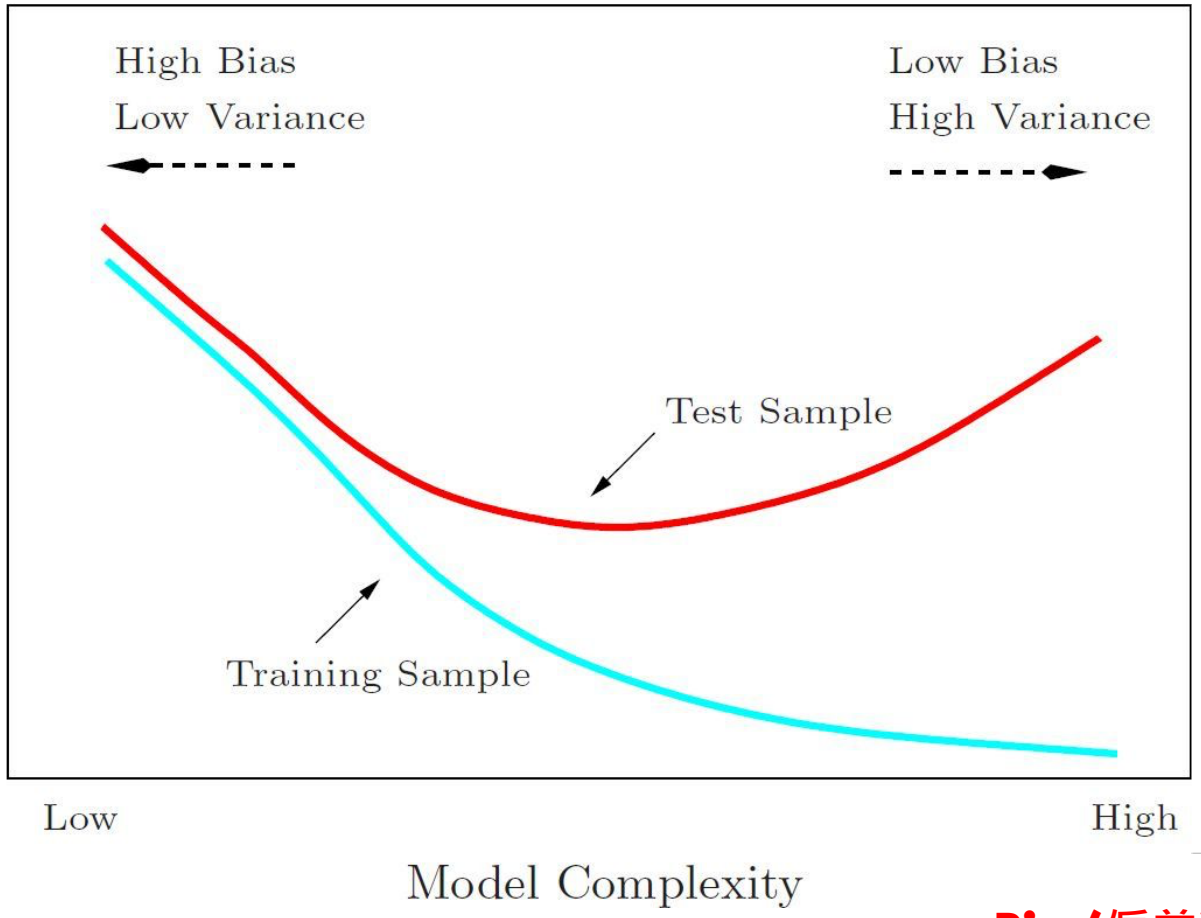

训练过程
原始数据
(训练)
数据标签
特征提取
特征矩阵
模型训练
模型
评估模型
预测过程
新数据
特征提取
特征向量
模型
预测
标签

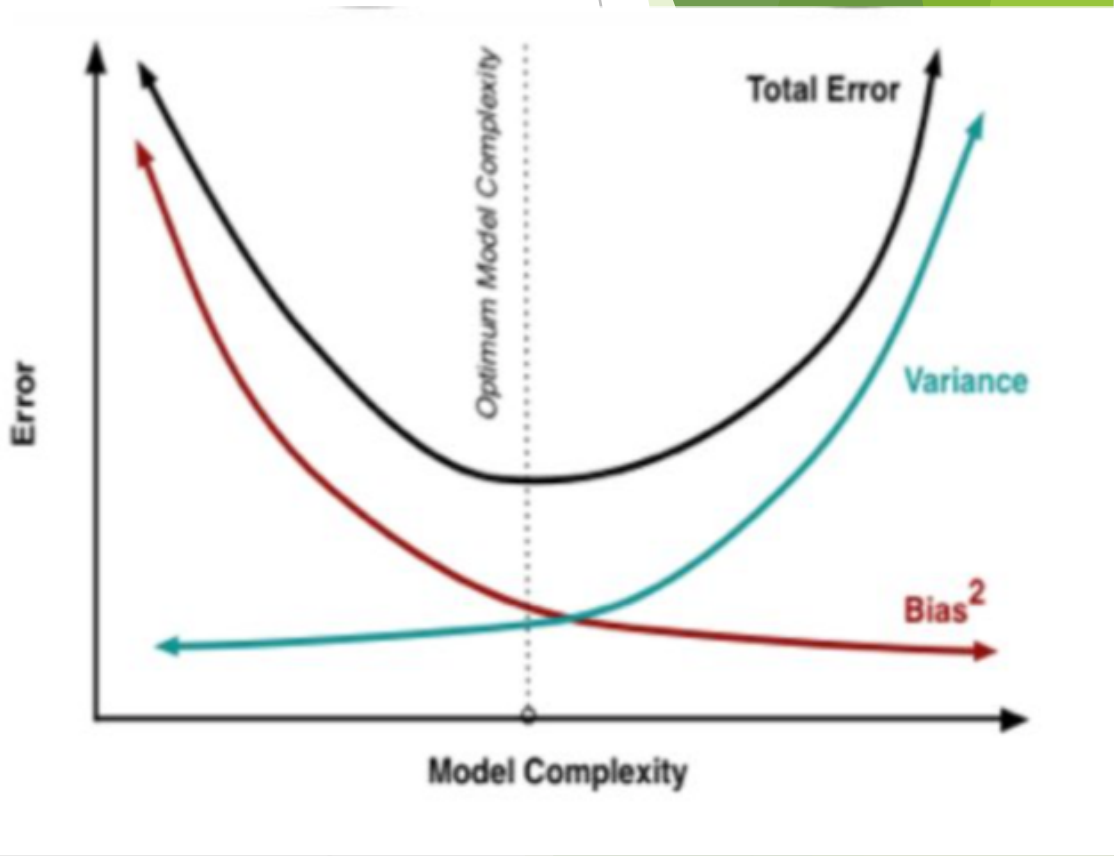
模型评估

模型评估在任何机器学习过程中都是一个关键性步骤。在监督学习和无监督学习中，评估过程也有所不同。对于监督学习模型而言，预测是模型的主要工作；而对于无监督学习模型而言，群组内的同质性和群组之间的异质性起着重要作用。

- 回归模型（包括交叉验证）的模型评估参数：
  - 确定系数
  - 均方根误差
  - 平均绝对误差
  - 赤池或贝叶斯信息量准则
- 分类模型（包括交叉验证）的模型评估参数：
  - 混淆矩阵（准确率、精确率、召回率以及F1值）
  - 增益或提升图
  - ROC（接收机工作特性）曲线面积
  - 和谐-不和谐比
- 无监督模型（聚类）的模型评估参数：
  - 列联表
  - 聚类对象与聚类中心或重心之间的平方误差之和
  - 轮廓值
  - 兰德指数
  - 匹配指数
  - 成对和调整配对精确率和召回率（主要用于NLP中）

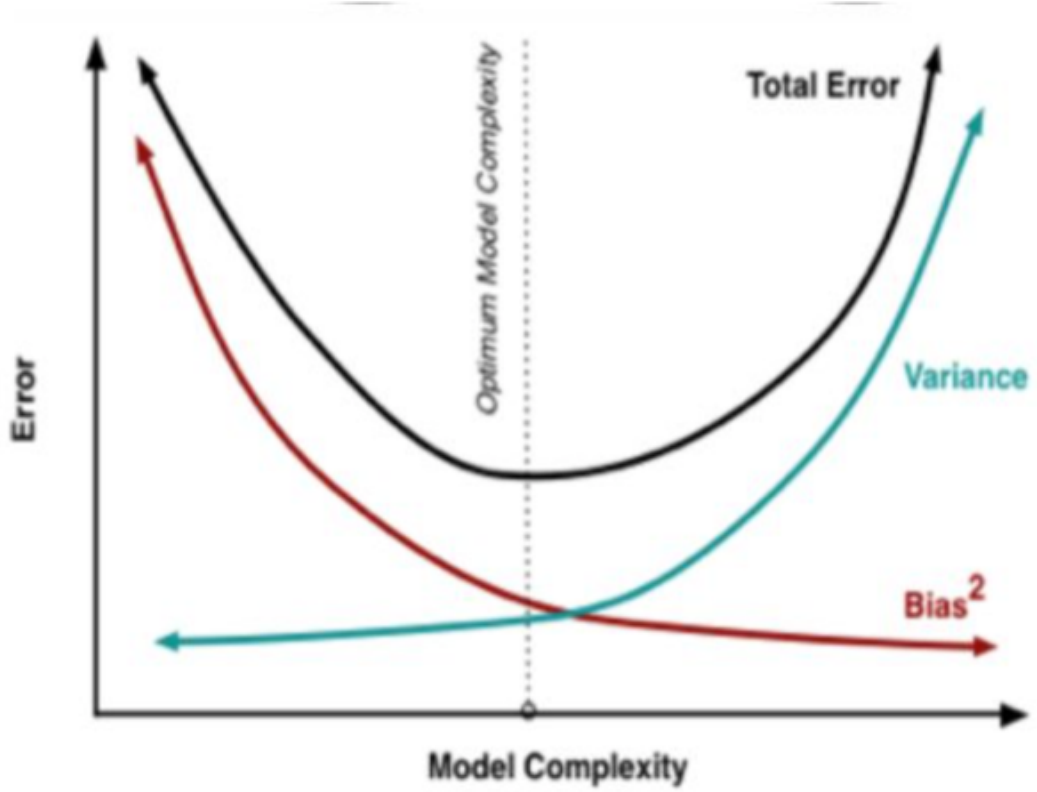
模型复杂度和准确率

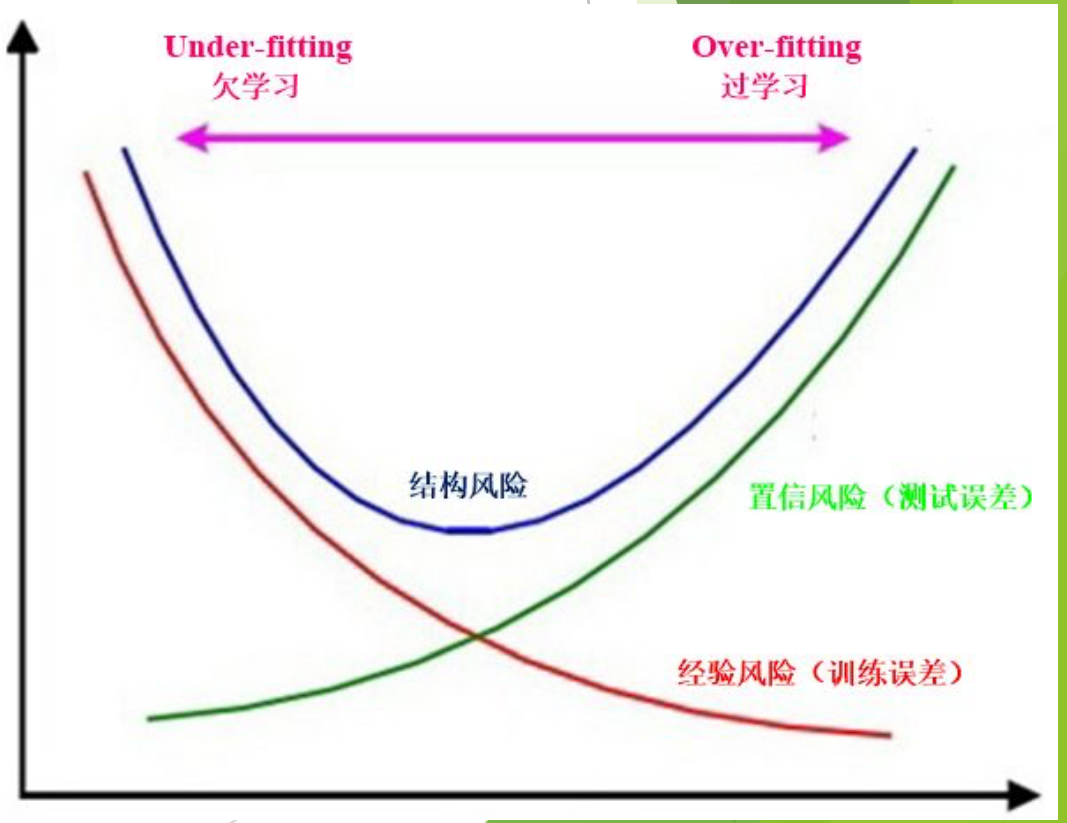




Bias(偏差)，Error(误差)，Variance(方差)

模型复杂度和准确率


Total Error
Optimum Model Complexity
Variance
Bias^2
Error
Model Complexity


Under-fitting
欠学习
Over-fitting
过学习
结构风险
置信风险（测试误差）
经验风险（训练误差）

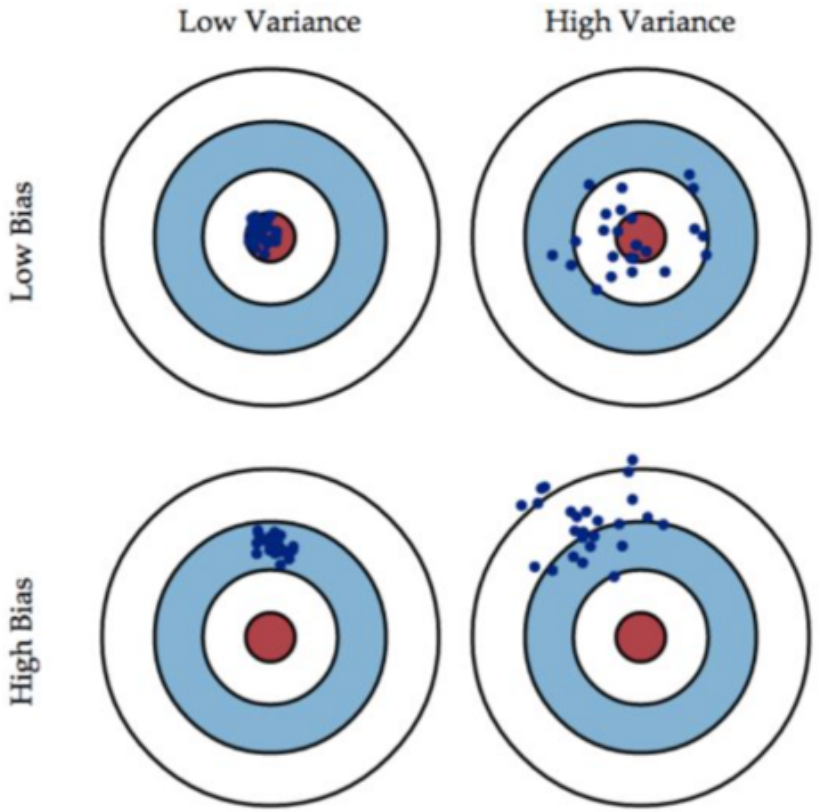
机器学习中的"准"与"确"

Bias

- 衡量模型拟合训练数据的能力（训练数据不一定是整个 training dataset，而是只用于训练它的那一部分数据，例如：mini-batch）
- bias 越小，拟合能力越高（可能产生overfitting）；反之，拟合能力越低（可能产生underfitting）
- bias是针对一个模型来说的

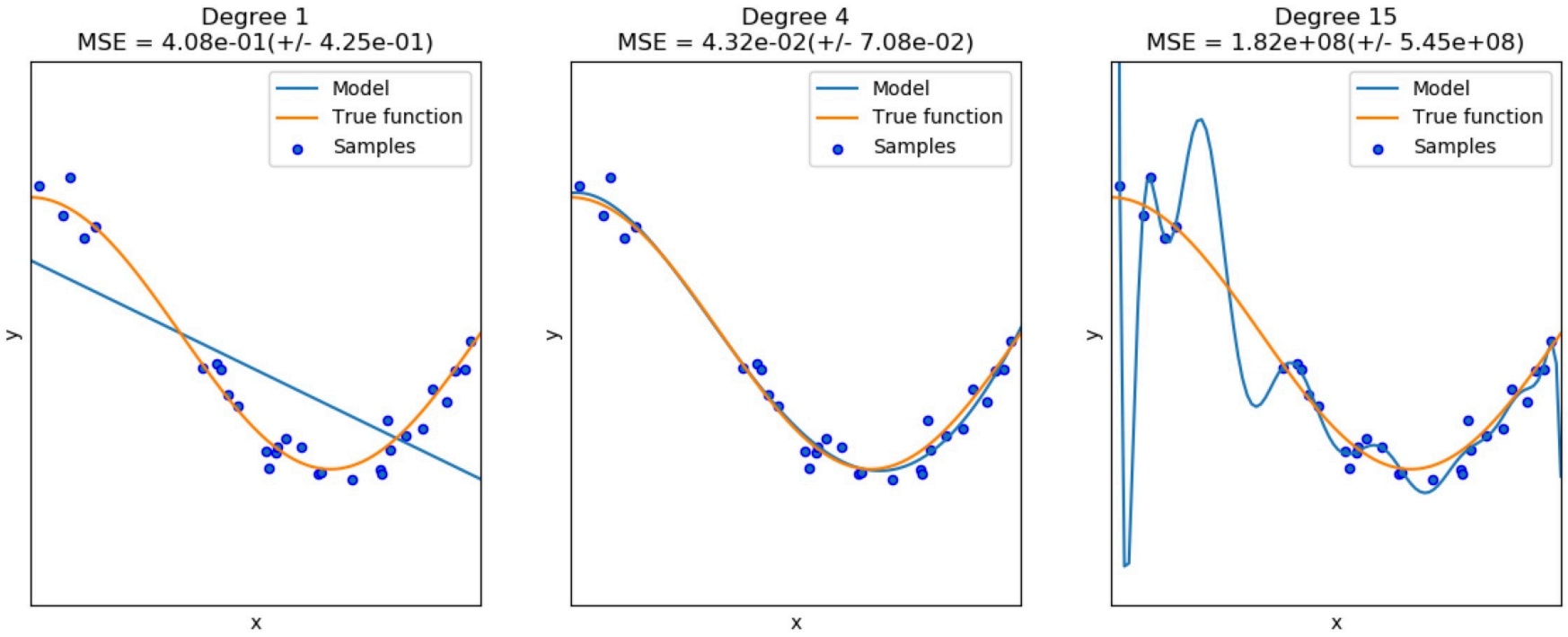
Variance

- 衡量模型的 generalization 的能力
- variance 越小，模型的 generalization 的能力越高；反之，模型的 generalization 的能力越低
- 它是针对多个模型来说的



Error：泛化误差，是误差、方差、噪声之和
Bias：偏差，对象是单个模型，期望输出与真实标记的差别
Variance：方差，对象是多个模型

欠拟合和过拟合



- 图一是欠拟合，拟合曲线是线性的，这是简单曲线，对数据欠拟合。真实值与预测值差别大，高偏差。预测值都在拟合曲线上，很明显是低方差的。所以这张图片对应欠拟合，高偏差，低方差。

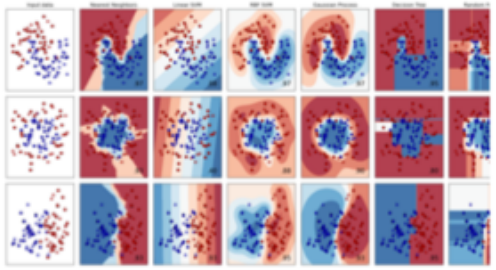
- 图三是过拟合，拟合曲线十分复杂，每次都命中训练数据，然而强行拟合每个数据导致预测值方差很大（波动大），预测值都等于真实值，所以偏差小。所以这张图片对应过拟合，低偏差，高方差。

机器学习Python库Scikit-learn

Classification

Identifying which category an object belongs to.

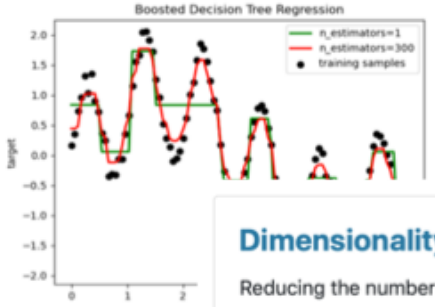
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...



Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering

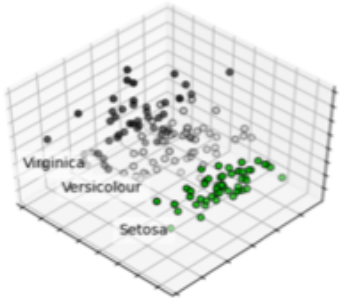
Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

Dimensionality reduction

Reducing the number of random variables to consider.

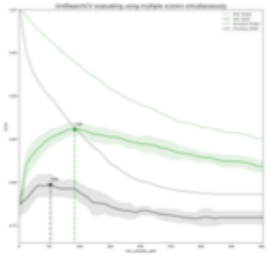
Applications: Visualization, Increased efficiency
Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...



Model selection

Comparing, validating and choosing parameters and models.

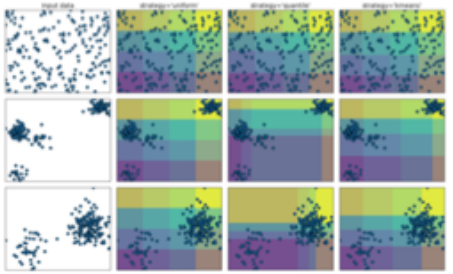
Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...



Preprocessing

Feature extraction and normalization.

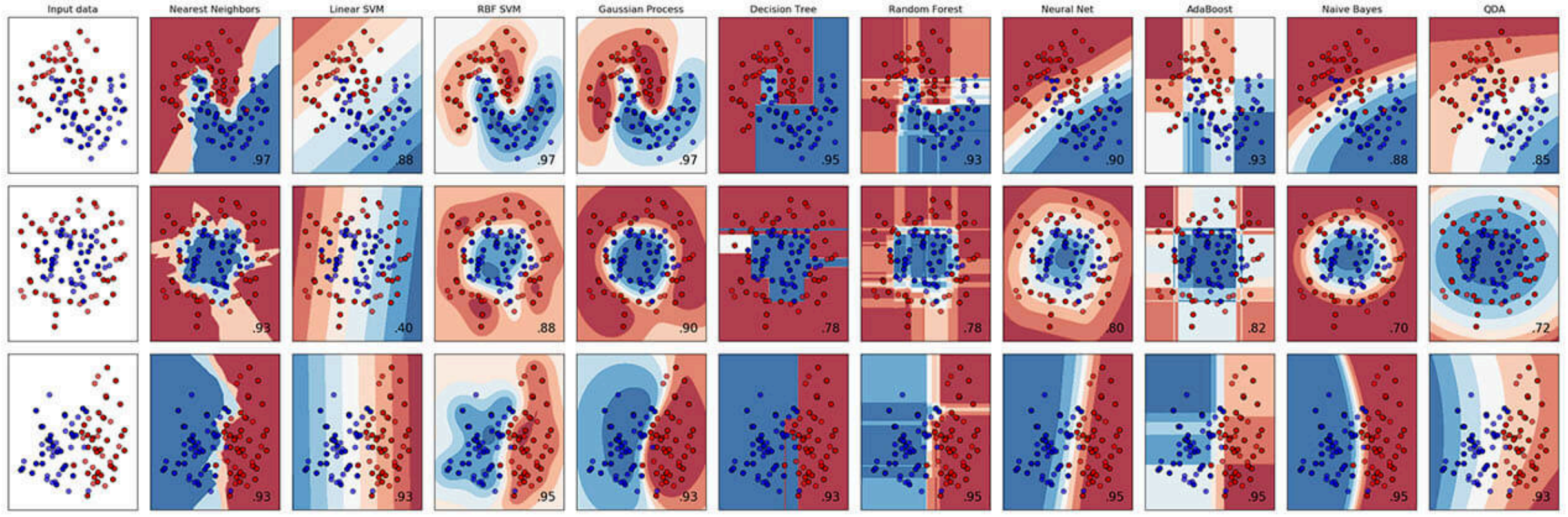
Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...



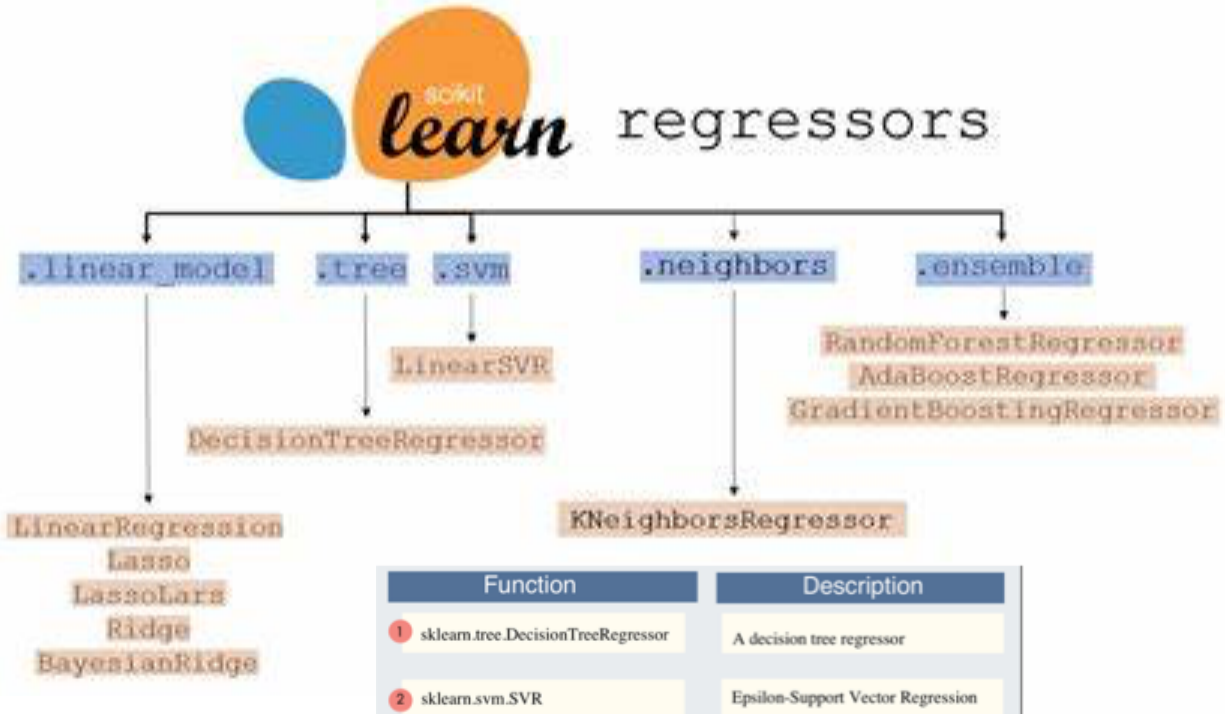
https://scikit-learn.org/stable/index.html

Scikit-learn框架介绍

- Scikit-learn的基本功能主要被分为六大部分：分类，回归，聚类，数据降维，模型选择和数据预处理。
- 专门面向机器学习的Python开源框架，Scikit-learn可以在一定范围内为开发者提供非常好的帮助。它内部实现了各种各样成熟的算法，容易安装和使用，样例丰富，而且教程和文档也非常详细。
- 不支持深度学习和强化学习， 不支持图模型和序列预测，不支持Python之外的语言，不支持GPU加速。



Sklearn中的regressors和classifiers



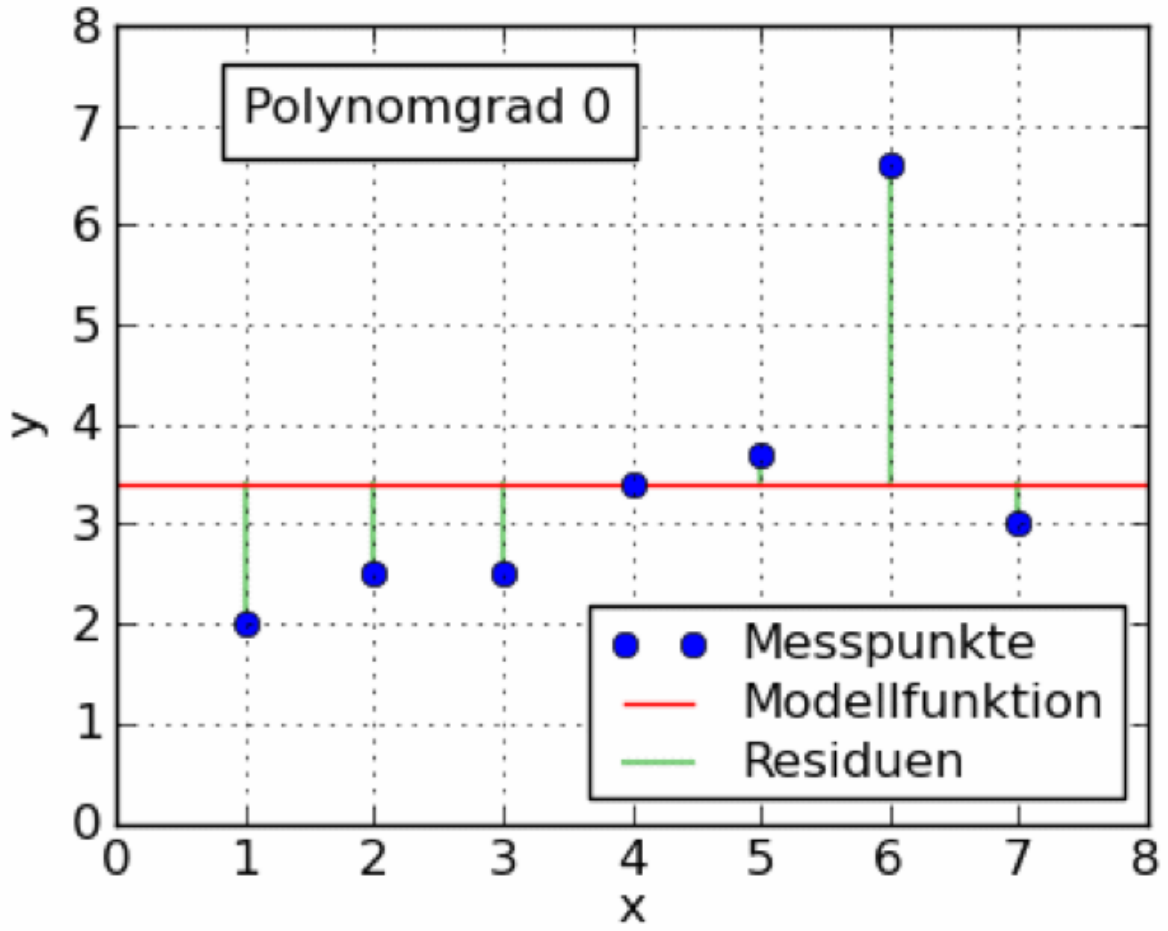
Function	Description
1 sklearn.tree.DecisionTreeRegressor	A decision tree regressor
2 sklearn.svm.SVR	Epsilon-Support Vector Regression
3 sklearn.linear_model.Linear Regression	Ordinary least squares Linear Regression
4 sklearn.linear_model.Lasso	Linear Model trained with L1 prior as regularizer (a.k.a the Lasso)
5 sklearn.linear_model.SGDRegressor	Linear model fitted by minimizing a regularized empirical loss with SGD
6 sklearn.linear_model.ElasticNet	Linear regression with combined L1 and L2 priors as regularizer
7 sklearn.ensemble.RandomForest Regressor	A random forest regressor
8 sklearn.ensemble.GradientBoosting Regressor	Gradient Boosting for regression
9 sklearn.neural_network. MLPRegressor	Multi-layer Perceptron regressor

Function	Description
1 sklearn.neural_network.MLP Classifier	Multi-layer Perceptron classifier
2 sklearn.tree.DecisionTreeClassifier	A decision tree classifier
3 sklearn.svm.SVC	C-Support Vector Classification
4 sklearn.linear_model.Logistic Regression	Logistic Regression (at.k.a logit, Max Ent) classifier
5 sklearn.linear_model.SGDClassifier	Linear classifiers (SVM, logistic regression, a.o.) with SGD training
6 sklearn.naive_bayes.GaussianNB	Gaussian Naive Bayes
7 sklearn.neighbors.KNeighbors Classifier	Classifier implementing the k-nearest neighbors vote
8 sklearn.ensemble.RandomForest Classifier	A random forest classifier
9 sklearn.ensemble.GradientBoosting Classifier	Gradient Boosting for classification

数据拟合和回归

- 曲线拟合是一个求解曲线模型或数学函数的过程，该曲线或数学函数在某种约束下最优表征该组数据点。曲线拟合可能要求拟合结果精确穿过数据（插值）或拟合的结果足够“平滑”。**拟合是一种数据建模的方法**。

- 回归分析（regression analysis）指的是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。回归分析按照涉及的变量的多少，分为一元回归和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。
- 回归是一种数据分析方法；拟合是一种数据建模方法；拟合侧重于调整曲线的参数，使得与数据相符。而回归重在研究两个变量或多个变量之间的关系。它可以用拟合的手法来研究两个变量的关系，以及出现的误差。



谢谢