Artificial Intelligence

AI

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

Models in Machine Learning

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

Contents:

Contents:

Objectives 教学目的

- This chapter will discuss in detail about the models that have been used in machine learning.
  这一章我们将详细讨论用于机器学习的一些模型。

What are Learning Models 什么是学习的模型

- The learning models are used to denote the approaches that can handle to fulfil a learning task.

  学习模型用于表示可以处理完成一个学习任务的方法。

- It is the algorithm-level but not data-level models (approaches).

  它是属于算法层面而不是数据层面的模型 (方法)。

Why Study Learning Models 为什么要研究学习的模型

- The result of machine learning is heavily dependent on the choice of an approach for solving the learning task.

  机器学习的效果在很大程度上取决于解决该学习任务时所选用的方法。

Typical Models for Machine Learning 机器学习的代表性模型

Models 模型	Brief Statements 简短描述	Sub-models 子模型	Typical Algorithm 典型算法
Probabilistic 概率	Use probabilistic models to denote the conditional dependence between random variables. 采用概率模式来表示随机变量之间的条件相关性。	Bayes 贝叶斯	Bayesian Network 贝叶斯网络
		Generative 生成	Probabilistic Program. 概率规划
		Statistic 统计	Linear Regression 线性回归
Geometric 几何	Use geometric models such as line, plane, distance or manifold to construct learning algorithms. 采用线、面、距离或流行等几何图形模型来构建学习算法。	Line 线	Linear Regression 线性回归
		Plane 面	SVM 支撑向量机
		Distance 距离	*k*-NN *k*-近邻
		Manifold 流行	Isomap 等距映射
Logical 逻辑	Use logical models to construct learning algorithms. 采用逻辑模型来构建学习算法。	Logic 逻辑	Inductive Logic Program. 归纳逻辑编程
		Rule 规则	Association Rule 相关规则
Networked 网络	Use networked models to construct learning algorithms. 采用网络模式构建机器学习算法。	Shallow 浅层	Perceptron 感知机
		Deep 深层	CNN 卷积神经网络

Contents:

Probabilistic Models

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

What are Probabilistic Models 什么是概率模型

- Probability theory is used to study the uncertainty problems, such as:
  概率论用于研究不确定性问题，例如：
  - what is the best prediction given some data?
    什么是给定某些数据的最佳预测？
  - what is the best model given some data?
    什么是给定某些数据的最优模型？
  - what measurement should I perform next?
    接下来应该执行什么度量？
- Probabilistic model is to use probability theory to express all forms of uncertainty.
  概率模型是采用概率论来表示所有不确定性的形式。
- Bayes rule allows us to infer unknown quantities, adapt our models, make predictions and learn from data.
  贝叶斯规则允许我们推断未知量，适配我们的模型，做出预测并从数据中学习。

Typical Probabilistic Methods 典型的概率方法

- Naive Bayes and Bayes Network — 朴素贝叶斯和贝叶斯网络
- Probabilistic program — 概率规划
- Gaussian processes — 高斯过程
- Hidden Markov models — 隐藏马可夫模型

Probability Used in Machine Learning 用于机器学习的概率

- Often the goal of machine learning is to determine the probability of an event, e.g., 机器学习的目标常常是要确定某个事件的概率，例如：

  P (a person is at risk for a disease)
  某人患某种疾病病的风险

- There are two main perspectives:
  有两个主要的学派：
  - Frequentist: probabilities are long-run frequencies, flip a coin a million times to determine if it’s fair.
    频率学派：概率是重复多次的频率，如投掷硬币一百万次以确定正反面出现频率是否相等。
  - Bayesian: probabilities quantify our uncertainty in events designed to get the closest to the truth given a specific set of data.
    贝叶斯学派：给定一组特定数据，概率对事件的不确定性进行量化，旨在最接近事实真值。

Bayes Rule 贝叶斯规则

- Bayes rule tells us how to do inference about hypotheses from data.
  贝叶斯规则告诉我们如何从数据推断假设。
- Learning and prediction can be seen as forms of inference.
  学习和预测可以看作是推理的形式。
- Bayes rule allows us to 贝叶斯规则允许我们

Thomas Bayes (1701-1761),
an English statistician, philosopher
托马斯·贝叶斯，英国统计学家、哲学家

infer unknown quantities，推断未知的量、 → **adapt our models 适配我们的模型、** → **make predictions, and 做出预测、以及** → **learn from data. 从数据中学习。**

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})\, P(\text{hypothesis})}{P(\text{data})}$$

Bayesian Machine Learning 贝叶斯机器学习

Everything follows from two simple rules:
任何事物都遵循如下**两个简单规则**：

Sum rule 求和规则： $P(x) = \sum_y P(x, y)$

Product rule 乘积规则： $P(x, y) = P(x)\, P(y \mid x)$

$P(\theta \mid ƛ, m)$ ： posterior probability of given data ƛ 给定数据D的后验概率

$P(ƛ \mid \theta, \mathrm{m})$ ： likelihood of parameters θ in model m 模型m中参数θ的似然性

$P(\theta \mid \mathrm{m})$ ： prior probability of θ θ的先验概率

$P(ƛ \mid \mathrm{m})$ ： evidence 证据

$$P(\theta \mid ƛ, m) = \frac{P(ƛ \mid \theta, m)\, P(\theta \mid m)}{P(ƛ \mid m)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Parametric vs Non-parametric Methods 参数与非参数方法

☐ Parametric Methods **参数方法**

- Assume some *finite set of parameters* θ. Given the parameters, future predictions *x*, are independent of the observed data D:

  设有限参数集 θ。给定该参数，则特征预测 x 独立于观测数据 D，即：

$$P(x \mid \theta, \lambda) = P(x \mid \theta)$$

  therefore capture everything there is to know about the data.

  因此，捕捉任何事物在于要知晓这些数据。

- So, the model's complexity is bounded even if the amount of data is unbounded.

  故，即使数据量是无限的，模型的复杂性也是有界的。

- This makes them not very flexible.

  这使得参数方法**不是很灵活**。

Parametric vs Non-parametric Methods 参数与非参数方法

Non-parametric Methods **非参数方法**

- Assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ.

  设数据分布不能用这样一个有限参数集来定义。但通常能够由假设的无限维度 θ 来定义。

  Usually we think of θ as a function.

  通常，我们把 θ 看作是一个函数。

- The amount of information that θ can capture about the data D can grow as the amount of data grows.

  θ 可以捕捉数据 D 的信息量可以随着数据量的增长而增长。

- This makes them more flexible.

  这使得非参数方法**更加灵活**。

Parametric vs Non-parametric Methods 参数与非参数方法

	Parametric 参数	Non-parametric 非参数
Brief Statements 简短陈述	The methods have a fixed number of parameters. 该方法具有**固定的**参数个数。	The number of parameters grow with the amount of training data. 参数个数随着训练数据的量而**增长**。
Typical Algorithm 典型算法	Support Vector Machine 支撑向量机	*k*-nearest neighbors *k* 近邻

Parametric vs Non-parametric Methods 参数与非参数方法

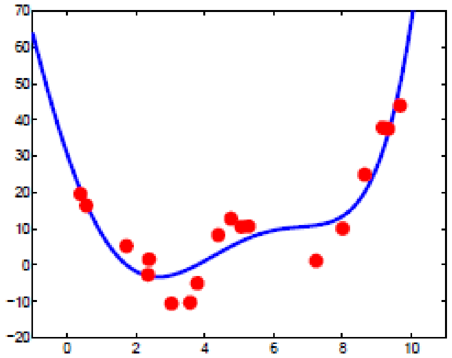
	Parametric 参数	Non-parametric 非参数
Feature 特征	Having a fixed number of parameters. 具有**固定的**参数个数	The number of parameters grow with the amount of training data. 参数的个数随着训练数据的量而**增长**。
Advantage 优点	Often being faster to use 通常**易于使用**	More flexible. 更加**灵活**
Disadvantage 缺点	Making stronger assumptions about the nature of the data distributions. 对数据分布的性质做出**严格的假设**。	Often computationally intractable for large datasets. 时常在计算上**难以应付大的数据集**。

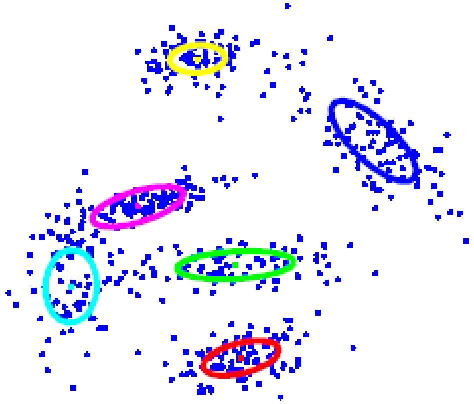
Why Non-parametric 为什么非参数

- Flexibility 灵活性
- Better predictive performance 更好的预测性能
- More realistic 更现实

Most methods are non-parametric 大多数算法是非参数的

- Kernel methods 核方法
  (SVM, Gaussian process 高斯过程)
- Deep networks 深度网络
- *k*-nearest neighbors *k* 近邻
- etc.





Non-parametric Methods and Uses 非参数方法及用途

- Bayesian non-parametric has many uses.
  贝叶斯非参数有许多用途。

Some modelling goals and examples of associated non-parametric Bayesian models.
一些建模目标和与贝叶斯非参数模型关联的实例

Non-parametric Methods 非参数方法	Examples 示例
Distributions on functions 功能分布	Gaussian process 高斯过程
Clustering 聚类	Chinese restaurant process 中餐馆过程
Sparse binary matrices 稀疏二元矩阵	Indian buffet processes 印度自助餐过程
Survival analysis 生存分析	Beta processes 贝塔过程
Distributions on measures 措施分布	Completely random measures 完全随机措施

Case Study: Probabilistic Program Induction 概率规划归纳

Brenden Lake, Ruslan Salakhutdinov, Joshua Tenenbaum, “Human-level concept learning through probabilistic program induction”凭借概率规划归纳法的人类层级概念学习, *Science*, Vol. 350, Dec. 2015.

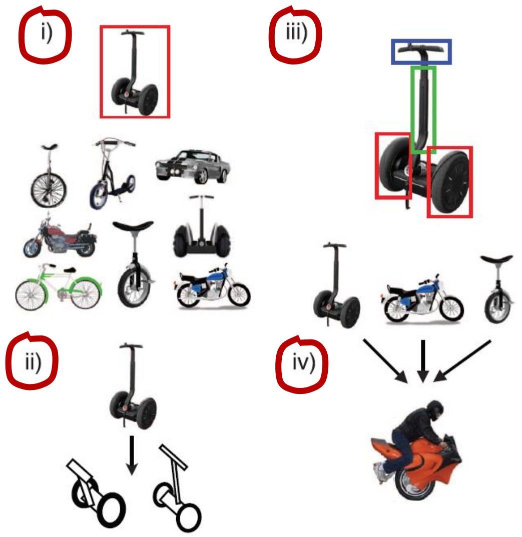
A single example of a new concept can be enough information to support:
一个新概念的单一实例可以有足够的信息来支持：

i. classification of new examples,
新样本的分类，

ii. generation of new examples,
新样本的生成，

iii. parsing an object into parts and relations,
将一个物体分成部件和关系，

iv. generation of new concepts from related concepts.
从相关概念生成新的概念。



People can learn rich concepts from limited data.
人们可以从有限数据中学习丰富的概念。

Case Study: Probabilistic Program Induction 概率规划归纳

☐ Bayesian Program Learning (BPL) **贝叶斯规划学习 (BPL)**

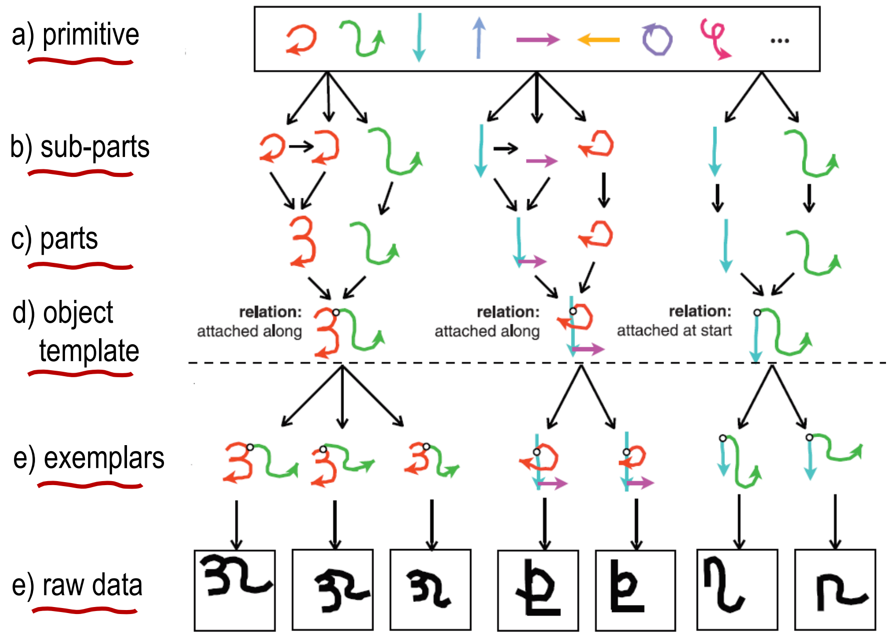
- BPL learns simple stochastic programs to represent concepts, and builds new types of concepts compositionally from parts, sub-parts, and spatial relations.
  BPL学习简单的随机规划来表示概念，并从部件、子部件、以及空间关系的组合来构建新类型的概念。
- Each new type is also represented as a generative model, and this lower-level generative model produces new examples (or tokens) of the concept.
  每个新类型还可以表示为一个生成模型，并且这个低层的生成模型产生概念的新样本 (或记号)。
- The joint distribution on types ψ, a set of *M* tokens $\theta^{(1)}, \ldots, \theta^{(M)}$, and the corresponding binary images $I^{(1)}, \ldots, I^{(M)}$ factors as
  联合分布，其参数：类型 ψ，一组 *M* 个记号 $\theta^{(1)}, \ldots, \theta^{(M)}$ ，和对应的二值图像 $I^{(1)}, \ldots, I^{(M)}$因子，即：

$$P(\psi, \theta^{(1)}, \ldots, \theta^{(M)}, I^{(1)}, \ldots, I^{(M)}) = P(\psi)\prod_{m=1}^{M} P(I^{(m)}|\theta^{(m)})P(\theta^{(m)}|\psi)$$

Case Study: Probabilistic Program Induction 概率规划归纳

One-shot learning 一次性学习

It incorporates three principles: compositionality, causality, and learning to learn.
它包含三个基本原理：**组合性、因果性、和学会学习**。



A generative model of handwritten characters.
一个手写字符的生成模型

a) generate new types by primitive,
由原语生成新的类型
b) extract sub-parts,
提取子部件
c) make parts,
生成部件
d) combine parts with relations to define object template,
根据关系组合部件来定义对象模板
e) generate new exemplars by running these templates,
通过运行这些模板生成新的样例
f) render them as raw data.
将其渲染成原始数据

Thank you for your attention!

AI