《数据挖掘方法》

⭐ CHX 统计学习方法总结

➡️ Create by *Wang JingHui*

➡️ Last Revision Time: *2021.05.16*

主要内容

1. 有监督学习
2. 无监督学习
3. 机器学习算法总结
4. 机器学习相关课程

有监督学习方法

方法=模型+策略+算法

监督学习, 非监督学习, 强化学习都有这样的三要素.

1. 模型
    i. 监督学习中, 模型就是所要学习的条件概率分布或者决策函数.

2. 策略
    i. 统计学习的目标在于从假设空间中选取最优模型.
    ii. 损失函数度量一次预测的好坏; **风险函数**度量平均意义下模型预测的好坏.
    iii. 经验风险最小化(ERM)与结构风险最小化(SRM)
    iv. 经验风险或者结构风险是最优化的目标函数.
3. 算法
    i. 统计学习基于训练数据集, 根据学习策略, 从假设空间中选择最优模型, 最后需要考虑用干什么样的计算方法求解**最优模型**.
    ii. 统计学习问题转化为最优化问题.
        a. 有显式解析解, 对应的最优化问题比较简单
        b. 通常解析解不存在, 需要通过数值计算的方式求解.
    iii. 算法需要解决的问题是如何找到**全局最优解**, 并且求解的过程非常高效.

模型

分类问题与标注问题都可以认为是从输入空间到输出空间的映射.

他们可以写成条件概率分布$P(Y|X)$或者决策函数$Y = f(x)$的形式.

概率模型和非概率模型

对应**概率模型**和**非概率模型**.

生成模型和判别模型

1. 生成模型
    - 直接学习条件概率分布$P(Y|X)$或者决策函数$Y = f(X)$的方法为判别方法, 对应的模型为判别模型.
    - 感知机, k近邻, 决策树, 逻辑斯谛回归模型, 最大熵模型, 支持向量机, 提升方法, 条件随机场
2. 判别模型
    - 先学习联合概率分布$P(X, Y)$, 从而求得条件概率分布$P(Y|X)$的方法是生成方法, 对应的模型是生成模型.
    - 朴素贝叶斯, 隐马尔科夫模型

线性模型和非线性模型

1. 线性模型
    - 感知机
2. 对数线性模型
    - 逻辑斯谛回归模型
    - 最大熵模型
    - 条件随机场
3. 非线性模型
    - k近邻
    - 决策树
    - 支持向量机(核函数)
    - 提升方法

策略

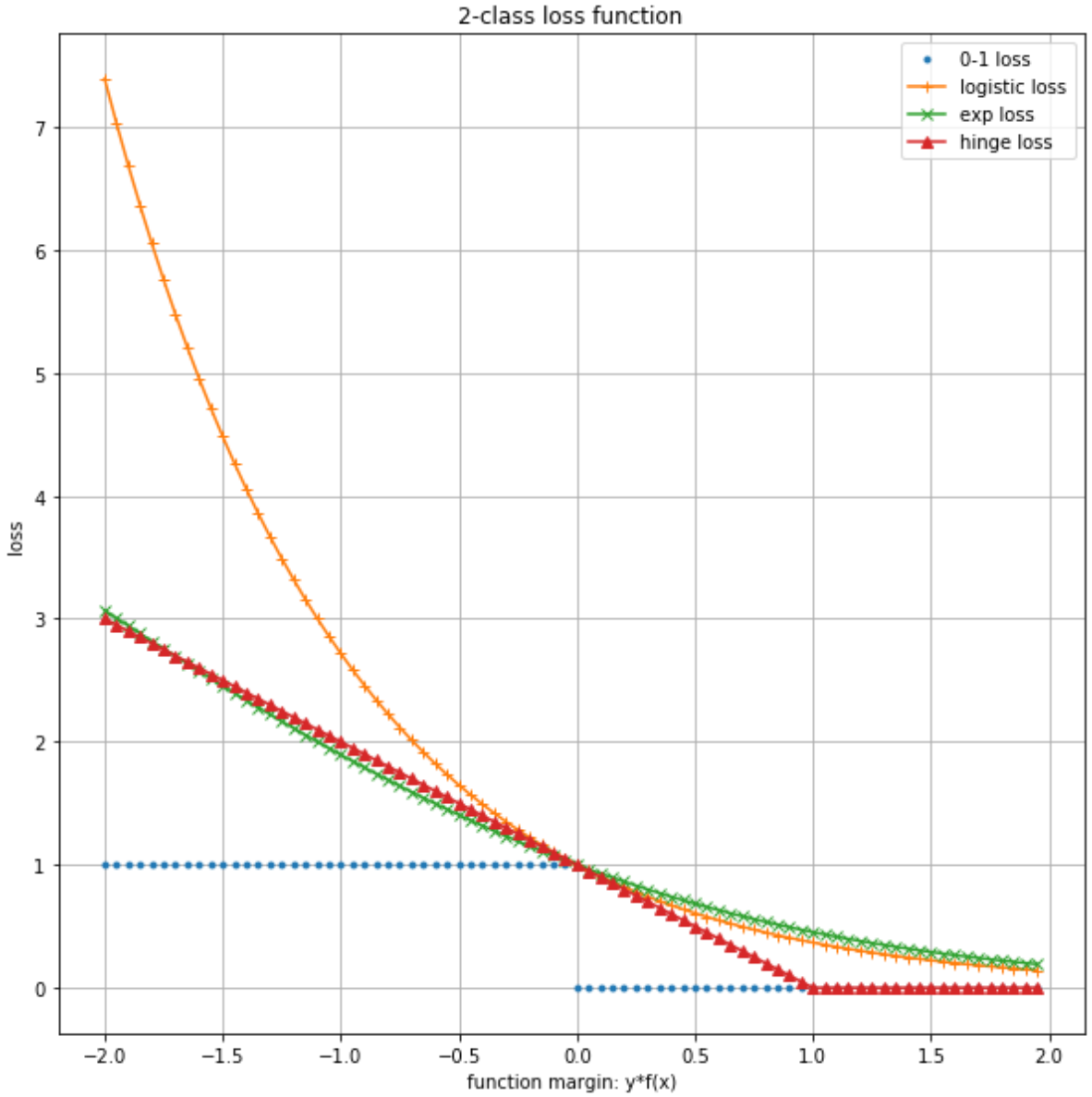
损失函数

1. 合页损失
   支持向量机 $max(0, 1 - yf(x))$
2. 逻辑斯谛损失函数
   逻辑斯谛回归模型与最大熵模型 $log(1 + exp(-yf(x)))$
3. 指数损失函数
   提升方法 $exp(-yf(x))$

三种损失函数都是0-1损失函数的上界.

上面这个图有几点要注意:

- logistic loss，里面的对数是2。
- 这些函数在0右侧的部分，都是有值的。
- 分类问题的损失，实现二分类任务



正则化方法

- 正则化方法根据算法的复杂度对算法进行调整，正则化方法通常对简单模型予以奖励而对复杂算法予以惩罚。
- 常见的算法包括：Ridge Regression， Least Absolute Shrinkage and Selection Operator（LASSO），以及弹性网络（Elastic Net）。

- 在分类问题上，引入**经验风险最小化**和**结构风险最小化**。
- 学习的策略是优化以下结构风险函数

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

> 第一项为经验风险(经验损失)，第二项为正则化项。
> 提升方法没有显式的正则化项，可以通过early stop控制停止

算法

梯度下降法

- 梯度下降法（Gradient Descent） 梯度下降法是最早最简单，也是最为常用的最优化方法。梯度下降法实现简单，当目标函数是凸函数时，梯度下降法的解是全局解。一般情况下，其解不保证是全局最优解，梯度下降法的速度也未必是最快的。梯度下降法的优化思想是用当前位置负梯度方向作为搜索方向，因为该方向为当前位置的最快下降方向，所以也被称为是"最速下降法"。最速下降法越接近目标值，步长越小，前进越慢。

牛顿法

- 牛顿法是一种在实数域和复数域上近似求解方程的方法，使用函数f(x)的泰勒级数的前面几项来寻找方程f(x) = 0的根。牛顿法最大的特点就在于它的收敛速度很快。
- 牛顿法优缺点
    - 优点：二阶收敛，收敛速度快；
    - 缺点：牛顿法是一种迭代算法，每一步都需要求解目标函数的Hessian矩阵的逆矩阵，计算比较复杂。
- 梯度下降法和牛顿法的比较
    - 从本质来说，梯度下降法是一阶收敛，牛顿法是二阶收敛，所以牛顿法的收敛速度更快。牛顿法是用一个二次曲面去拟合当前所处位置的局部曲面，而梯度下降法使用一个平面去拟合当前的局部曲面，通常情况二次曲面拟合会比平面更好，所以牛顿法的下降路径会更符合真实的最优下降路径。

拟牛顿法（DFP、BFGS）

- 拟牛顿法的本质思想是改善牛顿法每次需要求解复杂的Hessian矩阵的逆矩阵的缺陷，它使用正定矩阵来近似Hessian矩阵的逆，从而简化了运算的复杂度。拟牛顿法和最速下降法一样只要求每一步迭代时知道目标函数的梯度。通过测量梯度的变化，构造一个目标函数的模型使之足以产生超线性收敛性。这类方法大大优于最速下降法，尤其对于困难的问题。另外，因为拟牛顿法不需要二阶导数的信息，所以有时比牛顿法更为有效。如今，优化软件中包含了大量的拟牛顿算法用来解决无约束，约束，和大规模的优化问题。

共轭梯度法（Conjugate Gradient）

- 共轭梯度法是介于最速下降法与牛顿法之间的一个方法，它仅需利用一阶导数信息，但克服了最速下降法收敛慢的缺点，又避免了牛顿法需要存储和计算Hesse矩阵并求逆的缺点，共轭梯度法不仅是解决大型线性方程组最有用的方法之一，也是解大型非线性最优化最有效的算法之一。在各种优化算法中，共轭梯度法是非常重要的一种。其优点是所需存储量小，具有步收敛性，稳定性高，而且不需要任何外来参数。

启发式优化方法

- 启发式方法是指人在解决优化问题时所采取的一种根据经验规则进行发现的方法。其特点是在解决问题时，利用过去的经验，选择已经行之有效的方法，而不是系统地、以确定的步骤去寻求答案。启发式优化方法种类繁多，包括经典的模拟退火方法，遗传算法、蚁群算法以及粒子群算法等等。
- 还有一种特殊的优化算法被称之多目标优化算法，它主要针对同时优化多个目标（两个及两个以上）的优化问题，这方面比较经典的算法有NSGAII算法、MOEA/D算法以及人工免疫算法等。

EM算法

- EM算法是一类算法的总称。EM算法分为E-step和M-step两步。EM算法的应用范围很广，基本机器学习需要迭代优化参数的模型在优化时都可以使用EM算法。
- EM算法结果不一定能保证获得全局最优解，但如果我们优化的目标函数是一个凸函数，那么一定能保证得到全局最优解。否则可能获得局部最优解。因为如果优化的目标函数有多个峰值点，则如果优化到某个不是最高的峰值点处，则会无法再继续下去，这样获得的是局部最优解。

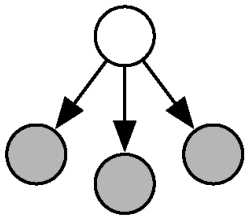
机器学习算法

1. 朴素贝叶斯法CH04和隐马尔科夫模型CH10
2. 感知机CH02, 逻辑斯谛回归模型CH06, 最大熵模型CH06, 条件随机场CH11
3. 支持向量机CH07
4. 决策树CH05
5. 提升方法CH08
6. EM算法CH09
7. NB和HMM的监督学习，最优解就是极大似然估计值，可以由概率计算公式直接计算。之前看NB其实就是计数查表，这种要有大的语料库进行统计，所谓学的多，就知道的多。

不同视角-从多个角度进行划分

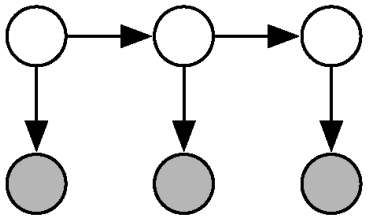
1. 简单分类方法
    i. 感知机
    ii. k近邻法
    iii. 朴素贝叶斯法
    iv. 决策树
2. 复杂分类方法
    i. 逻辑斯谛回归模型
    ii. 最大熵
    iii. 支持向量机
    iv. 提升方法
3. 标注方法
    隐马尔科夫模型 / 条件随机场

不同视角-从多个角度进行划分



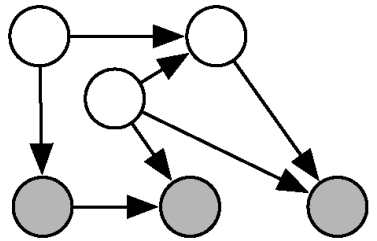
Naive Bayes

SEQUENCE



HMMs

GENERAL GRAPHS

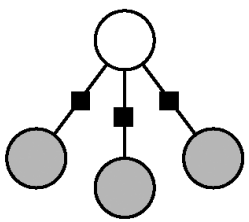


Generative directed models

CONDITIONAL

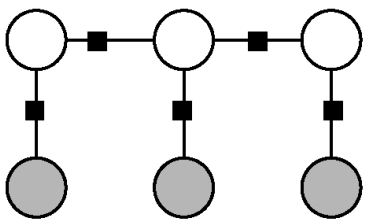
CONDITIONAL

CONDITIONAL



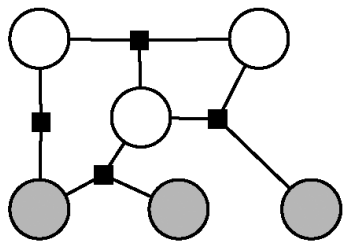
Logistic Regression

SEQUENCE



Linear-chain CRFs

GENERAL GRAPHS



General CRFs

方法	适用问题	模型特点	模型类型	学习策略	学习的损失函数	学习算法
感知机	二类分类	分离超平面	判别模型	极小化误分点到超平面距离	误分点到超平面距离	随机梯度下降
k 近邻法	多类分类，回归	特征空间，样本点	判别模型			
朴素贝叶斯法	多类分类	特征与类别的联合概率分布，条件独立假设	生成模型	极大似然估计，极大后验概率估计	对数似然损失	概率计算公式，EM 算法
决策树	多类分类，回归	分类树，回归树	判别模型	正则化的极大似然估计	对数似然损失	特征选择，生成，剪枝
逻辑斯谛回归与最大熵模型	多类分类	特征条件下类别的条件概率分布，对数线形模型	判别模型	极大似然估计，正则化的极大似然估计	逻辑斯谛损失	改进的迭代尺度算法，梯度下降，拟牛顿法
支持向量机	二类分类	分离超平面，核技巧	判别模型	极小化正则化合页损失，软间隔最大化	合页损失	序列最小最优化算法（SMO）
提升方法	二类分类	弱分类器的线性组合	判别模型	极小化加法模型的指数损失	指数损失	前向分步加法算法
EM 算法[8]	概率模型参数估计	含隐变量概率模型		极大似然估计，极大后验概率估计	对数似然损失	迭代算法
隐马尔可夫模型	标注	观测序列与状态序列的联合概率分布模型	生成模型	极大似然估计，极大后验概率估计	对数似然损失	概率计算公式，EM 算法
条件随机场	标注	状态序列条件下观测序列的条件概率分布，对数线性模型	判别模型	极大似然估计，正则化极大似然估计	对数似然损失	改进的迭代尺度算法，梯度下降，拟牛顿法

算法思想

基于实例的算法

- 基于实例的算法常常用来对决策问题建立模型，这样的模型常常先选取一批样本数据，然后根据某些近似性把新数据与样本数据进行比较。通过这种方式来寻找最佳的匹配。因此，基于实例的算法常常也被称为“赢家通吃”学习或者“基于记忆的学习”。常见的算法包括 k-Nearest Neighbor(KNN)， 学习矢量量化（Learning Vector Quantization， LVQ），以及自组织映射算法（Self-Organizing Map ， SOM）

决策树学习

- 决策树算法根据数据的属性采用树状结构建立决策模型， 决策树模型常常用来解决分类和回归问题。
- 常见的算法包括：分类及回归树（Classification And Regression Tree， CART），ID3 (Iterative Dichotomiser 3)， C4.5， Chi-squared Automatic Interaction Detection(CHAID)， Decision Stump， 随机森林（Random Forest）， 多元自适应回归样条（MARS）以及梯度推进机（Gradient Boosting Machine， GBM）

贝叶斯方法

- 贝叶斯方法算法是基于贝叶斯定理的一类算法，主要用来解决分类和回归问题。常见算法包括：朴素贝叶斯算法，平均单依赖估计（Averaged One-Dependence Estimators， AODE），以及Bayesian Belief Network（BBN）。

基于核的算法

- 基于核的算法中最著名的是支持向量机（SVM）了。 基于核的算法把输入数据映射到一个高阶的向量空间， 在这些高阶向量空间里， 有些分类或者回归问题能够更容易的解决。 常见的基于核的算法包括：支持向量机（Support Vector Machine, SVM）， 径向基函数（Radial Basis Function ， RBF)， 以及线性判别分析（Linear Discriminate Analysis ， LDA)等

关联规则学习

- 关联规则学习通过寻找最能够解释数据变量之间关系的规则，来找出大量多元数据集中有用的关联规则。常见算法包括 Apriori算法和Eclat算法等。

人工神经网络

- 人工神经网络算法模拟生物神经网络，是一类模式匹配算法。通常用于解决分类和回归问题。人工神经网络是机器学习的一个庞大的分支，有几百种不同的算法。重要的人工神经网络算法包括：感知器神经网络（Perceptron Neural Network）， 反向传递（Back Propagation）， Hopfield网络，自组织映射（Self-Organizing Map，SOM）。学习矢量量化（Learning Vector Quantization， LVQ）

深度学习

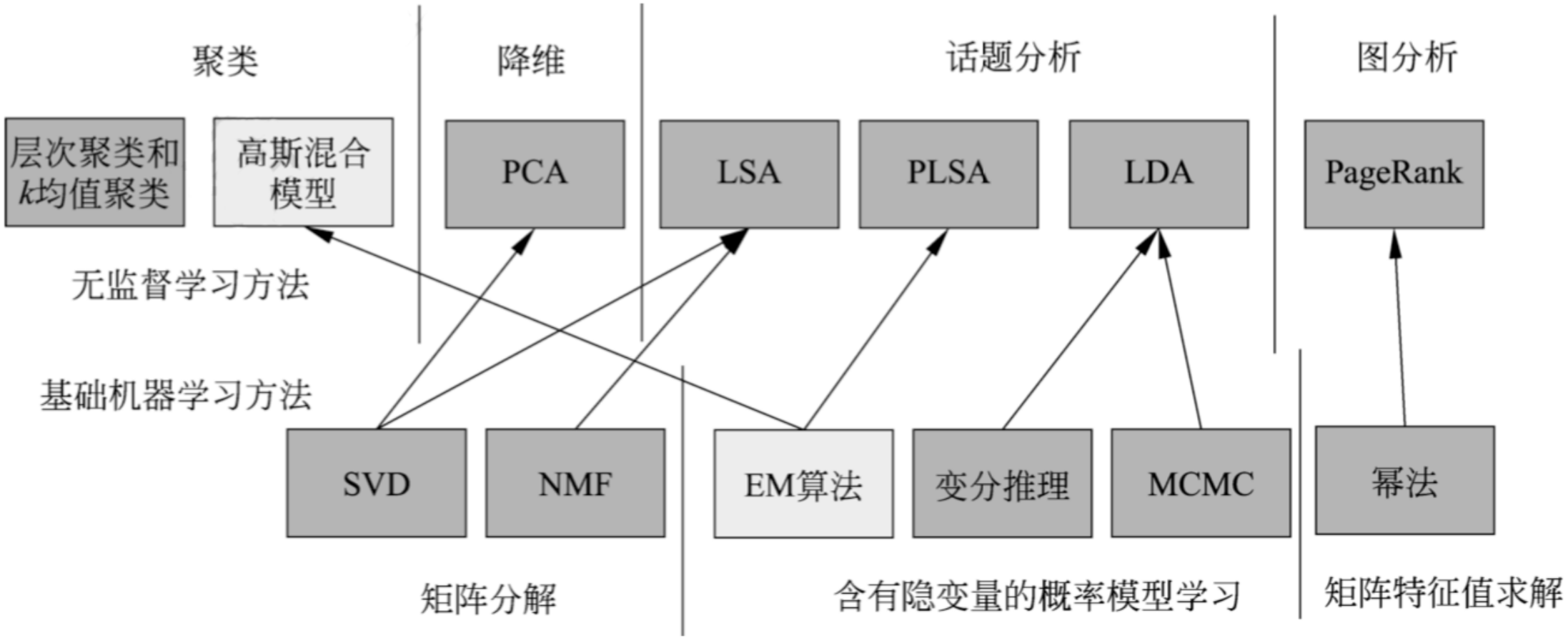
- 深度学习算法是对人工神经网络的发展。 在计算能力变得日益廉价的今天，深度学习试图建立大得多也复杂得多的神经网络，很多深度学习的算法是半监督式学习算法，用来处理存在少量未标识数据的大数据集。常见的深度学习算法包括：受限波尔兹曼机（Restricted Boltzmann Machine， RBN）， Deep Belief Networks（DBN），卷积网络（Convolutional Network）， 堆栈式自动编码器（Stacked Auto-encoders）。

集成算法：

- 集成算法用一些相对较弱的学习模型独立地就同样的样本进行训练，然后把结果整合起来进行整体预测。集成算法的主要难点在于究竟集成哪些独立的较弱的学习模型以及如何把学习结果整合起来。这是一类非常强大的算法，同时也非常流行。
- 常见的算法包括：Boosting， Bootstrapped Aggregation（Bagging）， AdaBoost，堆叠泛化（Stacked Generalization， Blending），梯度推进机（Gradient Boosting Machine， GBM），随机森林（Random Forest）。

无监督学习学习方法

- 无监督学习是指从无标注数据中学习模型的机器学习问题。无标注数据是自然得到的数据，模型表示数据的类别、转换或概率无监督学习的本质是学习数据中的统计规律或潜在结构，主要包括聚类、降维、概率估计。
- 无监督学习可以用于对已有数据的分析，也可以用于对未来数据的预测。学习得到的模型有函数$z=g(x)$，条件概率分布$P(z|x)$，或条件概率分布$P(x|z)$。
- 无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构，假定损失最小的压缩得到的结果就是最本质的结构。可以考虑发掘数据的纵向结构，对应聚类。也可以考虑发掘数据的横向结构，对应降维。还可以同时考虑发掘数据的纵向与横向结构，对应概率模型估计。


聚类
降维
话题分析
图分析
层次聚类和k均值聚类
高斯混合模型
PCA
LSA
PLSA
LDA
PageRank
无监督学习方法
基础机器学习方法
SVD
NMF
EM算法
变分推理
MCMC
幂法
矩阵分解
含有隐变量的概率模型学习
矩阵特征值求解

聚类

- 聚类是将样本集合中相似的样本（实例）分配到相同的类，不相似的样本分配到不同的类。聚类分硬聚类和软聚类。聚类方法有层次聚类和k均值聚类。
- 降维是将样本集合中的样本（实例）从高维空间转换到低维空间。假设样本原本存在于低维空间，或近似地存在于低维空间，通过降维则可以更好地表示样本数据的结构，即更好地表示样本之间的关系。降维有线性降维和非线性降维，降维方法有主成分分析。

概率模型估计

- 概率模型估计假设训练数据由一个概率模型生成，同时利用训练数据学习概率模型的结构和参数。概率模型包括混合模型、率图模型等。概率图模型又包括有向图模型和无向图模型。
- 话题分析是文本分析的一种技术。给定一个文本集合，话题分析旨在发现文本集合中每个文本的话题，而话题由单词的集合表示。话题分析方法有潜在语义分析、概率潜在语义分析和潜在狄利克雷分配。
- 图分析的目的是发掘隐藏在图中的统计规律或潜在结构。链接分析是图分析的一种，主要是发现有向图中的重要结点，包括 PageRank算法。

降维

降维方法分为线性和非线性降维，非线性降维又分为基于核函数和基于特征值的方法。

- 线性降维方法：PCA 、ICA LDA、LFA、LPP(LE的线性表示)
- 非线性降维方法：
  （1）基于核函数的非线性降维方法：KPCA 、KICA、KDA
  （2）基于特征值的非线性降维方法（流型学习）：ISOMAP、LLE、LE、LPP、LTSA、MVU

图分析

- 图算法是图分析的工具之一。图算法提供了一种最有效的分析连接数据的方法，它们描述了如何处理图以发现一些定性或者定量的结论。
- 图算法基于图论，利用节点之间的关系来推断复杂系统的结构和变化。
- 可以使用这些算法来发现隐藏的信息，验证业务假设，并对行为进行预测。
- 图分析和图算法具有广泛的应用潜力：从防止欺诈，优化呼叫路由，

无监督学习方法的特点

	方法	模型	策略	算法
聚类	层次聚类	聚类树	类内样本距离最小	启发式算法
	k 均值聚类	k 中心聚类	样本与类中心距离最小	迭代算法
	高斯混合模型	高斯混合模型	似然函数最大	EM 算法
降维	PCA	低维正交空间	方差最大	SVD
话题分析	LSA	矩阵分解模型	平方损失最小	SVD
	NMF	矩阵分解模型	平方损失最小	非负矩阵分解
	PLSA	PLSA 模型	似然函数最大	EM 算法
	LDA	LDA 模型	后验概率估计	吉布斯抽样，变分推理
图分析	PageRank	有向图上的马尔可夫链	平稳分布求解	幂法

机器学习算法总结

监督式学习：

- 在监督式学习下，输入数据被称为“训练数据”，每组训练数据有一个明确的标识或结果。

非监督式学习：

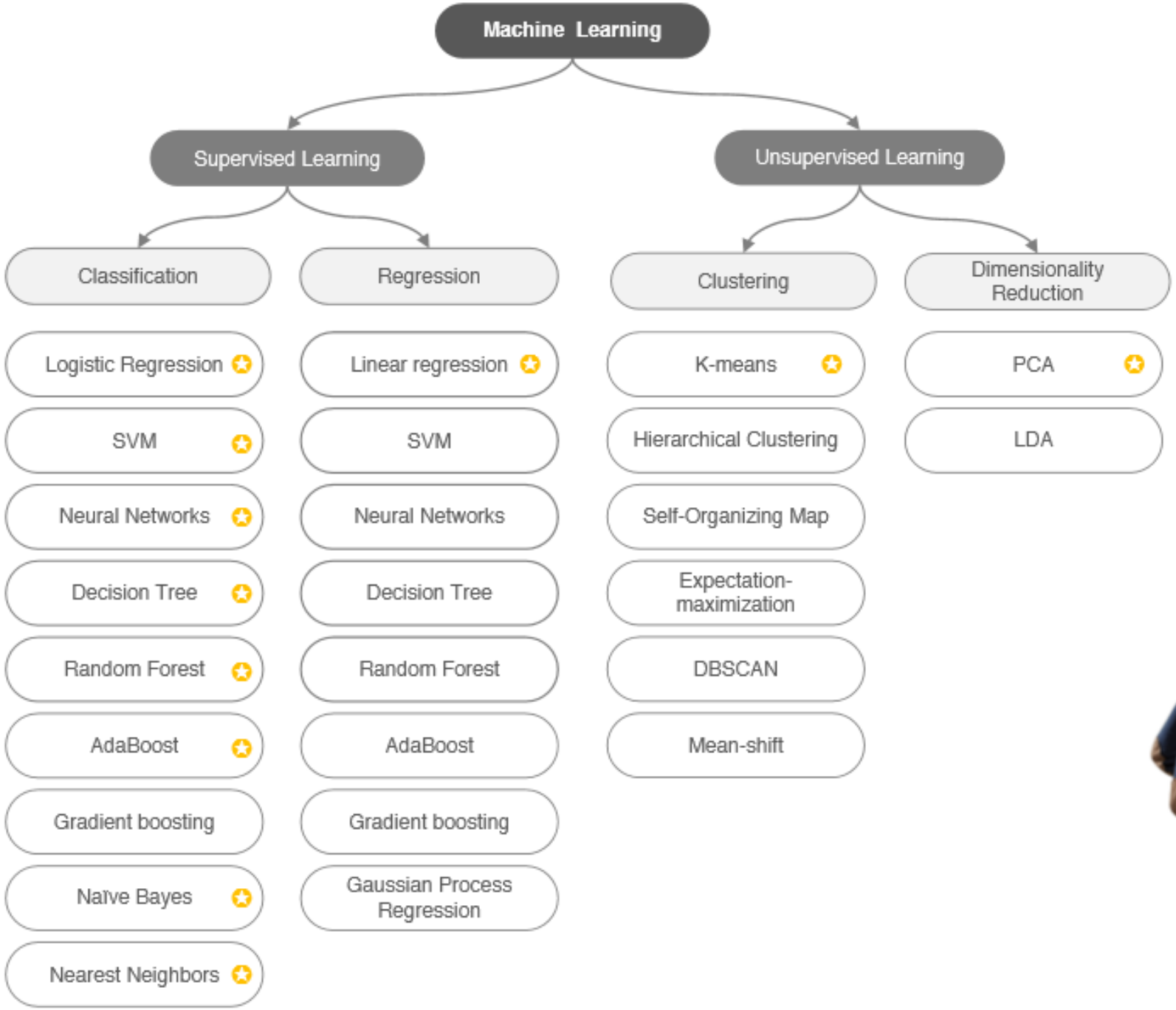
- 在非监督式学习中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构。常见的应用场景包括关联规则的学习以及聚类等。常见算法包括Apriori算法以及k-Means算法。

半监督式学习：

- 在此学习方式下，输入数据部分被标识，部分没有被标识，这种学习模型可以用来进行预测，但是模型首先需要学习数据的内在结构以便合理的组织数据来进行预测。应用场景包括分类和回归，算法包括一些对常用监督式学习算法的延伸，这些算法首先试图对未标识数据进行建模，在此基础上再对标识的数据进行预测。如图论推理算法（Graph Inference）或者拉普拉斯支持向量机（Laplacian SVM.）等。

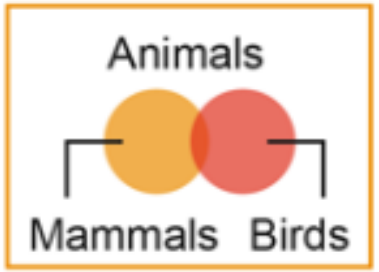
强化学习：

- 在这种学习模式下，输入数据作为对模型的反馈，不像监督模型那样，输入数据仅仅是作为一个检查模型对错的方式，在强化学习下，输入数据直接反馈到模型，模型必须对此立刻作出调整。常见的应用场景包括动态系统以及机器人控制等。常见算法包括Q-Learning以及时间差学习（Temporal difference learning）


Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering
Dimensionality Reduction
Logistic Regression
SVM
Neural Networks
Decision Tree
Random Forest
AdaBoost
Gradient boosting
Naïve Bayes
Nearest Neighbors
Linear regression
SVM
Neural Networks
Decision Tree
Random Forest
AdaBoost
Gradient boosting
Gaussian Process Regression
K-means
Hierarchical Clustering
Self-Organizing Map
Expectation-maximization
DBSCAN
Mean-shift
PCA
LDA

What are the five tribes?

Symbolists



Use symbols, rules, and logic to represent knowledge and draw logical inference

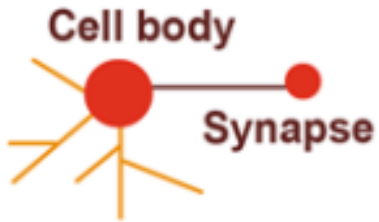
Favored algorithm
Rules and decision trees

Bayesians

Assess the likelihood of occurrence for probabilistic inference

Favored algorithm
Naive Bayes or Markov

Connectionists



Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm
Neural networks

Evolutionaries

Generate variations and then assess the fitness of each for a given purpose

Favored algorithm
Genetic programs

Analogizers

Optimize a function in light of constraints ("going as high as you can while staying on the road")

Favored algorithm
Support vectors

Source: Pedro Domingos, *The Master Algorithm,* 2015

机器学习相关课程

- 数值优化课程，参考教材乔治·诺塞达尔（Jorge Nocedal）和史蒂芬·赖特（Stephen J. Wright）的第二版《数值优化》（Numerical Optimization），或者开设数值分析，建议采用蒂莫西·索尔的《数值分析》（Numerical Analysis）为教材。
- 算法课程，参考教材是迈克尔·米曾马克（Michael Mitzenmacher）和伊莱·阿普法（Eli Upfal）的《概率与计算：随机算法与概率分析》（Probability and Computing： Randomized Algorithms and Probabilistic Analysis）。

- 在程序设计方面，增加或加强并行计算的内容。特别是在深度学习技术的执行中，通常需要GPU加速，可以使用戴维·柯克 （David B. Kirk） 和胡文美（Wen-mei W. Hwu）的教材 《大规模并行处理器编程实战》（第二版）（Programming Massively Parallel Processors： A Hands-on Approach， Second Edition）；可以参考优达学城（Udacity）上英伟达（Nvidia）讲解CUDA计算的公开课。
- 以计算机科学为主导，联合统计和应用数学专业，围绕理论机器学习、概率与随机图模型、贝叶斯方法、大规模优化算法、深度学习等基础机器学习领域。建议开设理论机器学习、概率图模型、统计推断与贝叶斯分析、凸分析与优化、强化学习、信息论等课程。

Don't be evil!

Thank you!

----- WangJingHui

Enjoy your machine learning!

https://github.com/wjssx/Statistical-Learning-Slides-Code

E-mail: csr_dsp@sina.com