前言

第1章 统计学习方法概论

1.1 统计学习

统计学习的特点

统计学习（statistical learning）：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与分析的一门学科。 因此统计学习也称为统计机器学习（statistical machine learning).

统计学习的主要特点

- 理论基础
  - 数学基础：微积分、线性代数、概率论、统计学、计算理论、最优化理论
  - 其他基础：信息论、计算机科学及应用相关的科学等多个领域的交叉学科
- 应用基础：以计算机及网络为平台
- 研究对象：数据
  - 统计学习是数据驱动的学科；
- 研究目的：对数据进行分类和预测；
- 研究手段：通过统计学习方法构建模型，并应用模型进行分类和预测；

统计学习的对象

统计学习的对象是数据（data）:从数据出发,提取数据的“特征”,抽象出数据的“模型”，发现数据中的“知识”，又回到对数据的“分类”与“预测”中。

数据的基本假设：

- 同类数据具有一定的统计规律性，所以可以用概率统计方法加以处理。

数据的基本分类：

- 连续型数据
- 离散型数据：本书主要关注的是“离散型”数据。

统计学习的目的

- 模型：学习什么样的模型
- 策略：如何学习模型 → 使模型能够对数据进行准确地分类和预测
- 算法：如何提高模型的学习效率

统计学习的方法

统计学习方法的三个要素 ：模型、策略、算法

统计学习方法的分类：

- 有监督学习 (supervised learning) （全书重点）
  - 从给定的、有限的、用于学习的训练数据 (training data) 集合出发；
    - 假设数据是独立同分布产生的；
  - 基于某个评价标准 (evaluation criterion)，从假设空间中选取一个最优的模型
    - 模型（model）：假设要学习的模型属于某个函数的集合，称为假设空间 (hypothesis space);
    - 策略（strategy）：使模型在给定的评价准则下，对已知的训练数据及未知的测试数据 (test data) 都有最优的预测；
    - 算法（algorithm）：最优模型的选取都由算法实现。
- 无监督学习 (unsupervised learning):
- 半监督学习 (semi-supervised learning):

- 强化学习 (reinforcement learning):

实现统计学习方法的步骤

- 数据：得到一个有限的、用于训练的数据集合；
- 统计学习方法的三要素：
  - 模型 (model): 模型的假设空间；
  - 策略 (strategy): 模型选择的准则；
  - 算法 (algorithm): 模型学习的算法。
- 寻优：通过学习的方式选择出最优模型；
- 预测：利用学习的最优模型对新数据进行分类或预测。

统计学习的研究

- 统计学习方法 (statistical learning method): 开发新的学习方法；
- 统计学习理论 (statistical learning theory): 探求统计学习方法的有效性与效率，以及统计学习的基本理论问题；
- 统计学习应用 (application of statistical learning): 将统计学习方法应用到实际问题中去，解决实际问题。

统计学习的重要性

- 是处理海量数据的有效方法；
- 是计算机智能化的有效手段；
- 是计算机科学发展的重要组成。

1.2 监督学习

监督学习的任务：是学习一个模型，使模型能够对任意给定的输入，及其相应的输出做出一个好的预测

基本概念

输入空间：输入数据所有可能取值的集合；集合中元素的个数可以有限，也可以是整个空间

输出空间：输出数据所有可能取值的集合；集合中元素的个数可以有限，也可以是整个空间

假设空间：由输入空间到输出空间的映射的集合，即可供选择的模型的集合构成的空间，空间的确定意味着学习范围的确定

特征空间：所有特征向量存在的空间，每个特征对应特征空间中的一个维度。

- 特征向量（feature vector）：表示每个具体输入的实例（instance）
- 输入空间有时直接作为特征空间使用，输入空间中的数据已经具备良好的特征表示
- 输入空间有时需要变换到特征空间中，将输入空间中的数据进行线性或者非线性变换，从而使得特征数据线性具有线性可分的特性或者其他特性。

统计学习中的有监督学习根据“解决的问题”主要包括

- 分类问题：判别模型，处理离散数据
- 预测问题：回归模型，处理连续数据
- 标注问题：既是分类问题的推广，又是预测问题的简化。

统计学习中的有监督学习根据“输入变量”和“输出变量”的不同主要包括

- 分类问题：输出变量为有限个离散变量的预测问题；
- 回归问题：输入变量与输出变量均为连续变量；
- 标注问题：输入变量与输出变量均为变量序列的预测问题；

联合概率分布：输入变量与输出变量遵循联合分布；

问题的形式化描述

在学习过程中，学习系统（也就是学习算法）试图通过给定的训练数据集合中的样本带来的信息来学习得到模型，再基于模型对测试样本集合进行预测来对模型的质量进行评价。

1.3 统计学习三个要素

统计学习方法 = 模型 + 策略 + 算法

模型

主要问题：学习什么样的模型？

模型的假设空间：包含所有可能的条件概率分布或决策函数，即由一个参数向量决定的函数族，也称为参数空间 (parameter space)。

模型分类

- 非概率模型：由决策函数表示的模型；
- 概率模型：由条件概率表示的模型；

策略

主要问题：按照什么样的准则，学习得到最优的模型，或者从假设空间中选择最优的模型。

基本概念：

- 损失函数 (loss function) 或代价函数 (cost function): 度量模型一次预测的好坏
  - 损失函数值越小，模型就越好。
  - 风险函数（risk function）或者期望损失（expected loss）：损失函数的期望是理论上

模型 f(X)关于联合分布 P(X,Y)的平均意义下的损失

- 风险函数（risk function）或期望损失（expected loss）：度量平均意义下模型预测的好坏
- 经验风险（empirical risk）或经验损失（empirical loss）：表示模型与训练数据的破例程度，即模型训练样本集的平均损失，当样本容量趋于无穷时，经验风险逼近期望风险
- 结构风险（structural risk）：表示模型先验知识
  - 模型复杂度的正则化项（regularizer）或惩罚项（penalty term）

常用的损失函数：

- 0-1 损失函数：L(Y,f(X))=0,Y=f(X);L(Y,f(X))=1,Y≠f(X)
- 平方损失函数：L(Y,f(X))=(Y-f(X))^2
- 绝对值损失函数：L(Y,f(X))=|Y-f(X)|
- 对数损失函数或对数似然损失函数:L(Y,P(Y|X))=-log P(Y|X)

学习目标

- 理想状态：就是选择期望风险或期望损失最小的模型，希望可以提供无限的数据训练；
- 现实状态：就是选择经验风险或经验损失最小的模型，因为只能提供有限的数据训练；

经验风险矫正：当样本容量过小时，容易出现“过拟合”问题，所以需要对经验风险进行矫正，经验风险最小化 + 结构风险最小化

- 经验风险最小化（empirical risk minimization, ERM）：
  - 基于经验风险最小化求解最优模型就是求解最优化问题
  - 当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化等价于极大似然估计
- 结构风险最小化（structural risk minimization, SRM）：解决“过拟合”问题
  - 结构风险最小化主要手段就是正则化（regularization）

- ◆ 正则化：为模型增加表示模型复杂度的正则化项或者惩罚项
- 当模型是条件概率分布，损失函数是对数损失函数时，模型复杂度由模型的先验概率表示时，结构风险最小化等价于最大后验估计

算法

统计学习是基于训练数据集，根据学习策略，从假设空间中选择最优模型，最后需要考虑用什么样的计算方法求解最优模型。 统计学习问题转化为最优化问题，统计学习的算法就转化为求解最优化问题的算法。

算法：即学习模型的具体计算方法。

- 有显式的解析解的最优化问题；
- 无显式的解析解的最优化问题，需要用数值计算的方法求解。
  - 如何保证找到全局最优解；
  - 如何保证求解的过程高效。

1.4 模型的评估与选择

- 1.4~1.7，与模型选择有关的问题。
- 1.8~1.10，与模型应用有关的问题。

模型评估（训练误差与测试误差）

学习方法评估的标准

- 基于损失函数的模型的训练误差 (training error)：用来评估一个学习问题是否容易学习
- 基于损失函数的模型的测试误差 (test error)：用来评估一个模型是否具备有效的预测
  - 泛化能力 (generalization ability)：学习方法对未知数据的预测能力

模型选择

过拟合 (over-fitting): 学习时选择的模型所包含的参数过多，以至于模型对已知数据预测较好，未知数据预测较差的问题

模型选择的常用方法："正则化"和"交叉验证"

1.5 正则化与交叉验证

正则化(regularization)

正则化: 结构风险最小化策略的实现，是在经验风险上加一个正则化项或惩罚项。

- 正则化项一般是模型复杂度的单调递增函数
  - 复杂度定义参考 Kolmogorov 复杂性理论 (complexity theory) [Haykin, 2011] P48
- Occam 剃刀原理：应用于模型选择时符合正则化的想法，即所有能够解释数据的模型中，复杂度越小越好。
- Bayes 估计：正则化项对应于模型的先验概率。数据较少时先验概率就可以抑制数据中噪声的干扰，防止出现过拟合问题。数据很多时，先验概率就让位于数据对模型的解释
- 正则化是优化学习算法，调整目标函数，增加先验知识的重要手段，是机器学习的核心之一。
  - 简单了解：[周志华，2018] P133
  - 深入理解：[Haykin, 2011] C07

交叉验证 (cross validation)

在数据充足时，随机地将数据切分成三个部分：训练集、验证集和测试集，选择对验证集有最小预测误差的模型。

- 训练集 (training set): 用来训练模型；
- 验证集 (validation set): 用来选择模型；
- 测试集 (test set): 用来评估模型。

交叉验证的常用方法

- 简单交叉验证：随机地将数据分成两个部分，70% 的数据为训练集，30% 的数据为测试集，选择测试误差最小的模型；
- S 折交叉验证
  - 随机地将数据分成 S 个互不相交的大小相同的部分
  - 然后利用 S-1 个部分的数据训练，1 个子集测试模型，
  - 再将这一个过程对所有可能的选择重复进行，
  - 最后选择 S 次评测中平均测试误差最小的模型。
- 留一交叉验证 :当 S=N 时采用的 S 折交叉验证 ,适用于数据极度缺乏的情况下。(N 为给定数据集的容量）

1.6 泛化能力

泛化误差(generalization error)

泛化误差: 是指学到的模型对未知数据预测产生的误差，反映了学习方法的泛化能力。

泛化能力 (generalization ability): 是指学习方法学习到的模型对未知数据的预测能力

泛化误差的上界(generalization error bound)

泛化误差的上界: 泛化误差的概率上界，通过比较两种学习方法的泛化误差概率上界来确定

泛化误差上界的性质：

- 是样本容量的函数，当样本容量增加时，泛化上界趋向于 0;
- 是假设空间的函数，当假设空间容量增加时，泛化误差上界就会变大，表示模型更难学

泛化误差上界定理及证明（建议跳过）

1.7 生成模型与判别模型

生成模型 (generative model): 模型表示了给定输入 X 产生输出 Y 的生成关系。

- 还原出联合概率分布；
- 学习收速度快；
- 样本容量增加时，能够更好地逼近真实模型；
- 存在隐变量时，仍然可以使用。

生成模型的主要案例：朴素 Bayes 方法和隐马尔可夫模型 (Hidden Markov Model, HMM);

注：生成模型是比较难理解的概念，HMM 是理解生成模型比较好的途径，如果对 HMM 感兴趣可以参考

- 简单了解：[周志华，2018] P320
- 深入理解：[Rabiner, 1989]

判别模型 (discriminative model): 由数据直接学习决策函数或条件概率分布作为预测的模型

- 直接学习得到条件概率分布或者决策函数；
- 直接面对预测，学习的准确率更高；
- 基于参数是直接学习得到的，因此可以对数据进行各种程度上的抽象、定义和使用特征，简化学习问题。

判别模型的案例：k 近邻法、感知机、决策树、Logistic 回归模型、最大熵模型、支持向量机、提升方法和条件随机场等

1.8 分类(classification)问题

分类: 利用分类器对新输入的数据进行输出的预测。

- 分类器 (classifier): 基于有监督学习从数据中学习得到的分类模型或者分类决策函数。

解决分类问题的两个过程

- 学习过程：根据已知的训练数据集利用有效的“学习方法”得到一个分类器；
- 分类过程：利用学习得到的分类器对新输入的实例进行分类。

评价分类器性能的指标：

- 分类准确率 (accuracy), 即对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。

二类分类问题常用的评价指标：

- 精确率 (precision)
- 召回率 (recall)
- F1 值：精确率和召回率的调和均值

解决分类问题的常用方法：k 近邻法、感知机、朴素 Bayes 法，决策树、决策列表、Logistc 回归模型、支持向量机、提升方法等

1.9 标注问题

标注问题：是分类问题的推广，也是更复杂的结构预测问题的简单形式。

- 数据：输入是一个观测序列； 输出是一个标记序列或状态序列。
- 目标：通过学习得到能够对观测序列给出标记序列作为预测的模型。
- 解决标注问题的两个过程：学习过程 和 标注过程
- 评价标注问题的指标：准确率、精确率和召回率。

- 解决标注问题的常用方法：HMM 模型和条件随机场（CRF）。

1.10 回归(regression)问题

回归模型：表示从输入变量到输出变量之间的映射关系的函数，等价于“函数拟合”。

- 回归: 用于预测输入变量（自变量）和输出变量（因变量）之间的关系。

解决回归问题的两个过程：学习过程 和 预测过程。

回归问题的分类

- 按输入变量的个数：一元回归和多元回归；
- 按输入变量和输出变量之间的关系：线性回归和非线性回归。

回归学习最常用的损失函数：平方损失函数，求解平方损失函数可以用最小二乘法。

1.11 本章概要

- 统计学习是计算机基于数据构建概率统计模型 ,并且运用这个模型对数据进行分析和预测的一门学科。
- 统计学习方法的三个要素：模型、策略、算法
- 有监督学习：从给定的、有限的训练数据出发，假设数据是独立同分布的，而且假设模型属于某个假设空间，应用某个评价标准，从假设空间中选取一个最优的模型，使得这个模型对于已知的训练数据和未知的测试数据在给定评价标准意义下有最准确的预测。
- 在统计学习中，选择最优的模型和提高学习的泛化能力是个重要问题
  - 如果只考虑减少训练误差，就会出现过拟合现象
  - 模型选择的方法和正则化和交叉验证
- 有监督学习的三个重要问题：分类问题、标注问题、回归问题

- 统计学习的主要方法：感知机、K 近邻、朴素贝叶斯、决策树、逻辑 Logistic 回归、最大熵模型、支持向量机、提升方法、EM 算法、HMM 模型、条件随机场

第2章 感知机

2.1 感知机模型

感知机，是根据输入实例的特征向量对其进行二类分类的线性分类模型，属于判别模型；

模型参数包括：

- 权值(weight)或权值向量(weight vector)
- 偏置(bias)

模型对应于输入空间（特征空间）中的分离超平面

2.2 感知机学习策略

假设：感知机学习的训练数据集是线性可分的；

目标：求得一个能够将训练集正实例点和负实例点完全正确分开的分离超平面；

策略：即定义（经验）损失函数，并将损失函数极小化；

- 损失函数定义为：误分类点的总数，损失函数不是连续可导函数，不易优化
- 损失函数定义为：误分类到分离超平面的总距离

2.3 感知机学习算法

感知机学习算法是基于误差——修正的学习思想，是由误分类驱动的；

学习算法的优化方法

- 批量学习：可以基于梯度进行优化
  - 一阶：最速下降法或梯度下降法；

- 二阶：牛顿法、共轭梯度法等等

- 在线学习：基于随机梯度下降法的对损失函数进行最优化 [Goodfellow, 2017] P95, P180
  - 原始形式：算法简单且易于实现。先任意选取一个超平面，然后随机选择一个误分类点使其用梯度下降法极小化目标函数
    - 例 2.1（比较简单，可以了解）
    - 定理 2.1（过于简略，建议跳过）
  - 对偶形式：（原始形式的另一种数学表示方式，通过对偶形式可以采用其他数学工具求解在线学习，求解的结果是一样的）

当训练数据集线性可分时，感知机学习算法是收敛的，且有无穷多个解。

2.4 本章概要

- 感知机是根据输入实例的特征向量对其进行二类分类的线性分类模型，感知机模型对应于输入空间（特征空间）中的分离超平面
- 感知机学习的策略是极小化损失函数
  - 损失函数对应于误分类点到分离超平面的总距离
- 感知机学习算法是基于随机梯度下降法的对损失函数的最优化算法，算法简单易于实现
  - 原始形式：任意选取一个超平面，然后使用梯度下降法不断极小化目标函数，优化过程中一次随机选取一个误分类点使其梯度下降
  - 对偶形式：
- 当训练数据集线性可分时
  - 感知机学习算法是收敛的，感知机算法在训练数据集上的误分类次数 $k \leq (R/\gamma)^2$
  - 感知机学习算法存在无穷多个解，其解由于初值不同或者迭代顺序不同而可能不同

2.5 学习总结

- 感知机是神经网络的基础，本章只有单个神经元模型，深入学习参考 [Haykin, 2011]
- 神经网络是深度学习的基础，深度学习参考 [Goodfellow, 2017]
- 距离度量是几何的概念，理论可参考 [Duda, 2003] P154
- 学习算法的优化是最优化理论，基本优化方法可参考 [Hyvarinen, 2007] P42

第3章 K 近邻法

K 近邻 (K-Nearest Neighbor, K-NN) 既可以用于分类，也可以用于回归。

- 输入为实例的特征向量，对应于特征空间的点；
- 输出为实例的类别，可以取多个类。

3.1 K 近邻算法

K 近邻法的基本思想：假设给定一个训练数据集，其中的实例类别已经确定； 对新输入的实例分类时 ,根据其 k 个最近邻的训练实例的类别 ,通过多数表决等方式进行预测。因此，K 近邻法不具有显式的学习过程。

K 近邻法的本质就是利用训练数据集对特征向量空间进行切分，并作为其分类的“模型”。

3.2 K 近邻法的模型

K 近邻法的模型：就是基于训练数据集对特征空间的一个划分。当训练集、距离度量、K 值及分类决策规则确定后，输入实例所属类别也唯一确定。

K 近邻法的三个要素：

- 距离度量：常用欧氏距离；(距离定义) [Duda, 2003]
  - L1 范数：Manhattan 距离
  - L2 范数：欧氏距离
  - 无穷范数：最大距离
- K 值的选择：反映了近似误差与估计误差之间的权衡。
  - K 值越大时，近似误差会增大，估计误差会减小，模型也越简单；

- 考虑数据的总体平均特性

- K 值越小时，近似误差会减少，估计误差会增大，模型也越复杂。
  - 考虑数据的每个数据点的特性
- 可以用交叉验证的方式选择最优 k 值。

- 分类决策规则：多数表决规则 (marjority voting rule), 等价于 经验风险最小化。

3.3 K 近邻法的实现基于 kd 树。

（了解即可，实际应用中大多使用的是已经成熟的软件包）

K 近邻法最简单的搜索方法是线性扫描，为了提高搜索的效率，可以使用 kd 树。

- kd 树是一种便于对 k 维空间中的数据进行快速检索的数据结构；
- kd 树是二叉树，表示对 k 维空间的一个划分；
- kd 树的每个节点对应于 k 维空间划分中的一个超矩形区域；
- 利用 kd 树可以省去对大部分数据点的搜索，从而减少搜索的计算量。

构造 kd 树的方法：

- 构造根结点，使根节点对应于 K 维空间中包含所有实例点的超矩形区域
- 对 K 维空间进行反复切分，生成子节点
- 在超矩形区域(节点)上选择一个坐标轴和在此坐标轴上的一个切分 ,确定一个超平面，这个超平面通过选定的切分点并垂直于选定的坐标轴 ,将当前的超矩形区域切分为两个子区域
  - 切分点一般选定坐标轴上的中位数，保证 kd 树是平衡的
  - 平衡的 kd 树在搜索时效率未必是最优的

搜索 kd 树的方法：

- 给定一个目标点，搜索其最近邻，找到目标点的叶节点；
- 从这个叶节点出发，依次退回到父节点；
- 不断查找与目标点最近邻的节点，当确定不可能存在更近的节点时终止。

3.4 学习总结

了解即可，因为面对高维问题效果很差，需要考虑降维操作。[周志华，2018] P225

第4章 朴素 Bayes 法

朴素 (naive) Bayes 法：是基于 Bayes 定理与所有特征都遵循条件独立性假设的分类方法。

朴素 Bayes 法是 Bayes 分类法的一种，遵循 Bayes 定理建模。[Mitchell, 2003] P112

朴素 Bayes 法基于的条件独立性假设：是说用于分类的特征在类别确定的条件下都是条件独立的。简化了计算复杂度，牺牲了分类准确率。

4.1 朴素 Bayes 法的学习与分类

朴素 Bayes 法是生成学习方法。

- 先验概率分布；
- 条件概率分布；
- 后验概率分布。后验概率最大化准则等价于期望风险最小化准则。

目标：由训练数据学习联合概率分布；

4.2 朴素 Bayes 方法的概率参数估计方法

- 极大似然估计 ：概率估计常用的方法；
  - 可能会出现所要估计的概率值为 0 的问题
- Bayes 估计 ：重点在于了解与极大似然估计的差别，才可以正确使用。
  - 解决了概率值为 0 的问题，因此也叫 Laplace 平滑。

4.3 学习总结

虽然不需要手工估计参数，但是对估计的理解很重要，书中的描述过于简单，具体内容请参

考 [Duda, 2003] P67

对于概念上的理解还可以参考 [周志华，2018] C07

第5章 决策树（decision tree)

决策树是一种既可以用于分类，也可以用于回归的方法

（本章主要讨论的是分类决策树）

5.1 决策树模型

决策树模型

分类决策树模型：是基于特征对实例进行分类的树形结构。

- 模型的组成结构：树形结构
  - 结点（node)
    - 内部结点（internal node)：路径上的结点，表示一个特征或者属性
      - 根结点（root node）
    - 叶结点（leaf node)：表示一个类别
  - 有向边（directed edge)
- 分类决策树可以转换成一个 if-then 规则的集合；
  - 决策树的根结点到叶结点的每一条路径构建一条规则；
    - 内部结点的特征对应着规则的条件，
    - 叶结点的类对应着规则的结论。
- 分类决策树的路径或者其对应的规则集合的性质：互斥并且完备，即全覆盖。
  - 覆盖是指实例的特征与路径上的特征一致或实例满足规则的条件。
  - 每个实例都被一条路径或者一条规则所覆盖

- 每个实例只被一条路径或者一条规则所覆盖

- 分类决策树可以看作是定义在特征空间与类空间上的条件概率分布。
  - 这个条件概率分布定义在特征空间的一个划分上，
    - 将特征空间划分为互不相交的单元或区域，
    - 每个单元定义一个类的概率分布就构成了一个条件概率分布。
    - 决策树分类时，将结点的实例分到条件概率大的类中。
- 决策树模型的主要优点：可读性强，分类速度快。
  - 学习时，利用训练数据，根据损失函数最小化的原则建立决策树模型
  - 预测时，对新的数据，利用决策树模型进行分类。

决策树学习

学习目标：根据给定的训练数据集，构建一个与训练数据拟合很好，并且复杂度最小的决策树，使之能够对实例进行正确的分类。

学习本质：

- 从训练数据集中归纳出一组分类规则。
  - 决策树与训练数据集不相矛盾的决策可能有多个，也可能一个也没有；
  - 因此，寻找的决策树要与训练数据的矛盾较小，同时还具有较好的泛化能力。
- 也可以看作由训练数据集估计条件概率模型
  - 基于特征空间划分的类的条件概率模型会有很多；
  - 选择的条件概率模型对训练数据拟合的效果很好；
  - 选择的条件概率模型对未知数据预测的效果很好。

决策树的学习算法包括 3 个部分

- 特征选择：递归地选择最优特征
- 决策树的生成：根据该特征对训练数据进行分割
- 决策树的剪枝：使之对各个数据集有一个最好的分类的过程
  - 学习准则：损失函数最小化
    - 损失函数是一种正则化的极大似然函数
  - 从所有可能的决策树中选取最优决策树是 NP 完全问题；
    - 现实中采用启发式方法学习次优的决策树。

5.2 特征选择

特征选择的目的：在于选取对训练数据能够分类的特征，提高决策树学习的效率；

特征选择的关键是其准则：

- 样本集合 D 对特征 A 的信息增益（Information Gain）最大
  - 信息增益：集合 D 的经验熵与特征 A 在给定条件下 D 的经验条件熵之差。
    - 熵：表示随机变量不确定性的度量。也称为经验熵。
    - 条件熵：定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望。也称为经验条件熵。
  - 信息增益表示得知特征 X 的信息而使得类 Y 的信息的不确定性减少的程度。
  - 信息增益等价于训练数据集中类与特征的互信息。
  - 信息增益依赖于特征，信息增益大的特征具有更强的分类能力。
- 样本集合 D 对特征 A 的信息增益比（Information Gain Ratio）最大
  - 为了避免信息增益对取值较多的特征的偏重，使用信息增益比来代替；
  - 信息增益比：特征 A 对训练数据集 D 的信息增益与训练数据集 D 关于特征 A

的值的熵之比。

- 样本集合 D 的基尼指数（Gini）最小

5.3 决策树的生成

决策树的生成过程：

- 确定特征选择的准则，根据准则计算指标
- 根据指标选取最优切分点
- 从根结点开发，递归地产生决策树。
  - 通过不断地选择局部最优的特征，得到可能是全局次优的结果。

常用的生成算法

- ID3: 在决策树的各个结点上应用信息增益准则选择特征，递归地构建决策树。
  - 特征选择准则：信息增益
  - 相当于用极大似然法进行概率模型的选择。
  - 具体方法：
    - 从根结点开始，对结点计算所有可能的特征的信息增益，选择信息增益最大的特征作为结点的特征，由这个特征的不同取值建立子结点
    - 再对子结点递归地调用以上方法，构建决策树
    - 直到所有特征的信息增益均很小或者没有特征可供选择
- C4.5: 在决策树的各个结点上应用信息增益比准则选择特征，递归地构建决策树。
  - 特征选择准则：信息增益比
  - 是对 ID3 算法的改进
- CART: 既可用于分类，也可用于回归。

- 等价于递归地二分每个特征
  - 将输入空间即特征空间划分为有限个单元
  - 在这些单元上确定预测的概率分布，在输入给定的条件下输出的条件概率分布
- CART 算法的两个过程
  - 决策树生成：基于训练数据集生成决策树，要尽量大；
    - 回归树生成
      - 用平方误差最小准则求解每个单元上的最优输出值。
      - 回归树通常称为最小二乘回归树。
    - 分类树生成
      - 用基尼指数选择最优特征，并决定该特征的最优二值切分点。
      - 算法停止计算的条件
        - 结点中的样本个数小于预定阈值；
        - 样本集的基尼小于预定阈值；
    - 决策树剪枝
    - 用验证数据集对已经生成的树进行剪枝，剪枝的标准为损失函数最小，基于标准选择最优子树。
    - 可以通过交叉验证法对用于验证的独立数据集上的子树序列进行测试，从中选择最优子树。
    - [Duda, 2003] P320, CART 作为通用的框架，定义了 6 个问题

5.4 决策树的剪枝

在决策树学习中将已经生成的树进行简化的过程称为剪枝（Pruning）。

- 目的：防止决策树存在过拟合问题
- 准则：极小化决策树整体的损失函数或代价函数
  - 等价于正则化的极大似然估计。
- 分类
  - 预剪枝：也叫分支停止准则。在决策树生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点；
  - 后剪枝：先从训练集生成一棵完整的决策树，然后自底向上地对内部结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点。

5.5 学习总结

- 算法（5.1, 5.2, 5.6）+ 例题（5.1, 5.2, 5.3, 5.4）通过算法和例题可以增强理解；
- 损失函数的定义可以进一步参考“不纯度”指标 [Duda, 2003] P320, 或“纯度”指标 [周志华，2018] P75
  - “不纯度”指标是求极小值，可以跟梯度下降法等最优化理论结合。

第6章 Logistic 回归与最大熵模型

对数线性模型：

- Logistic 回归模型：应用 Logistic 函数的分类方法。
- 最大熵模型（Maximum Entropy）：应用最大熵准则的分类方法
  - 最大熵是概率模型学习的一个准则

6.1 Logistic 回归模型

Logistic 回归模型，也称为对数几率回归模型

- 模型结构：
  - 输入是的线性函数
  - 输出的是对数几率模型
    - 基于 Logistic 分布建立的，表示条件概率的分类模型
      - Logistic 分布是 Sigmoid 函数，定义 6.1
    - 对数几率 (log odds) 或 logit 函数
      - 一个事件的几率是指该事件发生的概率与该事件不发生的概率的比值。
- 模型分类：
  - 二项 Logistic 回归模型是二类分类模型，定义 6.2
  - 多项 Logistic 回归模型是多类分类模型
- 模型参数估计
  - 极大似然估计法

6.2 最大熵模型

最大熵模型：

- 模型定义：
  - 是基于最大熵原理推导的，
  - 表示条件概率分布的分类模型，
  - 可以用于二类或多类分类。
- 模型准则：最大熵原理是概率模型学习或估计的一个准则。
  - 最大熵原理认为，在所有可能的概率模型（分布）的集合中，熵最大的模型是最好的模型。
- 最大熵模型的学习方法
  - 最大熵模型的学习过程就是求解最大熵模型的过程
  - 最大熵模型的学习可以形式化为有约束的最优化问题（对偶问题）
    - 拉格朗日乘子参考附录 C
  - 例 6.1, 6.2 方便理解最大熵模型的算法原理。
- 模型算法
  - 学习方法
    - 极大似然估计
    - 正则化极大似然估计
    - 形式化为无约束最优化问题
  - 求解无约束最优化问题的算法
    - 迭代尺度法
    - 梯度下降法

- 牛顿法或者拟牛顿法，计算量大，收敛速度快

6.3 学习总结

- Logistic 模型与最大熵模型都属于对数线性模型。[周志华，2018] C03
- 极大似然估计：写的比较简单，没有原理性的说明
  - 参考（[周志华，2018] P149, [Duda, 2003] P67）
- 模型学习的最优化算法：写的不好理解。
  - 参考（[周志华，2018] P403, [Hagan, 2006] C09）

第7章 支持向量机

支持向量机（Support Vector Machine， SVM）是一种二分类模型。 基本模型是定义在特征空间上的间隔最大的线性分类器

SVM 的基本概念：

- 支持向量决定了最优分享超平面
- 最终判别时，只需要很少的“重要”训练样本，大幅减少计算量。
- 间隔（看懂数学公式就可以理解间隔，判别在数据的维度上又增加了一个维度）
- 与其他模型的比较
- 与感知机的区别：间隔最大化产生最优超平面；
- 与线性模型的区别：使用核技巧成为非线性分类器。

SVM 的分类：

- 线性可分支持向量机：又称为硬间隔支持向量机。当训练数据线性时，通过硬间隔最大化，学习一个线性的分类器。
- 线性支持向量机：又称为软间隔支持向量机，是最基本的支持向量机。当训练数据近似可分时，通过软间隔最大化，学习一个线性的分类器。
- 非线性支持向量机：当训练数据线性不可分时，通过使用核函数（Kernel Function）及软间隔最大化，学习一个非线性的分类器。
    - 输入空间是欧氏空间或者离散集合；
    - 特征空间是 Hilbert 空间；
    - 核函数表示将输入从输入空间非线性映射到特征空间 ,从而使特征空间内的特征是线性可分的，从而可以使用支持向量机来学习线性分类器。

- 通过使用核函数学习非线性支持向量机，等价于隐式地在高维的特征空间中学习线性支持向量机。
- 核方法是比支持向量机更为一般的机器学习方法。

SVM 的学习：

- 学习在特征空间进行的
- 学习策略是间隔最大化

7.1 线性可分支持向量机

- 学习条件：训练数据线性可分
- 学习策略：硬间隔最大化。对训练数据集找到几何间隔最大的超平面，即以充分大的确信度对训练数据进行分类。
  - 求解能够正确划分训练数据集并且几何间隔最大的分离超平面
  - 对训练数据集找到几何间隔最大的超平面意味着以充分大的确信度对训练数据进行分类
  - 这样的超平面对未知原新实例有很好的分类预测能力
- 解的特征：
  - 最优解存在且唯一；（唯一性证明，建议跳过）
  - 支持向量由位于间隔边界上的实例点组成

7.2 线性支持向量机

- 学习条件：
  - 训练数据近似线性可分

- 训练数据中存在一些特异点 (outlier)

- 学习策略：软间隔最大化，即允许支持向量机在一些样本上出错
  - 误判的代价：惩罚参数 C 乘以 替代损失函数 f
  - 替代损失（Surrogate Loss）函数，一般具有较好的数学性质，是凸的连续函数并且是“0/1 损失函数”的上界。
    - hinge 损失（合页损失函数）：$\max(0,1-z)$
      - 保持了稀疏性
    - 指数损失：$\exp(-z)$
    - 对率损失：$\log(1+\exp(-z))$
      - 相似于对率回归模型
  - 目标是使间隔尽量大，误分类点尽量少。

- 解的特征
  - 权值唯一，偏置不唯一；
  - 支持向量由位于间隔边界上的实例点、间隔边界与分离超平面之间的实例点、分离超平面误分一侧的实例点组成；
  - 最优分享超平面由支持向量完全决定。

7.3 非线性支持向量机

- 基本概念
  - 线性空间：满足线性性质的空间
  - 距离：是一种度量
    - 距离的集合 → 度量空间 + 线性结构 → 线性度量空间

- 范数：表示某点到空间零点的距离
  - 范数的集合 → 赋范空间 + 线性结构 → 线性赋范空间
- 内积空间：添加了内积运算的线性赋范空间
  - 线性赋范空间 + 内积运算 → 内积空间
- 欧氏空间：有限维的内积空间
- 希尔伯特空间：内积空间满足完备性，即扩展到无限维
  - 内积空间 + 完备性 → 希尔伯特空间
- 巴拿赫空间：赋范空间满足完备性
  - 赋范空间 + 完备性 → 巴拿赫空间

- 条件：
  - 训练数据非线性可分；
  - 通过非线性变换（核函数）将输入空间（欧氏空间或离散集合）转化为某个高维特征空间（希尔伯特空间）中的线性可分；
  - 在高维特征空间中学习线性支持向量机。
- 学习策略：核技巧 + 软间隔最大化

7.4 最大间隔法

- 间隔概念
  - 函数间隔：表示分类的正确性及确信度
  - 几何间隔：规范化后的函数间隔，实例点到超平面的带符号的距离
- 分类
  - 硬间隔最大化 (hard margin maximization)

- 软间隔最大化（soft margin maximization）

- 间隔最大化的形式化
  - 求解凸二次规划问题
    - 最优化算法
  - 正则化的合页损失函数的最小化问题

- 求解过程
  - 原始最优化问题应用拉格朗日对偶性；
  - 通过求解对偶问题得到原始问题的最优解。
  - 中间也可以根据需要自然引入核函数。

7.5 核技巧（kernel method）通用的机器学习方法

- 应用条件
  - 非线性可分训练数据可以变换到线性可分特征空间；
  - “目标函数”中的内积可以使用“非线性函数”的内积替换；“非线性函数”的内积可以使用“核函数”替换；
  - 核函数使非线性问题可解。

- 常用的核函数，即正定核函数（Positive Definite Kernel Function）
  - 线性核：对应于线性可分问题
  - 多项式核函数：$K(x,z)=(x\cdot z+1)^p$
  - 高斯核函数：$K(x,z)=\exp(\frac{||x-z||^2}{2\sigma^2})$
  - 字符串核函数：定义在离散数据集合上的核函数。
    - 字符串核函数 :给出了字符串 S 和 T 中长度等于 n 的所有子串组成的特征向量

的余弦相似度。

- Sigmoid 核函数：
- 函数组合得到的核函数
  - 两个核函数的线性组合仍然是核函数，$k1(x,z)$ 和 $k2(x,z)$是核函数， $c1$ 和 $c2$是任意正数，则 $k(x,z)=c1k1(x,z)+c2k2(x,z)$也是核函数。
  - 两个核函数的直积仍然是核函数，$k1(x,z)$ 和 $k2(x,z)$是核函数，则 $k(x,z)=k1(x,z)k2(x,z)$也是核函数。
  - $k1(x,z)$是核函数 ,$g(z)$ 是任意函数 ,则 $k(x,z)=g(z)k1(x,z)g(z)$也是核函数。

7.6 SMO 算法

- 序列最小最优化（Sequential Minimal Optimmization，SMO）算法，是一种启发式算法
  - 基本思路 :如果所有变量的解都满足最优化问题的 KKT 条件 ,那么这组解就是最优化问题的解
- 算法组成：
  - 求解两个变量的二次规划的解析方法
  - 选择珠启发式方法
- 算法流程
  - 将原二次规划问题分解为只有两个变量的二次规划子问题；
    - 选择的第一个变量的过程为外层循环，外层循环在训练样本中选取违反 KKT 条件最严重的样本点，并将其对应的变量作为第一个变量；
    - 选择的第二个变量的过程为内层循环 ,内层特征是使目标函数增长最快的变量；
    - 目标是使两个变量所对应样本之间的间隔最大。

- 对子问题进行解析分解；
- 直到所有变量满足 KKT 条件为止。

- 算法特点：通过启发式的方法得到原二次规划问题的最优解。
  - 因为子问题总有解析解，所以每次求解子问题的速度很快，虽然子问题数量很大，但是总体上依然是高效的。

7.7 学习总结

- 支持向量机与神经网络是两大重要的机器学习算法；
- 结合周老师的书一起看，对于理解支持向量机会有较大帮助。[周志华，2018] C06
- 深入了解支持向量机的理论分析。[Haykin, 2011] C06

第8章 提升方法（集成学习）

提升方法是一种统计学习方法，也是一种提升模型学习能力和泛化能力的方法，还是一种组合学习（集成学习）的方法，是统计学习中最有效的方法之一。

1. 为什么要将各种学习方法组合起来？

- 强可学习方法与弱可学习方法的等价性；
  - 在概率近似正确（Probably Approximately Correct，PAC）学习的框架下
    - 一个概念（一个类），如果存在一个多项式的学习算法能够学习它，并且正确率很高，那么就称这个概念是强可学习的
    - 一个概念，如果存在一个多项式的学习算法能够学习它，并且学习的正确率只比随机猜测略好，那么就称这个概念是弱可学习的
- 将各种弱可学习方法组合起来就可以提升（boost）为强可学习方法

2. 如何将各种学习方法组合起来？

- Boosting 算法：个体学习器之间存在强依赖关系，一系列个体学习器串行生成。
  - 是一种通用的组合算法，可以将各种分类算法进行组合。
  - 算法追求的两个目标
    - 每一轮如何改变训练数据的权值或者概率分布
    - 如何将弱分类器组成成一个强分类器
  - 算法的主要代表：
    - AdaBoost 算法
    - 提升决策树模型
      - 以分类树或回归树为基本分类器的提升方法（组合算法）

- 提升树是统计学习中性能最好的方法之一

- Bagging 算法 :不存在强依赖关系 ,可以并行生成(书中没有 ,参考[周志华 ,2018] C8.3)
  - 算法的主要代表：随机森林（ Random Forest ）

8.1 AdaBoost（Adaptive Boost）自适应提升算法

- 模型：加法模型
  - 如何改变训练数据的权值和概率分布
    - 采用“分而治之”的方法。提高那些被前一轮弱分类器错误分类的样本的权值，从而保证后一轮的弱分类器在学习过程中能够更多关注它们。
  - 如何将弱分类器组合成一个强分类器
    - 采用“加权多数表决”的方法。加大分类误差率小的弱分类器的权值，从而保证它们在表决中起较大的作用。
- 策略：指数损失函数极小化，即经验风险极小化。
- 算法：前向分步算法来优化分步优化指数损失函数的极小化问题。
- 算法的训练误差分析
  - AdaBoost 能够在学习过程中不断减少训练误差，即减少训练数据集的分类误差率
- 算法的优化过程分析
  - 因为学习的是加法模型，所以能够从前向后，每一步只学习一个基函数及基系数，逐步逼近优化目标函数，简化优化的复杂度。
  - 前向分步算法与 AdaBoost 的关系：AdaBoost 算法是前向分布算法的特例。
    - AdaBoost 算法的模型是基本的分类器组成的加法模型；损失函数是指数函数

8.2 提升树模型

- 模型：采用加法模型（基函数的线性组合），以决策树为基函数的提升方法
  - 分类问题：决策树是二叉树
  - 回归问题：决策树是二叉回归树
  - 决策树的加法模型：$f_M(x) = \sum_{m=1}^{M} T(x;\Theta_m)$，其中$T(x;\Theta_m)$表示决策树，$\Theta_m$为决策树的参数，$M$为树的个数
- 策略：损失函数
  - 分类问题：指数损失函数
  - 回归问题：平方误差函数
  - 一般决策问题：一般损失函数
- 算法：前向分步算法
  - 梯度提升算法（Gradient Boosted Decision Tree ,GBDT）：利用下降法的近似方法，解决离散数据的优化问题，原理参考、[Friedman, 2001]

学习总结

- 学习基础
  - 熟悉重要的分类算法：神经网络和支持向量机
  - 熟悉常用的分类算法：k 近邻法和决策树
- 学习目标：
  - 组合各种分类算法，从而产生质量更好的学习能力和泛化能力模型
- 胡思乱想：
  - 全连接的深度神经网络就是理论上最完美的组合模型，问题在于维度灾难带来的计算复杂度问题。

- 为了解决计算复杂度问题，就需要了解其他分类模型，因为其他分类模型就是具备了先验知识的神经网络模型，将那些分类模型转化为神经网络模型后就可以大幅减少连接的数量。

- 参考资料
  - 概率近似正确 (probably approximately correct, PAC) 来自计算学习理论，可参考[周志华，2018] C12, [Mitchell, 2003] C07
  - 集成学习 (ensemble learning) 也被称为多分类器系统、基于委员会的学习等，可参考[周志华，2018] C08

第9章 EM 算法及推广

学习基础

- 概率模型求解
  - 如果概率模型的变量都是观测变量，那么给定数据，可以使用极大估计方法或者贝叶斯估计方法求解模型参数
  - 如果概率模型的变量中含有隐含变量，可以使用 EM 算法求得某个解
    - 因为解空间可能存在若干个极值点，而 EM 算法受初始值的影响只能得到其中的一个极值点作为解，不一定是全局最优解。

EM 算法：是对含有隐变量的概率模型进行极大似然估计或者极大后验估计的迭代算法。

- E 步，求期望；利用数据和假设的初值，求得一个隐变量的条件概率分布的期望，即“Q 函数”。（因为无法求得条件概率分布的具体值）

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z\left[\log(P(Y, Z|\theta)|Y, \theta^{(i)})\right] \\ &= \sum_{\mathrm{Z}} \log P(Y, Z|\theta) P\left(Z\middle|Y, \theta^{(i)}\right) \end{aligned}$$

- M 步，求极值。利用“Q 函数”来求极值，这个极值可以帮助拟合的概率分布更加逼近真实分布。

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)})$$

- Q 函数的定义（理解 Q 函数的涵义可以更好地推广到应用中）
- EM 算法的推导（如果书上的无法理解，还可以参考本文中的其他文献）
  - EM 算法是收敛的，但是有可能收敛到局部最小值。
  - EM 算法可以看成利用凸函数进行概率密度逼近；
  - 如果原概率密度函数有多个极值，初值的不同就可能逼近到不同的极值点，所以无

法保证全局最优。

- EM 算法的应用（下面的两个应用都是重点）
  - 高斯混合模型
  - HMM（隐 Markov 模型） 参考 C10
- EM 算法的推广（建议跳过，对了解 EM 算法帮助不大，只有研究 EM 算法才需要）
  - F 函数的极大 - 极大算法
  - 广义 EM 算法（GEM）

学习总结

- EM 算法的详细推导。[Borman, 2004]，或者[Determined22, 2017]的 EM 算法简述及简单示例（三个硬币的模型）
- EM 算法的概率分析。[Friedman, 2001]，或者[苏剑林，2017]的梯度下降和 EM 算法
- EM 算法的深入理解。可以参考[史春奇，2017]的 Hinton 和 Jordan 理解的 EM 算法

第10章 隐 Markov 模型（HMM）

学习基础

- 随机过程：用于理解 Markov 链的数学含义
- EM 算法：用于计算 HMM 的学习问题

10.1 HMM 的基本概念

Markov 链的定义

- 随机过程
  - 研究对象是随时间演变的随机现象。[盛骤，2015] C12
  - 设 T 是一无限实数集，对依赖于参数 t（t 属于 T）的一族（无限多个）随机变量称为随机过程。
  - 我的理解
    - 随机过程在任一个时刻 t，被观测到的状态是随机的，但是这个随机状态是由一个确定的函数控制的。
    - 例如：有 3 块金属放在箱子里面，任一个时刻 t 取出的金属是随机的，但是每块金属衰退的速度是由这块金属自身的函数控制的。
    - 随机变量刻画的是数值的随机性（某个数出现的概率）
    - 随机过程刻画的是函数的随机性（某个函数出现的概率）
- Markov 过程
  - Markov 性或无后效性：
    - 过程（或系统）在时刻 t_0 所处的状态为已知的条件下，过程在时刻 t>t_0 所

处状态的条件分布与过程在时刻 t_0 之前所处的状态无关。

- 即在已经知道过程“现在”的条件下，其“将来”不依赖于“过去”。[盛骤，2015] C13

- Markov 过程：具有 Markov 性的随机过程，称为 Markov 过程。

- Markov 链
  - 时间和状态都是离散的 Markov 过程称为 Markov 链，简称马氏链。
  - 深入理解可参考 [Rabiner, 1989]
- HMM
  - 关于时序的概率模型
    - 用于描述一个被观测到的随机序列，
      - 观测序列 O：每个状态生成一个观测，一个状态序列生成一个观测序列
    - 这个随机序列是由不可观测的状态随机序列生成的，
      - 状态序列 Q：隐藏的 Markov 链随机生成的状态序列；
    - 这个状态随机序列是由隐藏的 Markov 链随机生成的。
    - 序列的每一个位置都可以看作一个时刻。

HMM 的基本假设

- 齐次 Markov 假设
  - 即假设隐藏的 Markov 链在任意时刻 t 的状态只依赖于前一个时刻的状态，而与其他时刻的状态及观测无关，也与时刻 t 无关；
- 观测独立性假设，
  - 即假设任意时刻 t 的观测只依赖于该时刻的 Markov 链的状态，与其他观测和状态无关。

HMM 的基本元素

- N，模型的状态数；

- M，每个状态生成的可观测的标志数；
- $A = [a_{ij}]_{N\times N}$，转移概率矩阵，a_{ij} 表示从状态 i 转移到状态 j 的概率；
- $B = [b_j(k)]_{N\times M}$，观测概率矩阵，$b_j(k)$ 表示状态 j 产生标志 k 的概率；
- $\Pi = (\pi_i)$，初始状态分布，π_i 表示一开始系统在状态 i 的概率。
- HMM 参数的数学表示：$\lambda = (A, B, \pi)$

HMM 的三个基本问题

- 概率计算问题
  - 给定观测序列 O 和模型参数 λ，计算基于这个模型下观测序列出现的概率 $P(O|\lambda)$；
- 预测问题
  - 给定观测序列 O 和模型参数 λ，寻找能够解释这个观测序列的状态序列，这个状态序列的可能性最大；
  - 除非是退化的模型，否则不会有“正确”的状态序列，因为每个状态序列都有可以生成观测序列；
  - 只可能是依据某个优化准则，使找到的状态序列尽可能的逼近真实的状态序列。
- 学习问题
  - 给定观测序列 O，寻找能够解释这个观测序列的模型参数 λ，使得 $P(O|\lambda)$ 最大。
  - 评测哪个模型能最好地解释观测序列。

HMM 的三个基本问题的解决方案

- 概率计算问题：前向算法；
  - 先了解直接计算法，理解 HMM 需要计算的概率的方法和目的，同时明白直接计算法存在的问题；

- 再了解前向算法，如果利用栅格方法叠加前面计算的成果，从而降低直接计算法的庞大计算量。

- 预测问题：
  - 近似算法：计算简单，不能保证预测的状态序列整体是最有可能的状态序列
  - Viterbi 算法（重点）：利用动态规划求解 HMM 的预测问题
    - 即利用动态规划求解概率最大路径问题（最优路径）

- 学习问题：
  - 监督学习算法：极大似然估计，需要训练数据。
  - 非监督学习算法（重点）：Baum_Welch 算法（前向 + 后向算法 + EM 算法）
    - 利用前向 + 后向算法计算转移概率矩阵；
    - 再基于 MLE 理论构造 P(O|λ) 函数；
    - 因为函数中有三个参数不可知，无法直接计算得到，因为采用 EM 算法迭代求解。

HMM 的基本类型

- 基本的 HMM 类型
  - 4 状态遍历 HMM；其他类型都是遍历 HMM 的特例。
  - 4 状态从左到右 HMM；
  - 6 状态从左到右并行路径 HMM。
- 观测序列的密度是连续函数的 HMM：增加了混合高斯作为约束；
- 自回归的 HMM：很适合语音处理；
- 无输出的 HMM：即某些状态转移时无观测输出，主要用于语音识别；
- 一组状态到另一组状态转换：组内状态无转移；

- 优化准则：利用概率理论（ML）或信息理论（MMI，MDI）刻画；
- 比较 HMM 模型：用于模型的测度和选择，常用的测度（交叉熵或散度或判别信息）

HMM 算法的具体实现方法

- 观测数据的尺度化，方便计算机处理，防止溢出；
- HMM 模型的训练：通过多个观测序列进行训练，估计模型的参数；
- HMM 模型参数的初始值设定，没有形式化方法，只能凭借经验；
- 观测数据数量过少，或者观测数据不完整
  - 扩大用于训练的观测集的大小（现实不可操作）；
  - 减少 HMM 模型的参数个数，即减小 HMM 模型的规模；
  - 利用插值的方法补齐或者增加数据。
- HMM 模型的选择
  - 确定 HMM 模型的状态（模型状态数，模型路径数）
  - 确定 HMM 观测的标志（连续还是离散，单个还是混合）
  - 无形式化方法，依赖于具体的应用。

学习总结

- 随机过程和 HMM 算法的基本概念的理解，特别是语音识别和语言处理方向的研究极为重要；
- HMM 算法的计算过程的了解，虽然可以调用成熟的模块，但是了解这个计算过程对于 HMM 计算的调优可能会有帮助；
- HMM 算法的学习极力推荐 [Rabiner, 1989]，本章的框架就是基于这篇文章写的。
- 概率上下文语法（Probabilistic Context-Free Grammar，PCFG）：是 HMM 模型的一种推广
  - HMM 的不可观测序列是状态序列

- PCFG 的不可观测数据是上下文无关语法树

- HMM 是动态贝叶斯网络（Dynamic Bayesian Network，DBN）的一种特例
  - DBN 是定义在时序数据上的贝叶斯网络

第11章 条件随机场（CRF）

11.1 基本概念

（下面的概念已经超出书本内容，包括了整个概率图模型的概念）

概率模型

- 提供了一种描述框架，将学习任务归结于计算变量的概率分布。
- 推断：利用已知变量推测未知变量的分布，核心是如何基于可观测变量推测出未知变量的条件分布。

生成（generative）模型

- 考虑联合分布，是所有变量的全概率模型；
- 由状态序列决定观测序列，因此可以模拟（“生成”）所有变量的值。
- 具有严格的独立性假设；
- 特征是事先给定的，并且特征之间的关系直接体现在公式中。
- 优点
  - 处理单类问题比较灵活；
  - 模型变量之间的关系比较清楚；
  - 模型可以通过增量学习获得；
  - 可以应用于数据不完整的情况。
- 缺点：模型的推导和学习比较复杂。

- 应用：n 元语法模型 、HMM 、Markov 随机场 、Naive Bayes 分类器 、概率上下文无关文法

判别（discriminative）模型

- 考虑条件分布，认为由观测序列决定状态序列，直接对后验概率建模；
- 从状态序列中提取特征，学习模型参数，使得条件概率符合一定形式的最优。
- 特征可以任意给定，一般利用函数进行表示。
- 优点：模型简单，容易建立与学习；
- 缺点：描述能力有限，变量之间的关系不清晰，只能应用于有监督学习。
- 应用：最大熵模型 、条件随机场 、最大熵 Markov 模型 (maximum-entropy Markov model, MEMM) 、感知机

概率图模型

概率图模型：是一类用图来表达变量相关关系的概率模型，

- 在概率模型的基础上 ,使用了基于图的方法来表示概率分布(或者概率密度、密度函数), 是一种通用化的不确定性知识表示和处理的方法。
- 图是表示工具
  - 结点表示一个或者一组随机变量
  - 结点之间的边表示变量间的概率依赖关系，即“变量关系图”。
- 有向图模型（Bayes 网）：使用有向无环图表示变量间的依赖关系，如：推导关系
  - 静态 Bayes 网络
  - 动态 Bayes 网络：适合处理一般图问题

- 隐 Markov 模型：结构最简单的动态 Bayes 网，适合处理线性序列问题，可用于时序数据建模，主要应用领域为语音识别、自然语言处理等。

- 无向图模型（Markov 网）：使用无向图表示变量间的依赖关系，如：循环关系
  - Markov 随机场：典型的无向图模型（Makrov 网）
  - Boltzman 机
  - 通用条件随机场：适合处理一般图问题
  - 线性链式条件随机场：适合处理线性序列问题

Bayes 网络（信念网，信度网，置信网）

- 目的：通过概率推理处理不确定性和不完整性问题
- 构造 Bayes 网络的主要问题
  - 表示：在某一随机变量的集合上给出其联合概率分布。
  - 推断：因为模型完整描述了变量及其关系，可以推断变量的各种问题。
    - 精确推理方法：变量消除法和团树法
    - 近似推理方法：重要性抽样法、MCMC 模拟法、循环信念传播法和泛化信念传播法等
  - 学习：决定变量之间相互关联的量化关系，即储存强度估计。
    - 参数学习常用方法：MLE、MAP、EM 和 Bayes 估计法。
    - 结构学习：

11.2 随机场

Markov 随机场 (Markov Random Field, MRF)

- 定义
  - 是一组有 Markov 性质的随机变量的联合概率分布模型，
  - 联合概率分布满足成对、局部和全局 Markov 性。
  - 由一个无向图 G 和定义 G 上的势函数组成。
- 基本概念
  - 团 (clique)：是图中结点的一个子集，团内任意两个结点都有边相连。也称为完全子图 (complete subgraph)。
  - 极大团 (maximal clique)：若在一个团 C 中加入任何一个结点都不再形成团，就说那个团 C 是最大团。极大团就是不能被其他团所包含的团。
  - 因子分解 (factorization)：将概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数的乘积形式的操作。
  - 分离集 (separating set) :若从结点集 A 中的结点到结点集 B 中的结点都必须经过结点集 C 中的结点，则称结点集 A 和 B 被结点集 C 所分离。
  - 全局 Markov 性：给定两个变量子集的分离集，则这两个变量子集条件独立
    - 局部 Markov 性：给定某变量的邻接变量，则该变量独立于其他变量
    - 成对 Markov 性：给定所有其他变量，两个非邻接变量条件独立 。
- 势函数
  - 用于将模型进行参数化的参数化因子，称为团势能或团势能函数，简称势函数
  - 定义在变量子集上的非负实函数，主要用于定义概率分布函数，亦称“因子”。

- 多个变量之间的联合概率可以基于团分解为多个因子的乘积。
- 指数函数经常被用于定义势函数。

条件随机场

- 条件随机场(Conditional Random Field, CRF)的定义
  - 是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型
  - 用来处理标注和划分序列结构数据的概率化结构模型。
  - 假设输出随机变量之间的联合概率分布构成概率无向图模型，即 Makrov 随机场
- 线性链条件随机场
  - 输入序列（观测序列）和输出序列（标注序列）为线性链表示的随机变量序列
  - 在给定“输入序列”的条件下的“输出序列”的条件概率分布，其数据表示为参数化的对数线性模型
    - 模型包含特征及相应的权值
    - 特征是定义在线性链的边和结点上
  - 是“输入序列”对“输出序列”预测的判别模型
- 构造 CRF 的主要问题
  - 特征的选取
  - 参数训练：
    - CRF 的概率计算问题：前向——后向算法
    - CRF 的学习算法：改进的迭代尺度法、梯度下降法、拟牛顿法
  - 解码或标注：CRF 的预测问题（Viterbi 算法），给定输入序列求条件概率最大的输出序列

- CRF 优点：相比 HMM 没有独立性要求，相比条件 Markov 模型没有标识偏置问题

11.3 学习总结

- 本书的描述概念性内容过少，不利于理解，建议阅读 [周志华，2018] C14
- 以概率图模型为基础来理解条件随机场会更加容易，也能够保证知识相互之间的联系，还可以加深对 HMM 的理解。
- CRF 可以看作 MEMM（最大熵马尔可夫模型）在标注问题上的推广
- CRF 的主要应用是自然语言处理，因此结合自然语言处理来理解概念也会更加深刻。[宗成庆，2018] C06
- 虽然国内几本书都写的不错，但是 CRF 都不是他们书中的重点，若想深入学习 CRF 还是请参考 [Sutton, 2012]

第12章 统计学习方法总结

方法	适用问题	模型特点	模型	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	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			
朴素贝叶斯	多类分类	特征与类别的联合概率分布， 条件独立假设	生成	极大似然估计 极大后验估计	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	正则化的极大似然估计	对数似然损失	特征选择， 生成， 剪枝
Logistic 回归 最大熵模型	多类分类	特征条件下类别的条件概率分布， 对数线性模型	判别	极大似然估计,正则化的极大似然估计	Logistic 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	极小化正则化的合页损失， 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	弱分类器的线形组合	判别	极小化加法模型的指数损失	指数损失	前向分布加法算法
EM 算法	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	极大似然估计， 极大后验估计	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	极大似然估计， 正则化的极大似然估计	对数似然损失	改进的迭代尺度算法， 梯度下降， 拟牛顿法

表格 12-1 统计学习 10 种方法特点的概括总结

参考文献

- [Borman, 2004] Borman S. The expectation maximization algorithm-a short tutorial [J]. Submitted for publication, 2004, 41.
- [Charles, 2011] Charles Sutton and Andrew McCallum, An Introduction to Conditional Random Fields [J]. Machine Learning 4.4 (2011): 267-373.
- [Determined22, 2017] Determined22, http://www.cnblogs.com/Determined22/p/5776791.html , 2017.
- [Duda, 2003] Duda R O, Peter E Hart, etc. 李宏东等译。模式分类 [M]. 机械工业出版社。2003.
- [Friedman, 2001] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics, vol. 29, no. 5, 2001, pp. 1189–1232.
- [Friedman, 2001] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning [M]. New York: Springer series in statistics, 2001.
- [Goodfellow, 2017] Goodfellow I, Bengio Y, Courville A. 深度学习 [M]. 人民邮电出版社。2017.
- [Hagan, 2006] Martin T. Hagan. 戴葵等译。神经网络设计 [M]. 2002.
- [Haykin, 2011] Haykin S . 神经网络与机器学习 [M]. 机械工业出版社。2011.
- [Hyvarinen, 2007] Aapo Hyvarinen, Juha Karhunen. 周宗潭译 独立成分分析 [M]. 电子工业出版社。2007.
- [Mitchell, 2003] Tom M.Mitchell. 肖华军等译。机器学习 [M]. 机械工业出版社。2003
- [Rabiner, 1989] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.
- [Samuel, 2007] Samuel Karlin M.Taylor 著 ,庄兴无等译。 随机过程初级教程。[M]. 人民邮电出版社， 2007.
- [Sutton, 2012] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." Foundations and Trends® in Machine Learning 4.4 (2012): 267-373.

- [周志华，2018] 周志华 机器学习 [M]. 清华大学出版社。2018.
- [苏剑林，2017] 苏剑林，https://spaces.ac.cn/archives/4277 ，2017.
- [史春奇，2017] 史春奇，https://www.jianshu.com/u/0cb11e6948c6，2017.
- [盛骤，2015] 盛骤等编，概率论与数理统计（第四版）。 [M]. 高等教育出版社。2015.
- [宗成庆，2018] 宗成庆著，统计自然语言处理（第二版）。 [M]. 清华大学出版社。 2018.

符号说明

- Pxx，代表第 xx 页；
- Cxx，代表第 xx 章；
- [M]，代表图书；
- [J]，代表杂志；