机器学习 1-3 章温习总结

第一章

机器学习所研究的内容 说白了是一种训练计算机学习的“学习算法”，即如何教会机器自己学习。

机器学习始于训练样本，但是目标是让这个模型很好的适用于“新样本”，即“泛化能力”，故而有时过拟合是难以避免而且对泛化能力产生不良影响的学习结果。

学习过程：再所有假设组合的空间中进行探索，求得与训练集最最 fit 的假设。最后这个假设往往就是学习所得的结果，但是有可能存在多个不同的假设都很 fit 训练集，这时则考虑归纳偏好。

归纳偏好由奥卡姆剃刀进行引导，即选择“更平滑”,“更简单”的一个假设，这是常用的，自然科学研究中最基本的原则。

归纳学习：从样例中学习，狭义的归纳学习是从训练数据中学得概念（应该是更为苛刻的要求，因为有限的训练集所得出的概念往往很难适用于所有样本，即很难有很强的泛化能力）

没有免费的午餐定理：看起来难以理解，实际上通过书上计算可得出对于不同学习算法，总误差的期望性能与学习算法无关。即所有的算法都和胡猜拥有一样的可能产生误差大小的可能性。所以需要针对具体问题具体分析，找到不同问题偏重的方面。

人工智能从“推理期”——“知识期”到面临“知识工程瓶颈”，机器学习是研究发展到一定阶段的必然产物。

第二章

NP 难：P 问题是可以在多项式时间内被确定机(通常意义的计算机)解决的问题，

NP(Non-Deterministic Polynomial, 非确定多项式)问题,是指可以在多项式时间内被非确定机(他可以猜,他总是能猜到最能满足你需要的那种选择,如果你让他解决 n 皇后问题,他只要猜 n 次就能完成----每次都是那么幸运)解决的问题。

即不是所有机器学习的任务都是能再多项式时间内解决（是 NP 问题），但是有效的学习算法是 P 问题，NP!=P。

自助法：用于数据集较小，难以有效划分训练/测试集是有用，对集成学习等方法有很大的好处，但是改变了初始数据集的分布，引入估计偏差。

留出法、交叉验证：用于初始数据量充足。

泛化能力性能度量：回归：均方误差

分类任务：accuracy 精度

Error-rate 错误率

Precision 查准率

Recall 查全率

F1 （查准率和查全率的调和平均）

比较检验：

假设检验：得到一次测试错误率：二项检验

由多次留出法和交叉验证多次训练、测试：t 检验

k 折交叉验证即可用“成对 t 检验”

若测试错误率小于某个临界值，则假设不可拒绝（应该是两个比较的学习器的性能没有显著差别）

多个算法比较时运用基于算法排序的 Friedman 检验；列表排序，给出序值。

偏差：算法本身拟合能力

方差：不同的数据集对学习性能的变化

噪声：学习问题本身的难度（所有算法所能达到的期望方差下界）

第三章

线性回归：输入属性只有一个，学得一个函数使得均方误差最小化

多元线性回归则利用矩阵得运算，可能有多租解，这时由学习算法的归纳偏好决定，可以引入正则化项。

“极大似然法”令每个样本属于其真实标记得概率越大越好

线性判别分析：“类内散度矩阵”同类样例投影点得协方差

“类间散度矩阵”一类阳历的投影点尽可能远离

应用到多类时，“全局散度矩阵”，“类内散度矩阵”重定义为每个类别散度矩阵之和。

多分类学习的基本思路：拆解法

拆分策略：一对一 OvO N 个类别两两配对 N(N-1)/2 个二分类任务

一对其余 OvR N 个分类任务

多对多 MvM 纠错输出码

编码：N 个类别 M 次划分，M 个训练集训练 M 个分类器

译码：用 M 个分类器分别对测试样本进行预测，返回其中距离最小的类别作为预测的类别结果。

类别划分：编码矩阵（二元码 分为正类反类，三元码 分为正类反类+停用类）

（对于同一个学习任务，ECOC 编码越长，纠错能力越强，但是编码越长，所训练的分类器越多）

Class-imbalance 分类任务中不同类别的训练样例数目差别很大。

基本策略：对预测值进行调整，进行“再缩放 rescaling”（也是代价敏感学习的基础）这是根据训练样例中正例与反例的比例来调整，而不再是直接用 0.5 作为阈值判断。

做法：1.对训练集的反类样例进行欠采样

对训练集里的正类样例进行过采样

直接基于原始训练集学习，但再用训练好的分类器进行预测时，将再缩放加入到决策过程中，——“阈值移动 threshold-moving

第四章

决策树，可以有双重含义，理解为一种分类的机器学习方法，或者是学习所得的树。

决策树类似人类面临决策问题的处理方法。

决策树的学习目的也是希望得到泛化能力强，处理未见示例能力强的决策树。

决策树和普通的分类学习方法有什么区别呢？我比较出来的是普通分类学习是直接和学习所得的假设相比较，每个属性难道是同时比较的？而决策树是对每个属性的值进行不断的分支？决策树似乎也能在学习过程中分出新的节点（即新的类别），处理未见新示例的能力更强。

学习的关键是要选择最优划分属性（即在不断的分支过程中分支节点的样本基本属于同一类，节点的“纯度”越来越高）应该也是为了分支的次数少，减少算法复杂性。

信息熵就是度量样本集合纯度最常用的一个指标。（可以利用信息增益优先选择能使分支节点的纯度更高的那个属性，优先进行分支，然后不断递归）

根据计算公式可以得出一个漏洞：信息增益准则对可取值数目较多的属性有所偏好。比如瓜的编号，每个瓜都对应一个不同的编号，这个属性虽然信息增益最大，却不对泛化能力产生贡献，即没有意义的。所以改用增益率来选择最优化分属性。

但是增益率也并非尽善尽美，它对可取值数目较少的属性有所偏好。为了中和两者偏好，可以先选择信息增益高于平均水平的，在选择增益率最高的属性。

基尼指数：本质是一个概率，其值越小数据集 D 的纯度越高，选择划分后基尼指数最小的属性作为最优划分属性。

剪枝对付过拟合，分为预剪枝和后剪纸。用留出法对泛化性能是否提升进行判断。

但是预剪枝可能出现的问题是，一旦使精度降低就禁止，导致后续分支可能会显著提高的机会被抹杀了。导致学习的不够充分，欠拟合。

后剪枝保留更多的分支，而且最终精度比预剪枝高，但是训练开销大的多。

连续值解决策略：采用二分法对连续属性进行处理（一分为二）

（与离散属性不同，当前节点划分属性为连续属性时，该属性还可以作为其后代结点的划分属性）

缺失值解决策略：（样本的某些属性值缺失）（1）推广信息增益（为每个样本赋予权重）（2）若样本 x 在划分属性 a 上取值未知，则将 x 同时划入所有子节点，同时调整样本在与属性 a 对应的子节点中的权值

多个属性描述的样本经过决策树分类即找到空间中的分类边界（轴平行）

解决策略：斜划分（不是对某个属性，而是对属性的线性组合进行测试）（将多个属性组合为一个变量易于比较）

第五章

感知机学习过程中可以将阈值看作一个固定输入为-0.1 的“哑结点”，（和离散数学中的哑元类比），这样可以将阈值和权重统一学习，是一个简化学习的方法。

对于感知机，如果学习的二分类模式线性可分问题，则存在一个线性超平面可以将这两个类别分开，对应到感知机学习上就是学习过程会收敛，从而可以求得适当的权向量（连接权重和阈值），否则学习过程会产生振荡。所以异或问题不能由感知机学习解决，因为它是非线性可分问题，而多层网络学习能力可以解决该类问题。

BP 算法用于训练多层前馈神经网络，是基于梯度下降的策略。（即沿着目标的负梯度方向进行参数调整，目标就是网络在训练示例上的均方误差）

学习率控制算法每一轮迭代的更新步长： 学习率过大会导致学习过程中震荡，太小则各个参数的收敛速度过慢。

但是为什么令连接权和阈值的学习率不同，可以作为后续的精细调节？可能是因为有时候两者变化的步长不同，可以有更多连接权和阈值的匹配，更容易达到最佳参数值。

累积 BP 算法

BP 算法的目标是最小化训练集 D 上的累积误差：

而标准 BP 算法每输入一个样例就进行一次更新，所以它的参数更新非常频繁，而且不同样例可能会对更新起到抵消效果，从而使得模型需要更多次迭代才能到达累积误差的极小点。

标准 BP 算法和累积 BP 算法的区别类似于随机梯度下降和标准梯度下降的区别。

BP 算法的强大表示能力可能导致过拟合。解决方法有早停和正则化两种方法。

为什么正则化方法可以减少过拟合呢？

下面是知乎的解答，应该就是利用正则化方法对模型参数添加先验，降低模型空间复杂度，使得其对于训练数据的噪声等干扰受到的影响较小。具体如何先验？高斯分布协方差这部分不太理解。

过拟合发生的本质原因，是由于监督学习问题的不适定：在高中数学我们知道，从n个（线性无关）方程可以解n个变量，解n+1个变量就会解不出。在监督学习中，往往数据（对应了方程）远远少于模型空间（对应了变量）。因此过拟合现象的发生，可以分解成以下三点：

1. 有限的训练数据不能完全反映出一个模型的好坏，然而我们却不得不在这有限的数据上挑选模型，因此我们完全有可能挑选到在训练数据上表现很好而在测试数据上表现很差的模型，因为我们完全无法知道模型在测试数据上的表现。
2. 如果模型空间很大，也就是有很多很多模型可以给我们挑选，那么挑到对的模型的机会就会很小。
3. 与此同时，如果我们要在训练数据上表现良好，最为直接的方法就是要在足够大的模型空间中挑选模型，否则如果模型空间很小，就不存在能够拟合数据很好的模型。

由上3点可见，要拟合训练数据，就要足够大的模型空间；用了足够大的模型空间，挑选到测试性能好的模型的概率就会下降。因此，就会出现训练数据拟合越好，测试性能越差的过拟合现象。

过拟合现象有多种解释，

- 经典的是bias-variance decomposition，但个人认为这种解释更加倾向于直观理解；
- PAC-learning 泛化界解释，这种解释是最透彻，最fundamental的；
- Bayes先验解释，这种解释把正则变成先验，在我看来等于没解释。

另外值得一提的是，不少人会用“模型复杂度”替代上面我讲的“模型空间”。这其实是一回事，但“模型复杂度”往往容易给人一个误解，认为是一个模型本身长得复杂。例如5次多项式就要比2次多项式复杂，这是错的。因此我更愿意用“模型空间”，强调“复杂度”是候选模型的“数量”，而不是模型本事的“长相”。

最后回答为什么正则化能够避免过拟合：因为正则化就是控制模型空间的一种办法。

深度学习使用到的一个训练多隐层网络手段是：无监督逐层训练，也叫做“预训练+微调”，这样可以将大量参数分组，先解决各个组的较优设置问题，最后在联合起来寻找全局最优。

第六章

求凸二次规划问题可以直接利用现成的优化计算包求解，也可以用拉格朗日乘子法求得其“对偶问题”。如何求得对偶问题中的拉格朗日乘子值？

这是帮助理解的博客网页，遗憾的是依旧不是很能理解推导过程，这是哪部分的数学知识？范数似乎是数论中的知识。

总之利用 SMO 高效算法解出对偶问题求得拉格朗日乘子后可求出 w 与 b（偏移项），其中偏移项 b 可以直接使用所有支持向量求解的平均值。（由图象理解，所有支持向量一般是均匀分布在超平面附近，利用他们求解后的平均值一般和超平面的偏移项的值相差无几）

https://www.cnblogs.com/huahuahu/p/la-ge-lang-ri-cheng-zi-fa-yu-dui-ou-wen-ti.html

对拉格朗日乘式的理解

假设目标函数在 $\boldsymbol{x}^*$ 处取得极值，那么对于这个点，所有的$h_i(\boldsymbol{x}^*)=0$，某些$g_j(\boldsymbol{x}^*)=0$，某些$g_j(\boldsymbol{x}^*)<0$。

1. 对于所有的$h_i(\boldsymbol{x})$都有如下结论：
   - 对于任意点，$\nabla h(\boldsymbol{x})$ 和约束曲面正交。
   - 在最优点，$\nabla f(\boldsymbol{x}^*)$ 也和约束曲面正交。

   结合以上两点，存在λ，使得

   $$\nabla f(\boldsymbol{x}^*)+\lambda\nabla h(\boldsymbol{x}^*)=0 \tag{4}$$

   对 (2) 式的前两项对 $\boldsymbol{x}$ 求导，即可得到上面的式子；对 λ 求导得到约束条件 $h(\boldsymbol{x})=0$。即当 (2) 式前两项取得极值时，必满足条件 $h_i(\boldsymbol{x})=0$。
2. 对于某些 $g_j(\boldsymbol{x})$，满足$g_j(\boldsymbol{x}^*)=0$，此时必定对应的 $\mu_j>0$
   由于在 $\boldsymbol{x}^*$处取得极值，因此必定有$\nabla f(\boldsymbol{x}^*)+\mu_j\nabla g(\boldsymbol{x}^*)=0$。
   如果$\mu_j<0$，即两个函数的梯度方向相同，那么可以沿着$g_j(\boldsymbol{x})\leqslant 0$ 的方向移动$\boldsymbol{x}$，此时$f(\boldsymbol{x})$必定会减小，不符合已取得极小值的假设。
3. 对于某些 $g_j(\boldsymbol{x})$，满足$g_j(\boldsymbol{x}^*)<0$，此时必定对应的 $\mu_j=0$
   若$\mu_j\neq 0$，则$\boldsymbol{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$和$f(\boldsymbol{x})$值不相同。

核函数是可以将非线性可分的问题，利用将样本从原始空间映射到更高维特征空间的方法。令样本在这个特征空间线性可分。模型的最优解可以通过训练样本的和函数展开——支持向量展式。

核函数和映射之间的对应关系由定理规定，对于一个半正定和矩阵，总能找到一个与之对应的映射。（虽然并不怎么理解其中原理，但作为一个定理的确可以免去一些存在性求解的麻烦）

软间隔支持向量的最终模型仅仅与支持向量有关，即通过采用 hinge 损失函数仍然保持了稀疏性。

为了缓解过拟合问题，引入软间隔的概念。

稀疏性指什么？是指支持向量机的解，即支持向量是有限的？

为什么只与支持向量有关？是因为只有支持向量的位置可以决定超平面的位置？

对于上面的问题没有查到相关答案，是否是一个总结性的定义？

正则化似乎的确可以削减复杂度，在优化支持向量机中降低了最小化训练误差的过拟合风险中也有体现。

“核化”即引入核函数，可用于将线性学习器拓展为非线性学习器。