（书籍为李航的《机器学习方法》）

ML（李航书）—note(只到第七章,深度学习转：cs231n 课程)

https://datawhalechina.github.io/statistical-learning-method-solutions-manual/#/README（书籍习题解答）

1 机器学习与监督学习方法概论

1. 机器学习三要素：**模型**（模型的假设空间（hypothesis space）包含所有的可能概率分布或决策函数）+ **策略**（strategy）(loss function，经验风险最小化（empirical risk minimization , ERM）：对 loss function 取期 s 望 ； 结构风险最小化(structural ,SRM)：ERM+正则化) + **算法**（怎么样在模型空间中以策略找到最好模型）
2. 一般的，极大似然估计 ——ERM ； 贝叶斯估计中的最大后验概率——SRM。

方法	目标	对应策略	特点
极大似然估计（MLE）	$\max P(\text{Data}\|\theta)$	**ERM**（无正则化）	只关注数据拟合，可能过拟合
最大后验估计（MAP）	$\max P(\theta\|\text{Data})$	**SRM**（带正则化）	结合先验，控制复杂度，提升泛化

MLE/ERM：只追求在训练数据上“表现最好”，不考虑模型复杂度。MAP/SRM：在追求训练数据表现的同时，通过先验（正则化）对模型复杂度进行惩罚，避免过拟合。

这种对应关系揭示了频率学派（MLE）与贝叶斯学派（MAP）在机器学习优化目标上的内在联系，也解释了为什么正则化在机器学习中如此重要：它等价于引入了一种先验信念，即参数应该较小或稀疏。

2 感知机（1957）

1. f = sign(w*x + b) 单层感知机不能表示异或
2. （算法原始形式）：当实例点误分类时，改变 w,b 将分离超平面向该点移动，减少距离，直到该点正确分类。
3. （对偶形式）：因为用梯度下降时，w 与 b 对应的梯度是可以提前算的，将提前的梯度代入 loss，引出 Gram 矩阵，用空间换时间。
4. 定理：样本集线性可分的充分必要条件是正实例点所构成的凸壳与负实例点所构成的凸壳互不相交。

> 设集合$S \subset R^n$，是由R^n中的k个点所组成的集合，即$S=\{x_1, x_2, \cdots, x_k\}$。定义$S$的凸壳$\text{conv}(S)$为：
>
> $$\text{conv}(S)=\{x=\sum_{i=1}^{k}\lambda_i x_i \Big| \sum_{i=1}^{k}\lambda_i=1, \lambda_i \geqslant 0, i=1,2,\cdots,k\}$$

3 K-NN(k-nearest neighbor)（1968）

1. 对拿到的新实例，跟训练数据集做距离度量，取 K 个最近实例类别，多数表决。（不具有显式的学习方法）
2. K 值选择（一般交叉验证）+距离度量+分类决策 是 K-NN 的三个基本要素
3. kd tree(是存储 k 维空间的树结构) (重点在建树+搜树主要用法)

4 朴素贝叶斯（naive Bayes）

1. 基于特征条件独立假设(naive 的由来)，学习输入输出的联合分布。基于此模型，给定的 x，求 max（后验概率）——>y
2. 取 0/1 loss 做 ERM——>后验概率最大化准则
3. 该算法，学习意味着估计 P(Y = Ck)和 P(X j = x j | Y = Ck)，极大似然估计→简单计频（可能出现 P= 0，不能除 P）；贝叶斯估计→加了平滑的计频 第 4 章 朴素贝叶斯法

5 决策树（decision tree）(1986-1993)

1. 特征选择→决策树生成→决策树修剪
2. 条件熵(conditional entropy) : H (Y|X) def :X 给定条件下，Y 的条件概率分布的熵 对 X 的数学期望。

$$E(Y|X) = \sum_{i=1}^{m} p(X = x_i)E(Y|X = x_i)$$

维度	ID3	C4.5	CART
任务类型	仅分类	仅分类	分类 + 回归
划分标准	信息增益	信息增益率	分类：基尼系数；回归：平方误差
处理连续值	❌ 不支持	✅ 支持（二分切点）	✅ 支持（二分切点）
处理缺失值	❌ 不支持	✅ 支持（权重法）	✅ 支持（代理分裂）
剪枝方式	无系统剪枝	悲观误差剪枝（后剪枝）	代价复杂度剪枝（后剪枝）
分支类型	多叉树	多叉树	严格二叉树
特征多次使用	✅ 可重复	✅ 可重复	✅ 可重复
树结构	树内节点可有多子树	树内节点可有多子树	每个内部节点仅左右两个孩子
输出结果	类别标签	类别标签 + 概率	分类：类别/概率；回归：实数值
算法作者	Quinlan 1986	Quinlan 1993	Breiman 1984

其中：ID3（Iterative Dichotomiser 3，迭代二分器 3）**CART（Classification and Regression Trees，分类与回归树）**

6 logistic regression(逻辑地斯回归) 与（maximum entropy model）

1. 理论来源与 logistic 分布，让其参数化

二项逻辑回归模型是如下的条件概率分布

$$P(Y=1|X)=\frac{exp(w\cdot x+b)}{1+exp(w\cdot x+b)}$$

$$P(Y=0|X)=\frac{1}{1+exp(w\cdot x+b)}$$

2. 最大熵原理：在学习概率模型时，该分布的熵最大，认为该模型最好。
3. **单层的 Logistic 回归无法表示异或（XOR）**；把两层 Logistic 级联（即加隐藏层）就可以。(同感知机一样)
4. 最大熵模型学习中的对偶函数极大化，等价于，模型极大似然估计。

1. 最大熵模型的学习目标

最大熵模型的学习目标是：在满足所有已知约束（特征函数的期望等于经验分布的期望）的条件下，选择熵最大的概率分布。

数学上，这可以写成：

$$\max_{P\in\mathcal{P}} H(P)=-\sum_x P(x)\log P(x)$$

其中，$\mathcal{P}$ 是满足以下约束的分布集合：

$$\mathbb{E}_P[f_i]=\mathbb{E}_{\tilde{P}}[f_i],\quad i=1,\ldots,n$$

这里，f_i 是特征函数，$\tilde{P}$ 是经验分布。

2. 对偶问题与极大似然估计

最大熵模型的原始问题是一个约束优化问题。通过引入拉格朗日乘子，我们可以得到它的对偶问题。

对偶函数（即拉格朗日对偶函数）为：

$$L(\lambda)=\log Z(\lambda)-\sum_i \lambda_i \mathbb{E}_{\tilde{P}}[f_i]$$

其中，$Z(\lambda)$ 是配分函数（归一化因子）：

$$Z(\lambda)=\sum_x \exp\left(\sum_i \lambda_i f_i(x)\right)$$

最大化对偶函数 $L(\lambda)$ 等价于最大化训练数据的对数似然函数。

具体来说，最大熵模型的对偶函数极大化，等价于如下形式的极大似然估计：

$$\max_\lambda \sum_x \tilde{P}(x)\log P_\lambda(x)$$

其中，$P_\lambda(x)$ 是最大熵模型的形式：

$$P_\lambda(x)=\frac{1}{Z(\lambda)}\exp\left(\sum_i \lambda_i f_i(x)\right)$$

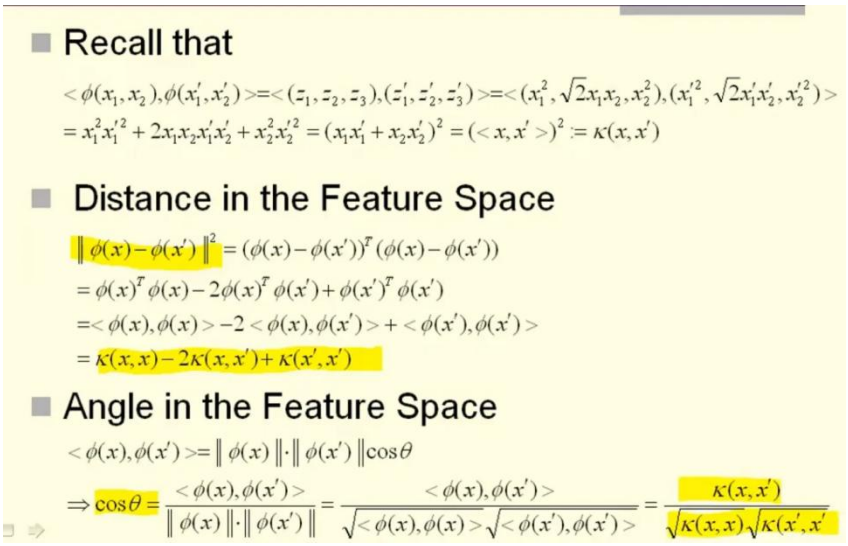
5. 同时，最大熵模型与 logistic regression 有类似的形式，对数线性模型，所以优化方法可以同用：迭代尺度法、梯度下降、牛顿法等（书附录）

7 支持向量机（support vector machines）(SVM)

1. 在特征空间中**间隔最大**的线性分类器**间隔最大就与感知机有区别了**

维度	线性可分 SVM	近似线性 SVM	非线性可分 SVM
数据分布	正负类能被一条硬带完全分开	大部分可分，但少量噪声/异常点重叠	本质重叠，任何直线都无法分开
核心思想	最大几何间隔	最大间隔 + 容忍违例	升维映射 → 在新空间做线性 SVM
关键公式	$\min \frac{1}{2}\|w\|^2$ s.t. $y_i(w\cdot x_i+b)\geq 1$	$\min \frac{1}{2}\|w\|^2 + C\sum\xi_i$ s.t. $y_i(w\cdot x_i+b)\geq 1-\xi_i,\ \xi_i\geq 0$	$\min \frac{1}{2}\|w\|^2 + C\sum\xi_i$ s.t. $y_i(w\cdot\varphi(x_i)+b)\geq 1-\xi_i$
引入变量	无	松弛变量 ξ_i（软间隔）	松弛变量 ξ_i + 非线性映射 $\varphi(\cdot)$
惩罚参数	无	C 越大→越不允许越界	C 同上，兼顾间隔与误分类
对偶体现	$\alpha_i(y_i\ y_j\ x_i\cdot x_j)$	$0\leq\alpha_i\leq C$	$0\leq\alpha_i\leq C$，内积→核函数 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$
常用核	不需要（线性核）	线性核即可	RBF、多项式、Sigmoid 等
支持向量	落在间隔带外边界上的点	间隔带内、误分类点都可能成为 SV	同左，但在高维特征空间计算
示意图	一条实线+两条虚线，无点越界	实线两侧少量点进入间隔或越界	二维圆圈数据，高维后线性分开
一句话记忆	硬间隔，最严格	软间隔，容小错	核技巧，曲线救国

2. SVM 可形式化成一个求解凸二次规划(convex quadratic programming)问题，也等价于正则化的合页损失函数的最小化问题。
3. 非线性可分问题：在 Rn 空间中能用一超曲面将正负实例分开。
4. 核函数（kernel function）如果支持向量机的求解只用到内积运算，而在低维输入空间又存在某个函数 K(x, x′) ，它恰好等于在高维空间中这个内积，即 K(x, x′) =ϕ(x) ·ϕ(x′) ; 。那么支持向量机就不用计算复杂的非线性变换，而由这个函数 K(x, x′) 直接得到非线性变换的内积，使大大简化了计算。这样的函数 K(x, x′) 称为核函数。



只要有了 kernel，那么就能有距离和角度可算。kernel 算子的具体形式并不重要。

5. 一函数 K，定义在 X*X,其对应的 Gram 矩阵，对于任意的 Xi 属于 Rn,是半正定的↔K 可以做核函数(正定核)。(书 P115-118 证明)

证明 必要性。由于 $K(x,z)$ 是 $\mathcal{X}\times\mathcal{X}$ 上的正定核，所以存在从 $\mathcal{X}$ 到希尔伯特空间 $\mathcal{H}$ 的映射 ϕ，使得

$$K(x,z)=\phi(x)\cdot\phi(z)$$

于是，对任意 $x_1,x_2,\cdots,x_m$，构造 $K(x,z)$ 关于 $x_1,x_2,\cdots,x_m$ 的 Gram 矩阵：

$$[K_{ij}]_{m\times m}=[K(x_i,x_j)]_{m\times m}$$

对任意 $c_1,c_2,\cdots,c_m\in\boldsymbol{R}$，有

$$\begin{aligned}\sum_{i,j=1}^{m}c_ic_jK(x_i,x_j)&=\sum_{i,j=1}^{m}c_ic_j(\phi(x_i)\cdot\phi(x_j))\\&=\left(\sum_i c_i\phi(x_i)\right)\cdot\left(\sum_j c_j\phi(x_j)\right)\\&=\left\|\sum_i c_i\phi(x_i)\right\|^2\geqslant 0\end{aligned}$$

表明 $K(x,z)$ 关于 $x_1,x_2,\cdots,x_m$ 的 Gram 矩阵是半正定的。

充分性。对任意 $x_1,x_2,\cdots,x_m\in\mathcal{X}$，已知对称函数 $K(x,z)$ 关于 $x_1,x_2,\cdots,x_m$ 的 Gram 矩阵是半正定的。根据前面的结果，对给定的 $K(x,z)$，可以构造从 $\mathcal{X}$ 到某个希尔伯特空间 $\mathcal{H}$ 的映射：

$$\phi:x\to K(\cdot,x)\tag{7.86}$$

由式 (7.83) 可知：

$$K(\cdot,x)\cdot f=f(x)$$

并且

$$K(\cdot,x)\cdot K(\cdot,z)=K(x,z)$$

由式 (7.86) 即得：

$$K(x,z)=\phi(x)\cdot\phi(z)$$

表明 $K(x,z)$ 是 $\mathcal{X}\times\mathcal{X}$ 上的核函数。

定理给出了正定核的充要条件，因此可以作为正定核，即核函数的另一定义。

6. SMO(sequential minimal optimization)序列最小优化算法（1998）。P121 第一遍学先略过。