机器学习原理

(基于统计学习方法)

温润泽

2023.3.22

统计学习方法

1 感知机

先任意选取一个超平面，然后用梯度下降法不断极小化目标函数.在这个过程中一次随机选取一个误分类点使其梯度下降，经过有限次搜索可以找到将训练数据完全正确分开的分离超平面

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

其中$x_i\in\mathcal{X}=\mathbf{R}^i$， $y_i\in\mathcal{Y}=\{-1,1\}$， $i=1,2,\cdots,N$,求参数w ,b ，以及是以下损失函数得最小值，其几何含义即为所有误分类点与分解面得距离之和最小

$$min_{w,b}\,L(w,b)=-\sum_{x_i\in M}y_i(w\bullet x_i+b)$$

如果 $y_i(w\cdot x_i+b)\leqslant 0$

$$w\leftarrow w+\eta y_i x_i$$
$$b\leftarrow b+\eta y_i$$

其中M为误分类点得集合。学习率（步长）$\eta(0\ <\ \eta\ \leq\ 1)$ 。本质上是不停调整分界线，使整个平面上不存在误分类点。当训练数据集线性可分时，感知机学习算法原始形式法代是收敛的，训练集线性不可分时，感知机学习算法不收敛， 法代结果 会发生震荡

计算方便可采用对偶形式：令w ,$b\ =0$，经过n次修改，$\alpha_i=n_i\eta$.

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i$$

$$b=\sum_{i=1}^{N}\alpha_i y_i$$

带入原始形式其余一致，初值为0，迭代次数应该选大一点

2 k 近邻法

定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的k个实例，这k个实例的多数属于某个类，就把该输入实例分为这个类。k近邻法没有显式的学习过程

k 近邻模型对应于基于训练数据集对特征空间的一个划分 ki丘邻法中， 当训练集、距离度量、 k 值 及分类决策规则确定后，其结果唯一确定

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. kd 树是二叉树，表示对 k 维空间 的一个划分，其每个结点对应于 k 维空间划分中的 一个超矩形区域

3 朴素贝叶斯法

首先基于特征条件独立假设学习输入/输出的联合概率分布:然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出y；后验概率最大即等价于0-1损失函数的期望风险最小

贝叶斯定理:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

其中A为类别，B为特征。

朴素：条件独立性假设指各个特征之间相互独立，但该假设较强让方法简明的同时，可能丧失部分准确性

$$P(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} \mid Y = c_k) = \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

朴素贝叶斯分类方法的基本公式:

$$P(Y = c_k \mid X = x) = \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}{\sum_k P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}, \quad k = 1, 2, \cdots, K$$

分类只关心类别$Y = c_k$,分母与Y的取值无关，为一个常数，所以定义朴素贝叶斯分类器:

$$y = \arg\max_{c_i} P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

算法:

1. 计算先验概率$P(Y = c_k)$和条件概率$P(X^{(j)} = a_{jl} \mid Y = c_k)$
2. 对给定的实例x计算$P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} \mid Y = c_k)$
3. 取最大后验概率确定x的类型c_k

4 决策树

4.1 概念

决策树是从训练数据中归纳出一套完整且互斥的分类规则，本质上是由数据集估计出条件概率模型；决策树学习的损失函数一般用正则化的极大似然函数，追求损失函数最小；最有决策树是NP complete问题，现实只能获得次优解 算法常用ID3、C4.5、CART

特征选择：可以在学习开始对特征进行选择，只留下足够训练的特征，节省计算量

决策树生成：构建过程为递归地选择最优特征并根据该特征分割训练数据，从根节点地数据选一个最优特征把根节点地数据分成子集，继续分别对各个子集进行各自最优特征地分割，直到把所有数据被分道叶节点上

剪枝：为防止过拟合，需要对生成地决策树进行修剪，使其具有更好的泛化能力。即去掉过于精细的子节点，把数据退回父节点上

4.2 特征选择

特征选择的准则是选择信息增益最大的特征，其往往有更强的分类能力；也可用信息增益比进行进一步校正，信息增益(information gain)表示知道特征A的信息而使得经验熵（不确定性）减小的程度

$$\text{增益：}\quad g(D,A)=H(D)-H(D\mid A)$$

$$\text{熵}H:\quad (p)=-\sum_{i=1}^{n}p_i\log p_i$$

$$\text{条件熵：}\quad H(Y\mid X)=\sum_{i=1}^{n}p_iH(Y\mid X=x_i)$$

$$\text{增益比：}\quad g_R(D,A)\ =\ \frac{g(D,A)}{H(D)}$$

p为随机变量X的概率分布，熵H(P)越大，变量的不确定性就大$0\leqslant H(p)\leqslant\log n$

随机变量 X 给定的条件下随机变量 Y 的是件熵 (conditionaJ entropy)$H(Y|X)$ ， 定义为 X 给定条件下 Y 的条件概率分布的娟对 X 的数学期望。

5 SVM支持向量机

5.1 基本概念

支持向量机(Support Vector Machine, SVM)是一种二分类学习模型,它的基本模型是定义在特征空间上的间隔最大的线性分类器,基本想法是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线性可分的训练数据集而言，线性可分离超平面有无穷多个(等价于感知机)，但是几何间隔最大的分离超平面是唯一 的。SVM适用于处理线性可分数据,通过使用不同的核函数可扩展到非线性分类问题

当输入空间为欧氏空间或离散集合、特征空间为希尔伯特空间时，核函数 （kernel function）表示将输入从输入空间映射到特征空间得到的特征向量之间的内积通过使用核函数可以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，这样的方法称为核技巧

函数间隔 单个数据点 $\hat{\gamma}_i = y_i(w \bullet x_i + b)$ 数据集 $\hat{\gamma} = \min\limits_{i=1,\cdots,N} \hat{\gamma}_i$

几何间隔 $\gamma_i = y_i \left(\frac{w}{\| w \|} \cdot x_i + \frac{b}{\| w \|} \right)$ 数据集 $\gamma = \min\limits_{i=1,\cdots,N} \gamma_i$

其中y为标记函数，在超平面正一侧是为+1，负一侧为-1

可知几何间隔为函数间隔的归一化 $\gamma = \frac{\hat{\gamma}}{\| w \|}$ 如果超平面参数 w 和 b 成比例地改变(超平面没有改变)，函数间隔也按比例改变，而几何间隔不变

支持向量（support vector） 线性可分的情况下训练数据集的样本点中与分离超平面距离最近的样本点的实例，使约束条件等于0，即满足 $y_i(w \cdot x_i + b) - 1 = 0$,分布在 H_1 和 H_2上

$$H_1 : w \cdot x + b = 1 \qquad y_i = +1$$

$$H_2 : w \cdot x + b = -1 \qquad y_i = -1$$

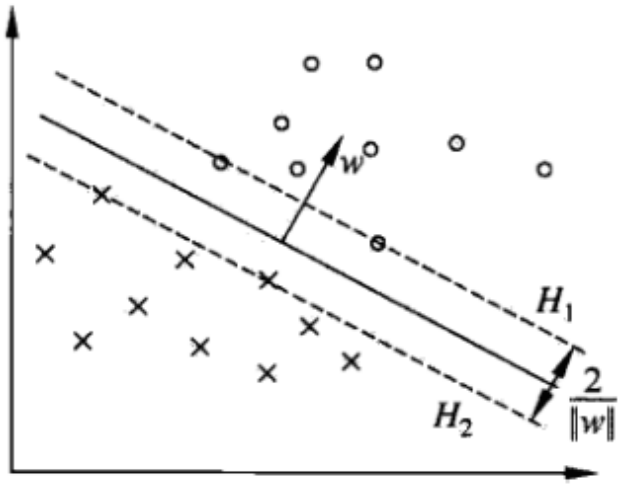

图 7.3　支持向量

在决定分离超平面时只有支持向量起作用

5.2 最大间隔法

$$\max_{w,b} \frac{\hat{\gamma}}{\| w \|} \qquad \text{s.t.} \quad y_i(w \bullet x_i + b) \geqslant \hat{\gamma}\,, \quad i = 1, 2, \cdots, N$$

希望最大化超平面(w, b) 关于训练数据集的几何间隔 ，约束条件表示的 是超平面(w, b) ，的关于每个训练样本点的几何间隔至少是γ

函数间隔 $\hat{\gamma}$ 的取值不影响优化问题的解，函数间隔等比例变化时，w，b也等比例变化，对约束和比例式均无影响，遂取 $\hat{\gamma}$=1；最大化$\frac{1}{\| w\|}$ 与最小化 $\frac{\|w\|^2}{2}$是等价的

得线性可分支持向量机学习的最优化问题：

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \quad y_i \in \mathcal{Y} = \{-1, +1\}\,, \ i = 1, 2, \cdots, N$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad \text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geqslant 0\,, \quad i = 1, 2, \cdots, N$$

可解得最优化问题的解 w^*, b^* ，最大间隔分离超平面 $w \cdot x + b = 0$ ，分类决策函数 $f(x) = \text{sign}(w^* \cdot x + b^*)$，即线性可分支持向量机模型

5.3 对偶学习算法

对偶转化的详细原理

等式约束可以用拉格朗日转变为非约束问题，不等式约束用KKT条件转化为等式约束问题；对偶问题更容易求解，自然地引入核函数；定义拉格朗日函数，其中 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{_N})^{\mathrm{T}}$为拉格朗日乘子向量

$$L(w, b, \alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{N} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i$$

原始问题的对偶问题是极大极小问题

$$\max_{\alpha} \min_{w,b} L(w, b, \alpha)$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \qquad \text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geq 0\,, \quad i = 1, 2, \cdots, N$$

根据对偶问题的解 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$. ，选择α^*的一个正分量$\alpha_j^* > 0$可以获得原始问题的解

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$$

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

分离超平面

$$\sum_{i=1}^{N} \alpha_i^* y_i (x \bullet x_i) + b^* = 0$$

分类决策函数只依赖于输入 x 和训练样本输入的内积

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i (x * x_i) + b^* \right)$$

对于线性不可分数据集，引入惩罚项和调节项C，进行软分界，核心思想是使间隔尽量大，误分项尽量少；

（1）选择惩罚参数 $C>0$，构造并求解凸二次规划问题

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1,2,\cdots,N$$

求得最优解 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)^{\mathrm{T}}$.

（2）计算 $w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$

选择 α^* 的一个分量 α_j^* 适合条件 $0 < \alpha_j^* < C$，计算

$$b^* = y_j - \sum_{i=1}^{N} y_i \alpha_i^* (x_i \cdot x_j)$$

（3）求得分离超平面

$$w^* \cdot x + b^* = 0$$

分类决策函数：

$$f(x) = \text{sign}(w^* \cdot x + b^*)$$

对于非线性问题，利用核变化升维后转变为线性问题，一般选择一致的核函数（gram矩阵半正定）

（1）选取适当的核函数 $K(x,z)$ 和适当的参数 C，构造并求解最优化问题

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i \tag{7.95}$$

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{7.96}$$

$$0 \leqslant \alpha_i \leqslant C, \quad i=1,2,\cdots,N \tag{7.97}$$

求得最优解 $\alpha^* = (\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$.

（2）选择 α^* 的一个正分量 $0<\alpha_j^*<C$，计算

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i K(x_i \cdot x_j)$$

（3）构造决策函数：

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i K(x \cdot x_i) + b^*\right)$$ ■

当 $K(x,z)$ 是正定核函数时，问题 (7.95) ～ (7.97) 是凸二次规划问题，解是存在的.