

机器学习导论

习题三

171840708, 张逸凯, zykhelloha@gmail.com

2020 年 4 月 25 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。¹

- (1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程；
- (2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用；**
- (3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

- (1) 请在LaTeX模板中第一页填写个人的学号、姓名、邮箱；
- (2) 本次作业需提交该pdf文件、问题4可直接运行的源码(.py文件)、问题4的预测结果(.csv文件)，将以上三个文件压缩成zip文件后上传。注意：pdf、预测结果命名为“学号_姓名”（例如“181221001_张三.pdf”），源码、压缩文件命名为“学号”，例如“181221001.zip”；
- (3) 未按照要求提交作业，提交作业格式不正确，**作业命名不规范**，将会被扣除部分作业分数；
- (4) 本次作业提交截止时间为**4月23日23:55:00**。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

¹参考尹一通老师高级算法课程中对学术诚信的说明。

1 [20pts] Decision Tree I

- (1) [5pts] 试分析使用“最小训练误差”作为决策树划分选择的缺陷。
- (2) [5pts] 树也是一种线性模型，考虑图(1)所示回归决策树， X_1, X_2 均在单位区间上取值， t_1, t_2, t_3, t_4 满足 $0 < t_1 < t_3 < 1, 0 < t_2, t_4 < 1$ ，试绘制出该决策树对于特征空间的划分。假设区域 R_i 上模型的输出值为 c_i ，试用线性模型表示该决策树。

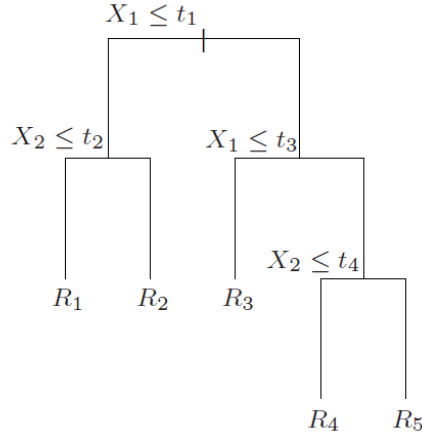


图 1: 回归决策树

- (3) [10pts] 对于回归树，我们常采用平方误差来表示回归树对于训练数据的预测误差。但是找出平方误差最小化准则下的最优回归树在计算上一般是不可行的，通常我们采用贪心的算法计算切分变量 j 和分离点 s 。CART回归树在每一步求解如下优化问题

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

其中 $R_1(j,s) = \{\mathbf{x} | x_j \leq s\}$, $R_2(j,s) = \{\mathbf{x} | x_j > s\}$ 。试分析该优化问题表达的含义并给出变量 j, s 的求解思路。

Solution. (1)

依据最小训练误差划分，也就是当前划分的依据(属性)可以带来最相似于训练集标签的结果。这样得到的决策树可以说是对训练集拟合最优的决策树。

对比其他方法(以ID3为例)，依据信息增益划分，信息增益是表示已知一个随机变量的信息后使得另一个随机变量的不确定性减少的程度，信息增益最大意味着对当前属性来进行划分所获得的“纯度提升”越大。所期望的是每个分支尽可能属于同一类别，这与分类问题中尽量提取数据的所有特征(信息)是一致的。

综上所述，依据最小训练误差划分可能使模型太“依赖”训练集了，也就是过拟合，导致模型对于测试集的泛化能力下降。

(2)

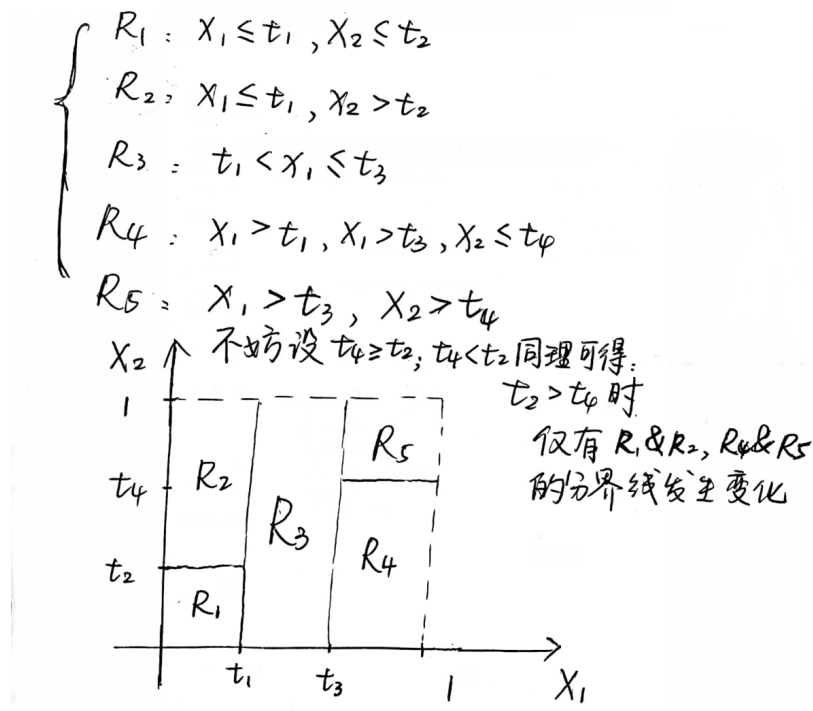


图 2: 绘制决策树对于特征空间的划分

线性模型表达: $f(x) = \sum_{i=1}^5 c_i \mathbb{I}(x \in R_i)$, 其中 $\mathbb{I}(x \in R_i) = \begin{cases} 1, & (x \in R_i) \\ 0, & (x \notin R_i) \end{cases}$

(3)

不妨令数据点 $x \in \mathbb{R}^d$, d 个属性构成了 d 维的特征空间, x 对应了 d 维特征空间一个数据点, CART 回归树的目标是将特征空间在每一维度划分成若干个子空间, 在树上叶节点 \in 某个子空间.

就像上题一样, 如果 $x \in R_i$, 那么就输出 R_i 内对应的数值.

题中所给优化问题:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

下面我们来解释这个优化问题的含义:

优化问题中 $R_1(j,s) = \{x | x_j \leq s\}$, $R_2(j,s) = \{x | x_j > s\}$, 也就是在第 j 个特征空间关于 s 的一个划分, R_1, R_2 是互补的两个区域, 我们可以发现上述优化问题就是遍历所有的划分属性 j , 然后递归地找到最优的划分点 s , 不断把当前特征空间划分成子空间直到满足终止条件.

下面给出变量 j, s 的求解思路:

- 求解内层 \min_{c_t} , $t \in \{1, 2\}$ (t 只有两个值表示不断把当前迭代步的特征空间按照划分点分成两个):

结合CART回归树的模型优化结果, 注意到在 R_t 的特征子空间里, 对上式关于 c_t 求偏导(这是一个简单的二次函数凸优化问题), 即发现对应数据点在子空间里的标签的均值:

$$\begin{aligned} & \frac{\partial}{\partial c_t} \sum_{x_i \in R_t(j,s)} (y_i - c_t)^2 \\ &= -2 \sum_{x_i \in R_t(j,s)} (y_i - c_t) \\ &= 0 \\ \Rightarrow c_t &= \sum_{x_i \in R_t(j,s)} \frac{y_i}{|R_t(j,s)|} \end{aligned}$$

• 求解外层 $\min_{j,s}$:

1. 因为特征空间的大小是有限的, 所以我们可以对 j, s 直接遍历求解: 即对题中优化问题遍历第 j 个特征空间遍历所有的切分点 s , 对应的 c_t 即用当前划分生成的子空间里的所有样本点取均值求得题中优化问题的最小值(外层 $\min_{j,s}$).

2. 上一步得到两个划分后的子空间:

$$R_1(j^*, s^*) = \{\mathbf{x} | x_j \leq s^*\}, R_2(j^*, s^*) = \{\mathbf{x} | x_j > s^*\}$$

3. 继续对这两个子空间递归地划分(执行1. 2.步骤), 直到满足停止条件, 即分支(子空间)全属于同一类别, 或者在该特征维度上取值相同, 或者包含的样本数据集合为空. 此时得到:

$$f(x) = \sum_{i=1}^K c_i \mathbb{I}(x \in R_i), \text{ 其中 } K \text{ 为子空间个数, } \mathbb{I}(x \in R_i) = \begin{cases} 1, & (x \in R_i) \\ 0, & (x \notin R_i) \end{cases}$$

(* 本题部分思路来自李航老师的统计学习方法, 但推导过程更不一样更细致)

2 [25pts] Decision Tree II

- (1) [5pts] 对于不含冲突数据 (即特征向量相同但标记不同) 的训练集, 必存在与训练集一致 (即训练误差为0) 的决策树。如果训练集可以包含无穷多个数据, 是否一定存在与训练集一致的深度有限的决策树? 证明你的结论。(仅考虑单个划分准则仅包含一次属性判断的决策树)
- (2) [5pts] 考虑如表1所示的人造数据, 其中“性别”、“喜欢ML作业”是属性, “ML成绩高”是标签。请画出使用信息增益为划分准则的决策树算法所有可能的结果。(需说明详细计算过程)
- (3) [10pts] 考虑如表2所示的验证集, 对上一小问的结果基于该验证集进行预剪枝、后剪枝, 剪枝结果是什么? (需给出详细计算过程)
- (4) [5pts] 比较预剪枝、后剪枝的结果, 每种剪枝方法在训练集、验证集上的准确率分别为多少? 哪种方法拟合能力较强?

表 1: 训练集

编号	性别	喜欢ML作业	ML成绩高
1	男	是	是
2	女	是	是
3	男	否	否
4	男	否	否
5	女	否	是

表 2: 验证集

编号	性别	喜欢ML作业	ML成绩高
6	男	是	是
7	女	是	否
8	男	否	否
9	女	否	否

Solution. (1)

存在与训练集一致的深度有限决策树;

由题中描述我们知道, 该训练集属性(特征)个数有限.

那么我们可以构造一个(带有冗余)且不满足信息增益最大等的决策树:

该树每一层都是对某一属性(特征)的完全划分, 即在该层划分之后, 这个属性(特征)将不再出现.

如下是一个例子:



因为每个样本都可以转化为有限次基于规则的属性(特征)判断, 构成树上一条根到叶子的有限路径, 所以这棵树是可以完全拟合训练集数据的.

注意, 该树的某层可能是无限宽的, 比如某个维度的属性(特征)是稠密的并且其他维度与这个维度独立不相关.

但是因为属性(特征)个数的有限的, 这就意味着这个存在冗余的(可能泛化能力非常差)决策树是有限深度并且完全拟合训练集的.

(★ 本次作业附录有关于训练集有限数据 \Rightarrow 存在有限深度的与训练集一致的决策树的数学证明.)

(2)

- $Gain(D, \text{喜欢ML作业})$:

$$Ent(D^1) = -\log_2 1 = 0$$

$$Ent(D^2) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3}\right) = 0.9183$$

$$\begin{aligned} Gain(D, \text{喜欢ML作业}) &= 0.9710 - \left(\frac{2}{5} \times 0 + \frac{3}{5} \times 0.9183\right) \\ &= 0.42002 \end{aligned}$$

- $Gain(D, \text{性别})$:

$$\begin{aligned} Gain(D, \text{性别}) &= Ent(D) - \sum_{v=1}^2 \frac{|D^v|}{|D|} Ent(D^v) \\ &= 0.42002 \end{aligned}$$

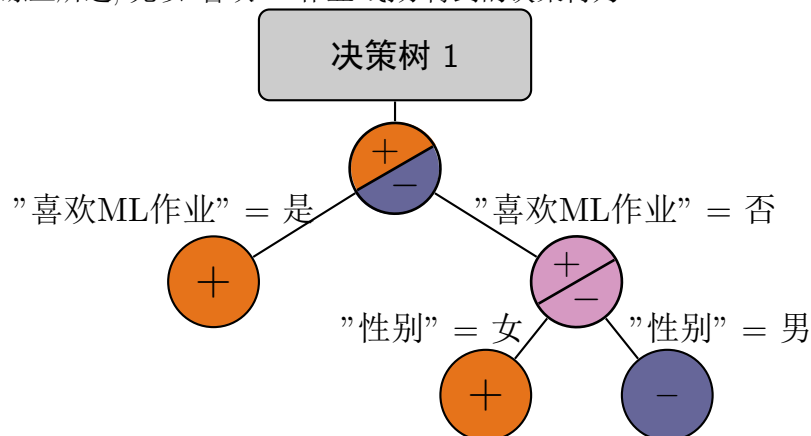
综上所述我们发现两个属性都可以先被划分.

数据集被划分为:

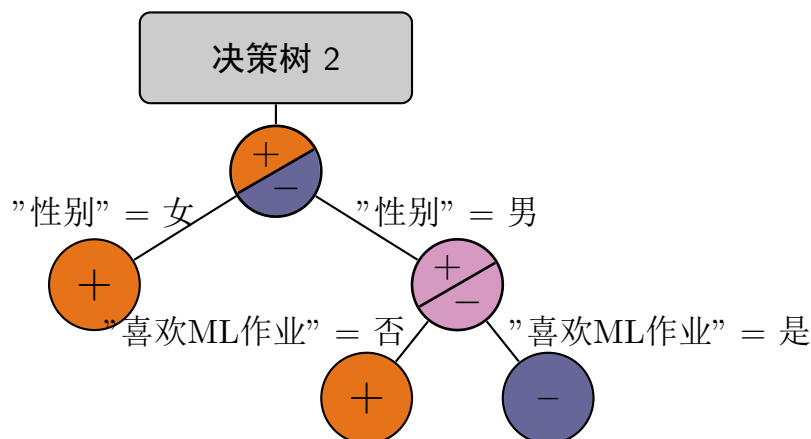
- ID = 1, 2; (此时样本全部属于同一类别, 无需划分)
- ID = 3, 4, 5;

再在“喜欢ML作业=否”分支上计算增益率继续划分, 令 $D' = \{x|x \in D \wedge x_{ID} = 3, 4, 5\}$, $Ent(D') = 0.9183 > 0$, 所以 $Gain(D', \text{性别}) > 0$, 即只能在“性别”上进行划分.

综上所述, 先以“喜欢ML作业”划分得到的决策树为:



同理可得:



(3)

对于决策树 1:

• 预剪枝:

– 对于“喜欢ML作业”结点:

划分前, 因为所有数据中正类更多, 所以这个结点被标记为正, 用验证集对这个结点进行评估: 编号6被分类正确, 所以精度为: $\frac{1}{4} = 25\%$.

划分后, “=是”的结点被标记为正, “=否”的结点因为包含{ 3, 4, 5 }, 所以被标记为负(因为负类更多), 所以在验证集上: 只有编号7的被分类错误, 所以精度为 $\frac{3}{4}$.

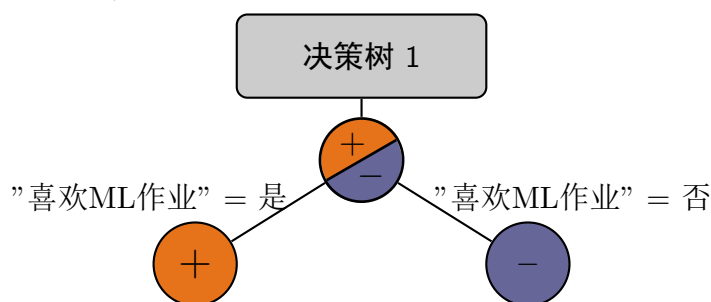
所以 对于“喜欢ML作业”结点 决策是划分.

– 对于“喜欢ML作业=否”结点做同样的操作:

划分前, 这个结点在训练集中负例更多, 所以被标记为负, 验证集精度为 100 %.

所以选择不划分(预剪枝).

综上所述, 预剪枝决策树为:



• 后剪枝: 未剪枝决策树精度为 50 %.

– “喜欢ML作业” = 否 结点:

将其替换为叶结点之后包含的训练样本为: { 3, 4, 5 }, 负类更多, 所以被标记为负.

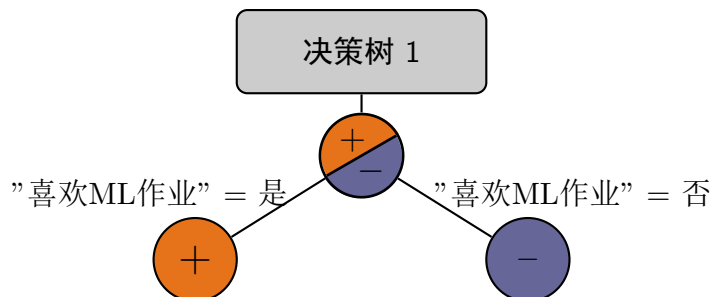
此时在验证集上的精度为 75 %. 所以剪掉这个分支.

— 根节点:

将其替换为叶子后包含所有训练集 \Rightarrow 被标记为正, 在验证集上精度: 25 %.

所以不能剪.

综上所述, 后剪枝决策树为:



同理, 我们对第二种决策树采用同样的操作:

• 预剪枝:

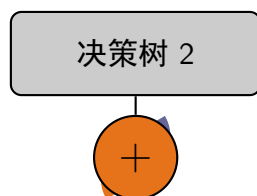
— 对于“性别”结点:

划分前, 训练数据中正类更多, 结点被标记为正, 用验证集对这个结点进行评估: 编号6被分类正确, 所以精度为: $\frac{1}{4} = 25\%$.

划分后, “=男”的结点被标记为负, “=女”的结点标记为正, 所以被标记为负(因为负类更多), 所以在验证集上: 只有编号6的被分类正确, 所以精度为 $\frac{1}{4} = 25\%$.

所以 对于“性别”结点 决策是不划分.

综上所述, 预剪枝决策树为:



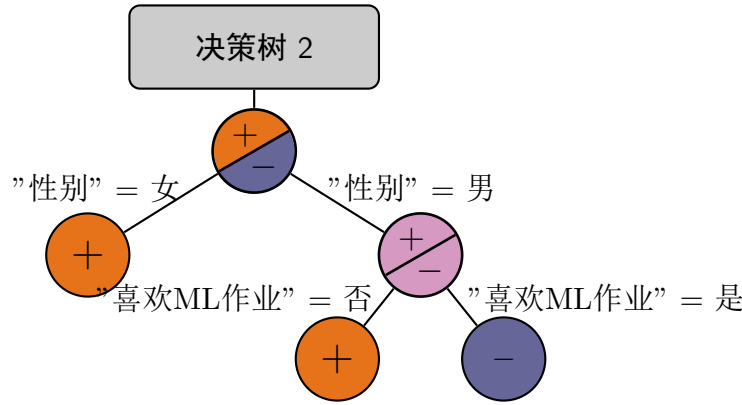
• 后剪枝: 未剪枝决策树精度为 50 %.

— “性别” = 男 结点:

将其替换为叶结点之后包含的训练样本为: { 1, 3, 4 }, 负类更多, 所以被标记为负.

此时在验证集上的精度为 25 %. 所以不剪.

综上所述, 后剪枝决策树为:



(4) 综上所述, 以**决策树 2**为例 (因为决策树2是两种剪枝策略不同的):

预剪枝决策树在训练集上准确率为: 60 %, 验证集上为25 %; 后剪枝决策树在训练集上100 %, 在验证集上为 50 %.

我们可以发现后剪枝决策树保留的分支更多, 拟合更好, 泛化性能往往优于预剪枝决策树.

3 [25pts] SVM with Weighted Penalty

考虑标准的SVM优化问题如下(即课本公式(6.35)),

$$\begin{aligned}
 \min_{\mathbf{w}, b, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\
 \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\
 & \xi_i \geq 0, i = 1, 2, \dots, m.
 \end{aligned} \tag{1}$$

注意到, 在(1)中, 对于正例和负例, 其在目标函数中分类错误或分对但置信度较低的“惩罚”是相同的。在实际场景中, 很多时候正例和负例分错或分对但置信度较低的“惩罚”往往是不同的, 比如癌症诊断等。

现在, 我们希望对负例分类错误(即false positive)或分对但置信度较低的样本施加 $k > 0$ 倍于正例中被分错的或者分对但置信度较低的样本的“惩罚”。对于此类场景下,

(1) [10pts] 请给出相应的SVM优化问题。

(2) [15pts] 请给出相应的对偶问题及KKT条件, 要求详细的推导步骤。

Solution. (1)

$$\min_{\mathbf{w}, b, \xi_i} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i \in D^+} \xi_i + k \sum_{i \in D^-} \xi_i \right).$$

其中 D^+ , D^- 分别是所有正类样本和负类样本

$$\begin{aligned}
 \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\
 & \xi_i \geq 0, i = 1, 2, \dots, m.
 \end{aligned}$$

(2)

由拉格朗日乘子法:

$$L(\mathbf{w}, b, \alpha, \xi, \mu) = \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\sum_{i \in D^+} \xi_i + k \sum_{i \in D^-} \xi_i \right) + \sum_{i=1}^m \alpha_i (1 - \xi_i - y_i (\mathbf{w}^T \mathbf{x}_i + b)) - \sum_{i=1}^m \mu_i \xi_i$$

开始求解, 分别求偏导:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha, \xi, \mu) &= \mathbf{w} - \sum_{j=1}^m \alpha_j y_j \mathbf{x}_j = 0 \\ \Rightarrow \mathbf{w} &= \sum_{j=1}^m \alpha_j y_j \mathbf{x}_j \\ \frac{\partial}{\partial \xi_i} L &= C \cdot (\mathbb{I}(i \in D^+) + k \cdot \mathbb{I}(i \in D^-)) - (\alpha_i + \mu_i) = 0 \\ \frac{\partial}{\partial b} L &= \sum_{i=1}^m \alpha_i y_i = 0 \end{aligned}$$

代入(1)中优化问题式子, 得到对偶问题:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C (\mathbb{I}(i \in D^+) + k \cdot \mathbb{I}(i \in D^-)), \quad i = 1, 2, \dots, m \end{aligned}$$

KKT条件:

$$\begin{cases} \alpha_i \geq 0, \quad \mu_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0 \\ \alpha_i (y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0 \\ \xi_i \geq 0, \mu_i \xi_i = 0 \end{cases}$$

- 如果 $\alpha_i < C (\mathbb{I}(i \in D^+) + k \cdot \mathbb{I}(i \in D^-))$: 则由拉格朗日函数对 ξ_i 的偏导为零(上面式子), 可得 $\mu_i > 0 \Rightarrow \xi_i = 0$. 此时样本恰好在最大间隔边界上.
- 如果 $\alpha_i = C (\mathbb{I}(i \in D^+) + k \cdot \mathbb{I}(i \in D^-)) \Rightarrow \mu_i = 0 \Rightarrow \begin{cases} \text{分类正确, } \xi_i \leq 1. \\ \text{分类错误, } \xi_i > 1. \end{cases}$

4 [30 pts] 编程题, Linear SVM

请结合编程题指南进行理解

SVM转化成的对偶问题实际是一个二次规划问题，除了SMO算法外，传统二次规划方法也可以用于求解对偶问题。求得最优拉格朗日乘子后，超平面参数 \mathbf{w} , \mathbf{b} 可由以下式子得到：

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i x_i \quad (2)$$

$$\mathbf{b} = \frac{1}{|S|} \sum_{s \in S} (y_s - \sum_{i \in S} \alpha_i y_i x_i^T x_s) \quad (3)$$

请完成以下任务：


- (1) [5pts] 使用QP方法求解训练集上的SVM分类对偶问题(不考虑软间隔情况)。
- (2) [10 pts] 手动实现SMO算法求解上述对偶问题。
- (3) [15 pts] 对测试数据进行预测，确保预测结果尽可能准确。

Solution. 采用十折交叉验证.

(3) 亮点:

- 尝试了多种核:
 - Polynomial kernel
 - Gaussian radial basis function
 - Gaussian kernel
 - Laplace RBF kernel
 - Sigmoid kernel
 - 吴建鑫老师提出的: power mean kernel

$$\bullet M_p(x_1, \dots, x_n) = \left(\frac{1}{n} \sum_{i=1}^n x_i^p \right)^{\frac{1}{p}}$$

$x_i > 0$ 

$$M_p(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d M_p(x_i, y_i), \quad p \leq 0 \text{ 为 Mercer 核}$$

- 尝试了可能的特征工程:
 - 标准化
 - 归一化
 - 区间缩放
 - 创造新特征: 对单独特征列的单调变换
- fine tuning 一些超参数, 比如 C , kernel 的参数(比如 rbf 核的 γ)

本题参考资料:

cvxopt document: <http://cvxopt.org/documentation/>
<https://cosmolearning.org/video-lectures/soft-margin-svm-kernels-with-cvxopt/>
https://en.wikipedia.org/wiki/Sequential_minimal_optimization
<http://web.cs.iastate.edu/~honaavar/smo-svm.pdf>
<http://pages.cs.wisc.edu/~dpage/cs760/SM0lecture.pdf>
 但是我可以保证理解思想并独立实现.

附录:

证明 训练集有限数据 \Rightarrow 存在有限深度的与训练集一致的决策树

- 首先说明属性(维度)大小与决策树划分深度的关系:

Lemma: 决策树就是 D 维空间的一个划分, 且深度 \leq 划分个数. 在子空间 C_i 内对应着分类标签 y_i 输出. 其中 D 维空间的每一个维度要么为连续有限区间, 要么为 $\{0, 1\}$ 二值变量.

证明如下:

不妨设数据 $\mathbf{x} \in \mathbb{R}^{D^*}$, 且由一般性可设 $D^* = 2$, $x^{(1)} \in [m, n]$, $x^{(2)} \in \{A, B, C\}$, (即第一个维度对应 $[m, n]$ 连续区间, 第二个维度对应 3 个离散值).

对 $x^{(2)}$ 进行 one-hot 处理, 目的是消除离散值的无序性, 即:

$$\mathbf{x}^{(2)} = \begin{cases} (1, 0, 0), & (\mathbf{x}^{(2)} = A). \\ (0, 1, 0), & (\mathbf{x}^{(2)} = B). \\ (0, 0, 1), & (\mathbf{x}^{(2)} = C). \end{cases}$$

这样 $D = D^* + 2 = 4$ 个维度, 同理对于多个离散值以及连续值特征也依此扩充维度, 使 $\mathbf{x} \in \mathbb{R}^D$ 且每一个维度要么为连续有限区间, 要么为 $\{0, 1\}$ 二值变量.

不妨设

$$\mathbb{S}(\mathbf{x}^{(j)}) = (\mathbf{x} \text{ 在第 } j \text{ 维上的取值个数})$$

举个例子:

$$\mathbf{x}_1 = (0.2, A)$$

$$\mathbf{x}_2 = (0.6, C)$$

$$\mathbf{x}_3 = (0.9, B)$$

$$\mathbf{x}_4 = (0.6, B)$$

$$\mathbf{x}_5 = (0.1, B)$$

$$\text{如上: } \mathbb{S}(\mathbf{x}^{(1)}) = 4, \mathbb{S}(\mathbf{x}^{(2)}) = 3;$$

则:

$$\text{划分后子空间个数} \leq \prod_{j=1}^D \mathbb{S}(\mathbf{x}^{(j)})$$

这是显然的, 因为第 j 维度上的划分最多只有 $\mathbb{S}(\mathbf{x}^{(j)})$.

又有:

决策树深度 \leq 上述划分子空间个数

由决策树原理可知, 决策树的每层如果只有一个分支, 那么上式取等.