SVM 算法

时间： April 29, 2019

目　　录

第 1 章 拉格朗日对偶性

1.1 约束优化问题

设目标函数 $f(x)$ 是定义在 R^n 上的连续可微函数，在变量 x 满足某些约束的条件下求 f(x) 的最优值的问题称为约束优化问题。如：

$$\begin{aligned} & \min_x f(x) \\ s.t. \quad & g_i(x) \le 0, i = 1, 2, ..., m; \\ & h_j(x) = 0, j = 1, 2, ..., l; \end{aligned}$$

表示在满足等式和不等式约束条件下，求得函数 f(x) 的最小值。这里，函数 g(x) 和 h(x) 均是定义在 R^n 空间上的连续可微函数。我们引入系数参数 α, β，定义如下的 Lagrange 函数：

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)$$

定义如下函数：

$$\theta_P(x) = \max_{\alpha, \beta; \alpha_i \ge 0} L(x, \alpha, \beta)$$

$\theta_P(x)$ 具有如下性质：

$$\theta_P(x) = \begin{cases} f(x) & \text{如果 } x \text{ 满足约束条件;} \\ \infty & \text{如果 } x \text{ 不满足约束条件;} \end{cases}$$

那么原始优化问题的等价表示是：

$$\min_x \theta_P(x)$$

这样

$$\min_x \max_{\alpha, \beta; \alpha_i \ge 0} L(x, \alpha, \beta)$$

表示 Lagrange 函数的极小极大问题，

$$p^* = \min_x \theta_P(x)$$

是原始问题的最优值。

1.2 约束优化问题的对偶问题

定义如下函数：

$$\theta_D(\alpha,\beta)=\min_x L(x,\alpha,\beta)$$

于是：

$$\max_{\alpha,\beta;\alpha_i\geq 0}\theta_D(\alpha,\beta)$$

称为原始问题的对偶问题。

这样

$$\max_{\alpha,\beta;\alpha_i\geq 0}\min_x L(x,\alpha,\beta)$$

表示 Lagrange 函数的极大极小问题，

$$d^*=\max_{\alpha,\beta;\alpha_i\geq 0}\theta_D(\alpha,\beta)$$

是对偶问题的最优值。

1.3 一些定理

定理 1.1

若原始问题和对偶问题都有最优值，p^* 是原始问题的最优值，d^* 是对偶问题的最优值，则：

$$d^*\leq p^*$$

♡

定理 1.2

若 x^* 是原始问题的可行解，α^*,β^* 是对偶问题的可行解，且满足

$$p^*=d^*$$

那么 x^*,α^*,β^* 是原始问题和对偶问题的最优解。

♡

定理 1.3

若函数 $f(x)$ 和 $g_i(x)$ 是凸函数，$h_i(x)$ 是仿射函数，并且不等式约束 $g_i(x)\leq 0$ 严格成立，则存在 x^*,α^*,β^*，使得 x^* 是原始问题的最优解，α^*,β^* 是对偶问

题的最优解，且：

$$p^* = d^* = L(x^*, \alpha^*, \beta^*)$$ ♡

定理 1.4

若函数 $f(x)$ 和 $g_i(x)$ 是凸函数，$h_i(x)$ 是仿射函数，并且不等式约束 $g_i(x) \leq 0$ 严格成立，x^*, α^*, β^* 分别是原始问题和对偶问题的最优解的充要条件是：x^*，α^*, β^* 需满足如下的 KKT(Karush-Kuhn-Tucker) 条件(定常方程式，原始可行性，对偶可行性，互补松弛性)。

$$\nabla_x L(x^*, \alpha^*, \beta^*) = 0$$

$$g_i(x^*) \leq 0,\ i = 1, 2, ..., m$$

$$h_j(x^*) = 0,\ j = 1, 2, ..., l$$

$$\alpha_i^* \geq 0,\ i = 1, 2, ..., m$$

$$\alpha_i^* g_i(x^*) = 0,\ i = 1, 2, ..., m$$ ♡

第 2 章 支持向量机的基本原理

2.1 线性可分

设给定一特征空间上的训练集：

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

x_i 是第 i 个样本的特征向量，y_i 是对应的类标记，当 y_i=+1 时，样本属于正类，当 y_i=-1 时，样本属于负类。数据集线性可分的含义是，存在一个超平面 $w^T x + b = 0$ 的，将特征空间划分成两个部分，一部分全是正类，一部分全是负类。

2.2 函数间隔

当确定一个超平面 $w^T x + b = 0$ 时，点 (x_i, y_i) 关于超平面的函数间隔定义如下：

$$\hat{\gamma}_i = y_i(w^T x_i + b)$$

数据集 D 关于超平面的函数间隔定义为数据集中和超平面距离最小的点的距离：

$$\hat{\gamma} = \min_i \hat{\gamma}_i$$

2.3 几何间隔

同时成比例的增加 w, b 时，超平面不变，但是函数间隔却也成比例增加了，需对超平面的法向量 w 施加规范约束，如约束法向量的长度为 1，$|w| = 1$，此时的函数间隔成为几何间隔。

一般的，定义样本点 (x_i, y_i) 到超平面 $w^T x + b = 0$ 的几何间隔如下：

$$\gamma_i = \frac{w^T}{|w|} x_i + \frac{b}{|w|}$$

于是数据集 D 关于超平面的几何间隔定义：

$$\gamma = \min_i \gamma_i$$

2.4 间隔最大化

支持向量机的基本思想是能够正确划分训练数据集并且几何间隔最大的超平面，这个问题可以表述为如下的约束最优化问题：

$$\max_{w,b} \gamma$$
$$s.t.\ y_i(\frac{w}{|w|}x_i + \frac{b}{|w|}) \geq \gamma, \quad i = 1, 2, ..., N$$

等价于下面用函数间隔表述的问题：

$$\max_{w,b} \frac{\hat{\gamma}}{|w|}$$
$$s.t.\ y_i(wx_i + b) \geq \hat{\gamma}, \quad i = 1, 2, ..., N$$

由之前的讨论可知，函数间隔 $\hat{\gamma}$ 的具体值不影响优化，我们可以取 $\hat{\gamma} = 1$，这样优化的目标函数是 $\frac{1}{|w|}$，这又等价于最小化函数 $\frac{1}{2}|w|^2$，于是线性可分支持向量机的优化问题可以表述为如下形式：

$$\min_{w,b} \frac{1}{2}|w|^2$$
$$s.t.\ y_i(wx_i + b) - 1 \geq 0, \quad i = 1, 2, ..., N$$

2.5 间隔最大化的对偶问题

引入 Lagrange 乘子 $\alpha_i \geq 0, i = 1, 2, ..., N$，定义如下的拉格朗日函数：

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

原始问题的对偶问题级极大极小问题：

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

首先，我们求解 $\min_{w,b} L(w, b, \alpha)$，即求如下方程组：

$$\begin{cases} \nabla_w L(w, b, \alpha) & = 0 \\ \nabla_b L(w, b, \alpha) & = 0 \end{cases}$$

解得：

$$\begin{cases} w = \sum_{i=1}^{N} \alpha_i y_i x_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

将上述结果带入 Lagrange 函数得：

$$\min_{w,b} L(w, b, \alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i x_j) + \sum_{i=1}^{N} \alpha_i$$

于是，原始问题的对偶问题可以表述为如下形式：

$$\begin{aligned} &\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i x_j) - \sum_{i=1}^{N} \alpha_i \\ &s.t. \sum_{i=1}^{N} \alpha_i y_i = 0 \\ &\alpha_i \geq 0, \quad i = 1, 2, ..., N \end{aligned}$$

2.6 核函数

设 X 是输入空间，H 是特征空间，如果存在一个从 X 到 H 的映射：

$$\phi(x) : X \rightarrow H$$

对所有的 $x_i, x_j \in X$，函数 $K(x_1, x_2)$ 满足：

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

则：$K(x_i, x_j)$ 是核函数，$\phi(x)$ 是映射函数。

通过引入核函数，我们可以处理非线性边界问题。其基本思想是，当数据集在输入空间中非线性可分，我们认为存在一个映射函数，将输入空间的样本点映射到一个更高维的特征空间。在这个高维的特征空间里，会存在一个超平面将映射后的样本集分割成正类和负类两部分。算法的目的是在这个更高为的特征空间做线性分割。核函数可以让我们不需知道映射函数的具体形式，就能计算在特征空

间两个样本的内积。我们重新回顾下线性可分支持向量机的对偶表述：

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

我们只需简单的将 $x_i, x+j$ 的内积 $x_i x_j$ 替换成核函数，这样我们便得到一个非线性支持向量机的对偶表示：

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

那么对于不同的分类问题，我们需要不同的核函数完成分类，常用的核函数有如下几种：

- 线性核函数

$$K(x_i, x_j) = x_i x_j$$

- 多项式核函数

$$K(x_i, x_j) = (x_i x + j + 1)^p$$

- 高斯核函数

$$K(x_i, x_j) = \exp^{-\frac{|x_i - x_j|^2}{2\sigma^2}}$$

2.7 支持向量机的学习算法——序列最小最优化算法(SMO)

待求变量 α_i 的数目等于样本数，一般的数据集规模使得待求变量数目多，难以得到解析解，计算机直接求解对小样本数据有效，当样本数据规模较大时，算法

会非常低效。高效的支持向量机的学习算法尤为重要,其中一种重要的算法 SMO 由 Platt 于 1998 年提出。SMO 算法的基本思想是,算法每次选择两个变量 α_1, α_2,固定其它变量,构建一个子二次规划问题,并求得这个子二次规划问题的最优解。然后在另选择两个变量,继续求解子问题。当所有变量的解都满足 KKT 条件后,那么问题的最优解便得到了。子二次规划问题中,由于约束条件的存在,实际上只有一个变量是自由的,一个变量可由另一个变量求得,实际上是一个单变量优化问题,这个问题可以解析求得解。这样,SMO 算法可以分成两个部分。

1. 对两个变量的二次规划问题的求解。
2. 选择两个变量的启发式算法。

第 3 章 在贫困生预测中的应用

SVM 算法的理论基础是凸二次规划问题的对偶问题的求解，通过引入核函数，实现高维空间非线性映射向量的内积计算，从而完成对非线性问题的求解。算法的目标是找到特征空间上的最优超平面，结果是找到支持向量（距离超平面最近的点），在分类决策中，起重要作用的是这些支持向量。

具有以下优点：

- 理论完善，有一套几乎完美的理论可以解释其原理。
- 少数支持向量决定了最终结果，具有较好的“鲁棒”性。
- 泛化能力强

缺点：

- 大样本数据量会消耗大量的内存和运算时间。主要改进有有 J.Platt 的 SMO 算法、T.Joachims 的 SVM、C.J.C.Burges 等的 PCGC、张学工的 CSVM 以及 O.L.Mangasarian 等的 SOR 算法
- 经典的支持向量机算法只给出了二类分类的算法，多分类问题需通过多个二类支持向量机的组合来解决

在贫困生预测中，我们的输入空间是学生的各项消费指标，输出是学生是否贫困的类标记，这里属于一个二分类问题，可以非常方便的应用 SVM 算法。另一方面，我们不仅想知道学生是否贫困，还需要知道学生是贫困生的概率，通过概率来计算贫困指数。在 SVM 算法的描述中，我们并未涉及到概率，模型是无法直接输出概率的，解决办法有两个。

1. 利用决策函数

$$f(x) = sgn(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*)$$

符号函数里面的部分是样本点 x 到决策边界的函数距离，这个距离可以用来衡量决策的确信度，作为概率的一种反映用于贫困指数的计算中。

2. Platt 标度

这是一种是一种参数化方法，使用 LR 模型（sigmoid 函数）对模型的输出值进行拟合，将模型的原始输出值映射为概率值，区间（0，1）。假设 f(x) 为模型的输出值，那么：

$$P(y = 1|f) = \frac{1}{1 + exp(A * f + B)}$$

A，B 通过训练集使用极大似然法解得。