支持向量机整理博客发布

王倩妮 2015112956 交通 2015-02 班

作为一个交通专业的学生加入“机器学习”课程已经有 12 周了，现在课程已经进展到 SVM 的部分。今天我们一起来整理 SVM（支持向量机）的相关内容。第一次接触 SVM 是大二下在熊老师的交通数据采集与处理课上，熊熊精简的 jupyter notebook 讲义让人“摸不着头脑”，不过图形化的可视化让我们对于 SVM 的过程有了“几何”的理解（说白了就是能通过看图看懂是怎么分类的，大体了解原理）；大三上的数据挖掘大作业，大家与 SVM 第二次相遇，不过同学们没有把重点放在原理的理解上，而是调用如 Python 中的 scikit-learn，MATLAB 中的 libsvm 直接解决问题。因此当我大三下学期再次遇到 SVM 时，更希望关注其数学原理层面的内容，搞清楚这套原理是怎么来的。

在学习的过程中发现，很多证明过程涉及运筹学的优化，那些理论在诸如网络交通流理论的搭建与求解中也有所应用，因此还需要抽空补起这部分知识。由于这部分知识的缺乏所造成的理解不到位的问题也希望大家指出并见谅。

一、SVM 基本内容

SVM 方法的基本思想

定义最优线性超平面，并把寻找最优线性超平面的算法归结为求解一个最优化（凸规划）问题。进而基于 Mercer 核展开定理，通过非线性映射，把样本空间映射到一个高维乃至于无穷维的特征空间（Hilbert 空间），使得在特征空间中可以应用线性学习机的方法解决样本空间中的高度非线性分类和回归等问题。简单地说就是实现升维和线性化。

SVM 的简单描述

听到以上一段内容可能会感觉“云里雾里”，但打个简单的比方，现在我们有一些数据（包含两个 attribute，和一个分类对应的标签），那么我们把这两个 attribute 的数据分别作 x、y 并绘制于一个二维平面上，用不同的颜色或形状区分不同的标签。

即**输入：**两组带标签的数据。

SVM 方法就是要通过一系列手段找到一条直线将两类数据分隔开，但就下图来说，能够分隔的直线有无数条，我们需要通过定义输出一条“最优”的直线。

即**输出：**一条最优的将两组数据分开的直线。

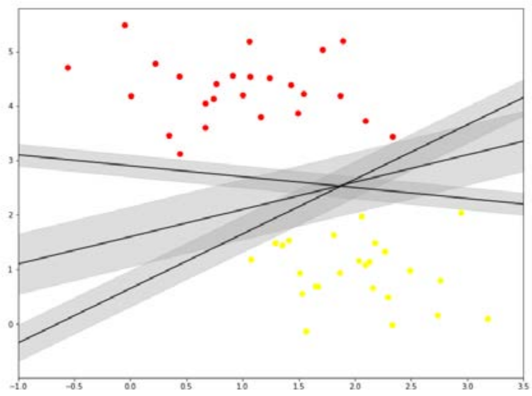


图 1

超平面的概念

这里用于分隔的直线拓展到不同维度后可称之为*超平面（hyperplane）*，在 N 维空间内，超平面维度为 N-1，超平面也就是分类问题的决策边界，使分布在超平面一侧的所有数据属于某个类别，分布在另一侧的数据属于另一类别。

间隔的概念；如何寻找到最优的分隔？

刚刚我们提到了超平面这个决策边界，那么如何找到所谓“最优”呢？这里的最优的直观解释为：找到离分隔超平面最近的点，确保它们离分隔面的距离尽量远。这里点到分隔面的距离就被称为*间隔（margin）*，在 SVM 之中我们需要**以最大化间隔为目标**，但该问题求解困难，还需一系列推证进行转化与求解。

支持向量的概念

SVM 是支持向量机的简称，那么什么是支持向量呢？*支持向量（support vector）*就是离分隔超平面最近的那些点；也就是离超平面最近的那些点支撑着这个分类器，确保分类器的分类规则实现。

二、SVM 支持向量机优缺点分析

SVM 是最常用的分类方法之一：

优点：其结果易于解释，有着坚实的理论基础，适用于小样本学习；少数支持向量决定于最优结果，达到抓住关键样本、剔除大量冗余样本，提高“鲁棒性”。

缺点：对于参数调节和核函数的选择较为敏感，原始分类器不加修改仅适用于处理二类问题。

SVM 适用于数值型数据与标称型数据。

三、SVM 数学原理推导

由于公式编辑较为困难，因此以下推导过程通过手写进行，并将推导过程附图进行分享。以下为线性 SVM 的推导：

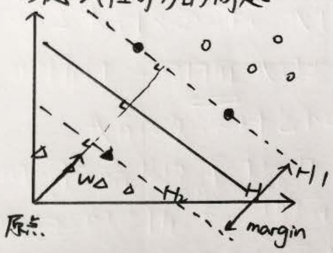
训练数据 $\{\vec{x_i}, y_i\}$，$i=1,2,\cdots,m$　$y_i \in \{-1,1\}$，$\vec{x_i} \in R^d$.

超平面方程　$\vec{w}\cdot\vec{x}+b=0$

上式中 $\vec{w}$ 为超平面的法向量，$\frac{|b|}{\|w\|}$ 为原点到该超平面的垂直距离.

令 $d_+(d_-)$ 表示正(负)样本(支持向量)到超平面的最近距离，则间隔 $margin = d_+ + d_-$

考虑线性可分的问题



我们需要寻找一个分割超平面，使 margin 最大.

依左图，有：

$$\begin{cases} H_1: \vec{w}\cdot\vec{x_i}+b \le -1 & \text{for } y_i=-1 \quad ① \\ H_2: \vec{w}\cdot\vec{x_i}+b \ge 1 & \text{for } y_i=1 \quad ② \end{cases}$$

进行了规范化处理

将①②式合并.

$y_i\cdot(w x_i+b) \ge 1 \cdots \forall i$　③

如上图，margin 为 H_1 与 H_2 间的距离，原点到 H_1 与 H_2 的距离可表示为 $\frac{|1-b|}{\|w\|}$ 与 $\frac{|-1-b|}{\|w\|}$，所以 $margin = \frac{|-1-b|}{\|w\|} - \frac{|1-b|}{\|w\|} = \frac{2}{\|w\|}$

因此目标即为最大化 $\frac{2}{\|w\|}$，现将目标进行修改，改为最小化 $\|w\|$，为了接下来的计算可将以上内容等同于最小化 $\|w\|^2$ 或 $\frac{1}{2}\|w\|^2$

所以我们将优化问题改为：

$$\begin{cases} \min\{\frac{1}{2}\|w\|^2\} \\ s.t.\ y_i(w x_i+b)-1 \ge 0 \quad \forall i. \end{cases}$$

此问题不能直接求解，因此利用拉格朗日乘子法进行改换

图 2

$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i(y_i(w x_i+b)-1)$　④　其中 α 为拉格朗日乘子，$\alpha \ge 0$

对 w 与 b 分别做微分(求偏导)

$$\frac{\partial L}{\partial w}=0 \Rightarrow \vec{w}=\sum_{i=1}^{m}\alpha_i y_i \vec{x_i} \quad ⑤$$

$$\frac{\partial L}{\partial b}=0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i=0 \quad ⑥$$

将⑤⑥结果代入④

$$J(\alpha)=\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y_i\vec{x_i}\cdot\sum_{i=1}^{m}\alpha_i y_i\vec{x_i}\right)+\sum_{i=1}^{m}\alpha_i-\left(\sum_{i=1}^{m}\alpha_i y_i\vec{x_i}\cdot\sum_{j=1}^{m}\alpha_j y_j x_j\right)-b\sum_{i=1}^{m}\alpha_i y_i.$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(\vec{x_i}\cdot\vec{x_j})$$

所以以上问题至此转化为

$$\begin{cases} \max J(\alpha)=\max\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\alpha_i\alpha_j y_i y_j(\vec{x_i}\cdot\vec{x_j})\}. \\ s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0 \quad (\alpha_i \ge 0\ \ i=1,2,\cdots m) \end{cases} \quad ⑦$$

通过比较，我们发现
- 原问题最小化已经改变为了对偶问题的最大化.
- 原问题所求 w 分量个数为 n，与样本维数一样，对偶问题所求的 α 分量个数为 m，与样本个数相同而与维数无关，避免了高维数计算

现有多种方法可以用于解决⑦，⑦求解通过数值方法求解后，得最优解：

$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)$　⑧　其中，非支持向量对应的 α_j^* 均为0

将⑧代入⑤，求得 $w^*=\sum_{i=1}^{m}\alpha_i^* y_i\vec{x_i}$，又将 w^* 代入下式

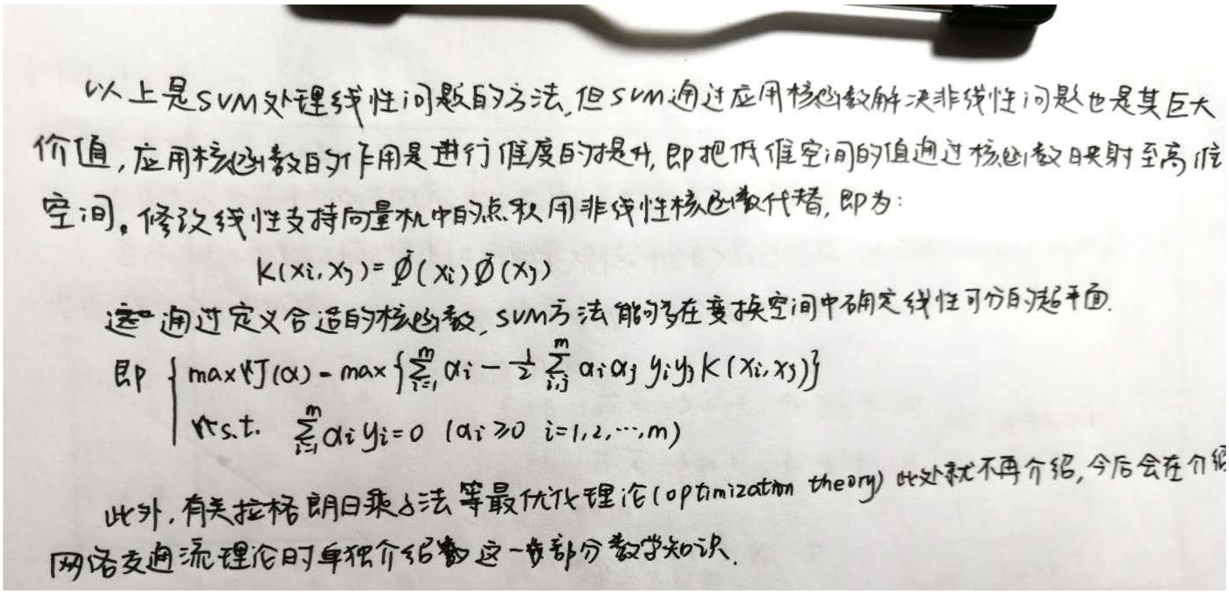
$\alpha_i(y_i(w^* x_i)+b-1)=0$　$i=1,2,\cdots,m$ 可以求得 b^*.

$b^*=-\frac{1}{2}[w^*(X_+ + X_-)]$ 其中 X_+ 与 X_- 分别是正类与负类中任意一个支持向量，再将 w^* 与 b^* 代回超平面.

得最优超平面　$M(\vec{x})=sgn(w^* x+b)=sgn(\sum\alpha_i^* y_i(x_i\cdot x)+b^*)$.

图 3

采用核函数的思想将线性 SVM 分类器进行扩展：


以上是SVM处理线性问题的方法，但SVM通过应用核函数解决非线性问题也是其巨大价值，应用核函数的作用是进行维度的提升，即把低维空间的值通过核函数映射到高维空间，修改线性支持向量机中的点积用非线性核函数代替，即为：

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

通过定义合适的核函数，SVM方法能够在变换空间中确定线性可分的超平面.

即
$$\begin{cases} \max W(\alpha)=\max\left\{\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\right\} \\ s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0 \quad (\alpha_i\geq 0 \quad i=1,2,\cdots,m) \end{cases}$$

此外，有关拉格朗日乘子法等最优化理论(optimization theory)此处就不再介绍，今后会在介绍网络交通流理论时单独介绍这一部分数学知识.

图 4

关于核函数的作用，我们也可以通过下图更为直观的理解，如左图的情况，无法利用直线直接将两类点集进行分隔，因此使用径向基核函数 Radial Basis Function 进行运算，将二维空间扩展至三维空间，至此可以直观看到，可用一平面将两类点进行分隔。

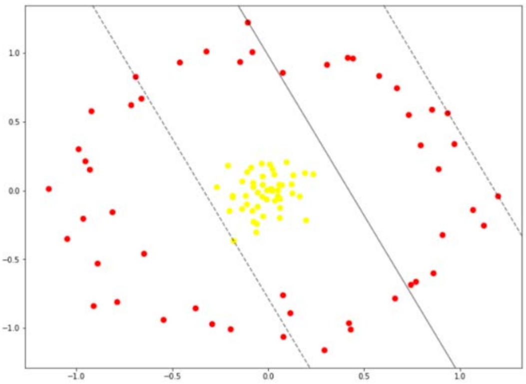


图 5

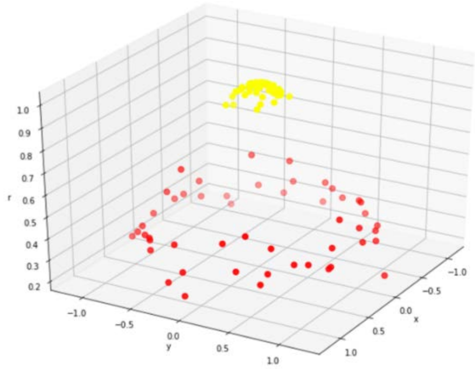


图 6