统计学习方法第二章： 感知机

JazzCharles

2019 年 4 月 1 日

1. 基本概念

统计学习方法的第二章是关于感知机（Perceptron）——一个线性的二分类模型。感知机将输入的样本集 $X \in R^d$ 输出对应的类别 +1 或-1。主要思想为利用一个超平面将样本划分成为正负两个集合。在数据集线性可分的前提下，感知机可以被证明在有限轮次的迭代中可以收敛。以下从模型，策略以及算法三方面简述感知机。

2. 模型

感知机的模型在空间中可以被表示为一个超平面 $w \cdot x + b = 0$，其中 w 为平面的法向量，b 为截距。用公式可以表达为

$$f(x) = sign(w \cdot x + b). \tag{1}$$

其中 $sign()$ 为符号函数，括号内的值大于 0 输出 1，否则输出-1。w, b 为学习算法需要学习的参数。

3. 学习策略

在数据是线性可分的条件下，感知机的目标为找到这样一个超平面分割正负样本。自然的想法是希望超平面能尽可能少的有误分类样本。考虑到误分类样本的个数总和并非参数的连续可导函数。对于误分类点自然希望能到超平面的另外一侧即正确分类一侧，那么误分类点和超平面的距离越近证明超平面的效果越好。因此采用最小化所有误分类点到超平面距离作为优化策略。

$$L(w, b) = -\sum_{x_i \in M_i} y_i * (w \cdot x_i + b) \tag{2}$$

其中 M_i 代表所有误分类点的集合，公式等价于最小化经验风险。

4. 学习算法

求解方法可以采用随机梯度下降（SGD)。损失函数对参数的导数分别为：

$$\partial L / \partial w = -\sum_{x_i \in M_i} y_i * x_i \tag{3}$$

$$\partial L / \partial b = -\sum_{x_i \in M_i} y_i \tag{4}$$

不同于普通随机梯度下降算法在于每次迭代仅仅随机选取一个误分类点的梯度来更新权重，而非采用多个误分类点或是全部误分类点！设学习率为 α，因此更新过程可以表示为：

$$w = w + \alpha * y_i * x_i \tag{5}$$

$$b = b + \alpha * y_i \tag{6}$$

文中提到为了更高效地计算梯度更新，考虑利用感知机的对偶形式。考虑到参数 w 每次都利用单个误分类点的信息进行更新，可以认为最后总共更新的总量等同于所有误分类点的线性组合。因此可以将 w 表示为（(b 同理)：

$$w = \sum_{x_i \in M_i} a_i * y_i * x_i \tag{7}$$

$$a_i = n_i * \alpha \tag{8}$$

其中 a_i 可以表示为线性组合的系数，n_i 代表误分类点被选中的次数，可以看出选中的次数越多，组合系数的权重越大。因此对偶形式的损失函数可以表达成。

$$L(w,b) = -\sum_{x_i \in M_i} y_i * (\sum_{x_j \in M_j} a_j * y_j * (x_j \cdot x_i) + b) \tag{9}$$

这样的表达方式使得 $x_i \cdot x_j$ 可以被提前预处理成 Gram 矩阵，提高计算效率。更新参数 w 的过程变换为每次更新 a_i 的过程，累计当前样本点被误分类的次数 $n_i = n_i + 1$，即 $a_i = a_i + \alpha$

5. 收敛性

在数据集为线性可分的条件下，更新步数 k 可以在有限步数内收敛。其满足以下条件：

$$k <= (R/\gamma)^2 \tag{10}$$

$$R = \max_{1<=i<=N} ||x_i|| \tag{11}$$

$$\gamma = \min_{1<=i<=N} y_i * (w \cdot x_i + b) \tag{12}$$

其中 R 代表所有样本中模最大的值，γ 近似代表最近的样本点到分割超平面的距离。