Online Learning[1]

简介

本章将探讨在线学习（online learning），之前探讨的模型大多是批量学习（batch learning），就是在给定的数据集上一次性训练出一个假设函数。而在线学习要根据新来的数据，边学习，边给出结果。

给定一个按序到达的数据集$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$，在线学习的工作流程是：首先根据点$x^{(1)}$预测$y^{(1)}$，并获得$y^{(1)}$的真实值修正模型。接着根据$x^{(2)}$预测$y^{(2)}$，并获得$y^{(2)}$的真实值继续修正模型……重复这一过程直至学习完所有数据。在在线学习中，我们关注的是在整个学习过程中算法预测错误的总数。

1 感知机（perception）

假设输入空间（特征空间）$\mathcal{X} \in \mathbb{R}^n$，输出空间$\mathcal{Y} \in \{-1, 1\}$，感知机的假设为

$$h_\theta(x) = g(\theta^T x)$$

其中

$$g(z) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

感知机是一种线性分类模型，属于判别模型。它的思想和SVM很相似，都是在寻找一个能够将所有正例和负例分开的超平面。事实上，感知机的损失函数

$$L(\theta) = -\sum_{x^{(i)} \in M} y^{(i)} \left(\theta^T x^{(i)}\right)$$

其中，M表示所有误分类点的集合。对于误分类点而言，$y^{(i)} \left(\theta^T x^{(i)}\right) < 0$，事实上这个就是在SVM中提到的函数间隔，所以感知机的策略是使在训练集上误分类点的函数间隔总和最小。由于

$$\nabla_{\theta_j}(-y(\theta^T x)) = -yx_j$$

故参数θ的更新法则（SGD，且令学习率为1）为

$$\theta := \theta + yx$$

注意，上式中$x \in M$。也就是说，感知机的每一步迭代只对误分类点修正模型，这正是online learning的思想。我们得到的更新法则虽然简单，但却很有效，下一节将证明这一点。

2 算法的收敛性

定理2.1（Novikoff） 给定训练集$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)\right\}$线性可分[2]时，假设$\|x^{(i)}\| \leq D; i = 1, \cdots, m$，且存在一个单位向量$\mu$使得$y^{(i)} \left(\mu^T x^{(i)}\right) \geq \gamma$对训练集中

所有数据成立，那么感知机误分类次数不超过$(D/\gamma)^2$。

证明 感知机只对误分类点更新权值，令$\theta^{(k)}$表示模型第k次误分类时的权值，那么$\theta^{(1)}=\vec{0}$（θ初始化为0向量）。假设$\left(x^{(i)},y^{(i)}\right)$是第k个误分类的样本，那么$g\left(\left(x^{(i)}\right)^T\theta^{(k)}\right)\neq y^{(i)}$，即$\left(x^{(i)}\right)^T\theta^{(k)}y^{(i)}\leq 0$，此时可得更新法则为

$$\theta^{(k+1)}=\theta^{(k)}+y^{(i)}x^{(i)}$$

推得

$$\begin{aligned}\left(\theta^{(k+1)}\right)^T\mu&=(\theta^{(k)})^T\mu+y^{(i)}x^{(i)}\mu\\&\geq(\theta^{(k)})^T\mu+\gamma\end{aligned}$$

归纳得

$$\left(\theta^{(k+1)}\right)^T\mu\geq\left(\theta^{(1)}\right)^T\mu+k\gamma=k\gamma$$

此外

$$\begin{aligned}\|\theta^{(k+1)}\|^2&=\|\theta^{(k)}+y^{(i)}x^{(i)}\|^2\\&=\|\theta^{(k)}\|^2+\|y^{(i)}x^{(i)}\|^2+2y^{(i)}\left(x^{(i)}\right)^T\theta^{(k)}\\&\leq\|\theta^{(k)}\|^2+\|x^{(i)}\|^2\\&\leq\|\theta^{(k)}\|^2+D^2\end{aligned}$$

归纳得

$$\|\theta^{(k+1)}\|^2\leq\|\theta^{(1)}\|^2+kD^2=kD^2$$

综合以上两个结论

$$\begin{aligned}\sqrt{k}D&\geq\|\theta^{(k+1)}\|\\&\geq\left(\theta^{(k+1)}\right)^T\mu\\&\geq k\gamma\\\Rightarrow\quad k&\leq\left(\frac{D}{\gamma}\right)^2\end{aligned}$$

命题得证。（注：$z^T\mu=\|z\|\cdot\|\mu\|\cos\phi\leq\|z\|$，其中$\phi$表示向量$z$与$\mu$之间的夹角。）定理表明，误分类次数k是有上界的，经过有限次探索[3]可以找到将训练数据完全分开的超平面。也就是说，当训练数据线性可分时，感知机的迭代是收敛的。

---

1. Writen by Jimmy on 2016/03/09.↩
2. 线性可分是指可以找到一个超平面将数据集的正例和负例完全正确地划分到超平面的两侧。↩

3. 有限次探索指的是感知机误分类次数有上限，而且此上限不依赖于数据集大小或者输入的维度，它由输入大小的最大值D及数据几何间隔的最小值γ决定。↩