在线学习(Online Learning)

JerryLead

csxulijie@gmail.com

原题目叫做 The perception and large margin classifiers,其实探讨的是在线学习。这里将题目换了换。以前讨论的都是批量学习(batch learning),就是给了一堆样例后,在样例上学习出假设函数 h。而在线学习就是要根据新来的样例,边学习,边给出结果。

假设样例按照到来的先后顺序依次定义为($x^{(1)},y^{(1)}$),($x^{(2)},y^{(2)}$),...,($x^{(m)},y^{(m)}$)。X 为样本特征,y 为类别标签。我们的任务是到来一个样例 x,给出其类别结果 y 的预测值,之后我们会看到 y 的真实值,然后根据真实值来重新调整模型参数,整个过程是重复迭代的过程,直到所有的样例完成。这么看来,我们也可以将原来用于批量学习的样例拿来作为在线学习的样例。在在线学习中我们主要关注在整个预测过程中预测错误的样例数。

拿二值分类来讲,我们用 y=1 表示正例,y=-1 表示负例。回想在讨论支持向量机中提到的感知算法(perception algorithm)。我们的假设函数为

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

其中 x 是 n 维特征向量, θ 是 n+1 维参数权重。函数 g 用来将 $\theta^T x$ 计算结果映射到-1 和 1 上。具体公式如下:

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

这个也是 logistic 回归中 g 的简化形式。

现在我们提出一个在线学习算法如下:

新来一个样例(x,y),我们先用从之前样例学习到的 $h_{\theta}(x)$ 来得到样例的预测值 y,如果 $h_{\theta}(x) = y$ (即预测正确),那么不改变 θ ,反之

$$\theta \coloneqq \theta + yx$$

也就是说,如果对于预测错误的样例, θ 进行调整时只需加上(实际上为正例)或者减去(实际负例)样本特征 x 值即可。 θ 初始值为向量 0。这里我们关心的是 $\theta^T x$ 的符号,而不是它的具体值。调整方法非常简单。然而这个简单的调整方法还是很有效的,它的错误率不仅是有上界的,而且这个上界不依赖于样例数和特征维度。

下面定理阐述了错误率上界:

定理(Block and Novikoff):

给定按照顺序到来的 $(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),...,(x^{(m)},y^{(m)})$ 样例。假设对于所有的样例 $\|x^{(i)}\| \leq D$,也就是说特征向量长度有界为 D。更进一步,假设存在一个单位长度向量 $\mathbf{u} (\|u\| = 1) \mathbf{1} \mathbf{y}^{(i)} \cdot (u^T x^{(i)}) \geq \gamma$ 。也就是说对于 y=1 的正例, $(u^T x^{(i)}) \geq \gamma$,反例 $(u^T x^{(i)}) \leq -\gamma$, \mathbf{u} 能够有 γ 的间隔将正例和反例分开。那么感知算法的预测的错误样例数不超过 $\left(\frac{\mathbf{D}}{\mathbf{D}}\right)^2$ 。

根据前面对 SVM 的理解,这个定理就可以阐述为:如果训练样本线性可分,并且几何

间距至少是 γ ,样例样本特征向量最长为 D,那么感知算法错误数不会超过 $\left(\frac{D}{\gamma}\right)^2$ 。这个定理是 62 年提出的,63 年 Vapnik 提出 SVM,可见提出也不是偶然的,感知算法也许是当时的热门。

下面主要讨论这个定理的证明:

感知算法只在样例预测错误时进行更新,定义 $\theta^{(k)}$ 是第 k 次预测错误时使用的样本特征权重, $\theta^{(1)} = \vec{0}$ 初始化为 0 向量。假设第 k 次预测错误发生在样例 $(x^{(i)}, y^{(i)})$ 上,利用 $\theta^{(k)}$ 计算 $y^{(i)}$ 值时得到的结果不正确(也就是说 $g((x^{(i)})^T\theta^k) \neq y^{(i)}$,调换 x 和 θ 顺序主要是为了书写方便)。也就是说下面的公式成立:

$$(x^{(i)})^T \theta^{(k)} y^{(i)} \le 0.$$

根据感知算法的更新方法,我们有 $\theta^{(k+1)} = \theta^{(k)} + y^{(i)}x^{(i)}$ 。这时候,两边都乘以 u 得到

$$\begin{array}{rcl} (\theta^{(k+1)})^T u & = & (\theta^{(k)})^T u + y^{(i)} (x^{(i)})^T u \\ & \geq & (\theta^{(k)})^T u + \gamma \end{array}$$

两个向量做内积的时候,放在左边还是右边无所谓,转置符号标注正确即可。 这个式子是个递推公式,就像等差数列一样 f(n+1)=f(n)+d。由此我们可得

$$(\theta^{(k+1)})^T u \ge k\gamma.$$

因为初始 θ 为 0。

下面我们利用前面推导出的 $(x^{(i)})^T \theta^k y^{(i)}) \le 0$ 和 $||x^{(i)}|| \le D$ 得到

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

也就是说 $\theta^{(k+1)}$ 的长度平方不会超过 $\theta^{(k)}$ 与 D 的平方和。 又是一个等差不等式,得到:

$$||\theta^{(k+1)}||^2 \le kD^2$$
.

两边开根号得:

$$\sqrt{k}D \geq ||\theta^{(k+1)}||
\geq (\theta^{(k+1)})^T u
\geq k\gamma.$$

其中第二步可能有点迷惑,我们细想 u 是单位向量的话,

$$z^T u = ||z|| \cdot ||u|| \cos \phi \le ||z|| \cdot ||u||$$

因此上面的不等式成立,最后得到:

$$k \le (D/\gamma)^2$$
.

也就是预测错误的数目不会超过样本特征向量 x 的最长长度与几何间隔的平方。实际上整个调整过程中 θ 就是 x 的线性组合。

整个感知算法应该是在线学习中最简单的一种了。