CS229 Lecture notes

原作者：Andrew Ng (吴恩达)，翻译：CylcesUser

1 感知器（perception）和大型边界分类器 (large margin classifiers)

本章是讲义关于学习理论的最后一部分，我们介绍另外机器学习的模式。在之前的内容中，我们考虑的都是批量学习的情况，即给了我们训练样本集合用于学习，然后学习得到的假设 h 来评估和判别测试数据。在本章，我们要讲义中新的机器学习模式：在线学习，这种情况下，我们的学习算法要在进行学习的同时给出预测。学习算法获得一个样本序列，其中内容为有次序的学习样本，$(x^1, y^1), (x^2, y^2), \cdots, (x^m, y^m)$. 最开始获得的就是 x^1, 然后需要预测 y^1. 在完成预测之后，在把 y^1 的真实值告诉算法（然后利用这个信息来进行某种学习）。接下来给算法提供 x^2, 在让算法对 y^2 进行预测，然后再把 y^2 的真实值告诉算法，这样的算法就又能学到一些信息了。这样的过程一直持续到最末尾的样本。(x^m, y^m). 在这种在线学习的背景下，我们关心的是算法在此过程中出错的总次数。因此，这适合需要一边学习一边给出预测的应用情形。

接下来，我们将对感知器学习算法（perceptron algorithm）的在线学习误差给出一个约束。为了让后续的推导（subsequent derivations）更容易，我们就用正负号来表征分类的标签，即假设 $y \in \{-1, 1\}$。回忆一下感知器算法（在第二章有讲到），其中参数 $\theta \in R^{n+1}$, 该算法根据下面的方程来给出预测：

$$h(\theta) = g(\theta^T x) \quad (1)$$

其中

$$g(z) = \begin{cases} 1 & if \quad z \le 0 \\ -1 & if \quad z < 0 \end{cases}$$

然后，给定一个训练样本 (x, y)，感知器学习规则（perception learning rule）就按照如下所示进行更新。如果 $h_\theta = y$，那么不改变参数。若二者相等关系不成立，则进行更[1]。

$$\theta = \theta + yx$$

当感知器算法作为在线学习算法运算的时候，每次对样本给出错误判断的时候，则更新参数，在下面的定理给出了这中情况下的在线学习误差的边界约束。要注意，下面的错误次数的约束边界与整个序列样本的个数 m 不具有特定的依赖关系（explicit dependence），和输入特征的维度 n 也无关。

定理（Block 1962 and Novikoff 1962）。设有一个样本序列:$(x^1, y^1), (x^2, y^2), (x^3, y^3), \cdots, (x^m, y^m)$。假设对于所有的 i，都有 $\| x^{(i)} \le \|$，更近一步存在一个单位长度的向量 $u(\| u \|_2 = 1)$ 对序列中的所有样本都满足 $y^{(i)} \cdot (u^T x^i) \ge \gamma$（例如，$u^T x^i \ge \gamma \quad if \quad y^{(i)} = 1$, 而 $u^T x^{(i)} \le \gamma$，若 $y^{(i)} = -1$，则 u 就以一个宽度至少为 γ 的分界分开了样本数据。）而此感知器算法针对这个序列给出错误预测的综述的上限为 $(D/\gamma)^2$

证明：感知器算法每次只针对出错的样本进行权重更新。设 $\theta^{(k)}$ 为犯了第 k 个错误（$k_t hmistake$）的时候的权重。则 $\theta^{(1)} = 0$（因为初始权重为零），若第 k 个样本发生了错误在样本 $(x^{(i)}, y^{(i)})$ 则 $g((x^{(i)})^T \theta^{(k)}) \ne y^i$，也就意味着：

$$(x^{(i)})^T \theta^{(k)} y^{(i)} \le 0 \quad (2)$$

另根据感知器算法的定义，我们知道 $\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$ 然后就得到：

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

[1]这和之前我们看到的跟新规则的写法稍微有一点点不一样，因为这里我们把分类标签（labels）改成了 $y \in \{-1, 1\}$, 另外学习速率参数（learning rate parameter）α 也被省去了。这个速来参数的效果只是使用某些固定的常数来对参数 θ 进行缩放，并不会影响生成器的行为效果。

利用一个简单的归纳法（straightforward inductive argument）得到：

$$(\theta^{(k+1)})^T u \geq k\gamma \quad (3)$$

还是根据感知器算法的定义可以得到：

$$\begin{aligned} \| \theta^{(k+1)} \|^2 &= \| \theta^{(k)} + y^{(i)} x^{(i)} \|^2 \\ &= \| \theta^{(k)} \|^2 + \| x^{(i)} \|^2 + 2y^{(i)} (x^{(i)})^T \theta^i \\ &\leq \| \theta^{(k)} \|^2 + \| x^{(i)} \|^2 \\ &= \| \theta^{(k)} \|^2 + D^2 \end{aligned} \quad (4)$$

上面这个推导过程，第三步用到了等式（2）。另外这里还要使用一次简单的归纳法，上面的不等式（4）表明：

$$\| \theta^{(k+1)} \|^2 \leq kD^2 \quad (5)$$

把上面的不等式 (3) 和不等式 (4) 结合起来：

$$\begin{aligned} \sqrt{k} D &\geq \| \theta^{(k+1)} \| \\ &\geq (\theta^{(k+1)})^T u \\ &\geq k\gamma \end{aligned}$$

上面的第二个不等式是基于 u，是一个单位长度的向量（$z^T u = \| z \| \cdot \| u \| cos\phi \leq \| z \| \cdot \| u \|$）其中的 ϕ 和向量 z 和向量 u 的夹角）。结果表明 $k \leq (D/\gamma)^2$。因此，如果感知器犯了第 k 个错误，则 $k \leq (D/\gamma)^2$。