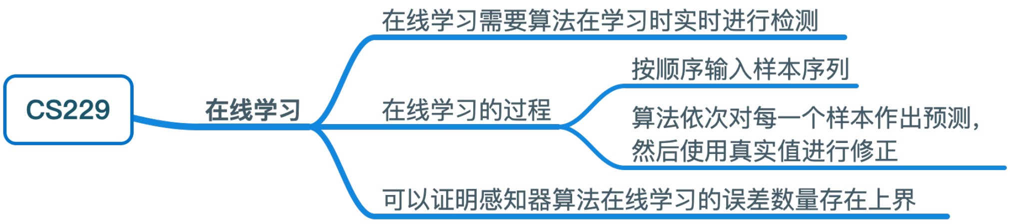
第八章：在线学习



在线学习

- 之前我们讨论的学习都是**批量学习**（batch learning）
  - 批量学习的特点是我们会基于一个训练集进行学习，然后在独立的测试数据上评估学习得到的假设 h
- 本节将讨论**在线学习**（online learning）
  - 在线学习的特点是算法需要在学习时实时进行预测
- 在线学习的过程如下：
  - 一个样本序列 $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$ 会按顺序输入算法
  - 算法依次对每一个样本作出预测，然后使用真实值进行修正
    - 具体来说，算法首先会看到 $x^{(1)}$，然后被要求预测 $y^{(1)}$ 的值，预测完成后 $y^{(1)}$ 的真实值会暴露给算法，对模型进行修正
    - 然后，算法会看到 $x^{(2)}$ ，同样被要求进行预测，重复上一步的操作，直至到达 $(x^{(m)}, y^{(m)})$
- 我们关心的是在线学习在整个过程中产生的误差数量
  - 因此，在线学习是对算法需要在学习过程中进行预测的应用的建模

感知器与大间隔分类器

- 下面将给出感知器算法的在线学习误差数量的上界
  - 为了简化推导，这里将分类标签定义为 $y \in \{-1, 1\}$
- 我们知道感知器算法的参数 $\theta \in \mathbb{R}^{n+1}$，其假设函数为：

$$h_\theta(x) = g(\theta^T x) \tag{1}$$

  - 其中：

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

- 给定一个训练样本 (x, y)，感知器算法的学习规则如下：
  - 如果 $h_\theta(x) = y$，那么参数不发生变化，否则：

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

    - 该规则与第二章的相比有所不同，因为这里分类标签为 $\{-1, 1\}$
    - 此外，学习速率被省略了，这只会影响到参数的大小，对算法本身的行为没有影响
- 下面的定理将给出感知器算法在线学习误差数量的上界

- 当其作为在线算法运行时，每一次得到分类样本错误的时候会进行一次更新
- 注意下面给出的误差数量上界与序列样本数量 m 和输入维度 n 并没有直接关系

- **定理**（Block, 1962, and Novikoff, 1962）：
  - 给定一个样本序列 $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$，假设对于所有的样本，都有 $||x^{(i)}|| \leq D$（欧几里得范数），并且存在一个单位长度向量 u（$||u|| = 1$），使得对于所有样本都有：$y^{(i)} \cdot (u^T x^{(i)}) \geq \gamma$ 成立（即 u 将数据以至少为 γ 的间隔分离）
  - 那么感知器算法对于该序列的总预测误差数量最多为 $(D/\gamma)^2$
- **证明**：
  - 感知器只有当发现错误时才会更新参数，定义 $\theta^{(k)}$ 为出现第 k 个错误时的权重
    - 那么有 $\theta^{(1)} = \vec{0}$（因为权重初始化为0）
    - 如果第 k 个错误出现时的样本为 $(x^{(m)}, y^{(m)})$，那么 $g((x^{(i)})^T \theta^{(k)}) \neq y^{(i)}$，即：

$$(x^{(i)})^T \theta^{(k)} y^{(i)} \leq 0 \tag{2}$$

  - 根据感知器算法的学习规则，我们有 $\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$，据此有：

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

  - 通过一个简答的数学归纳法证明，可以得到：

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

  - 此外，我们还有：

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

    - 第三步使用了公式 (2) 的结论
    - 基于数学归纳法可以得到：

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

  - 将 (3) 和 (5) 式结合起来可以得到：

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

    - 第二步的推导来自于 $z^T u = ||z|| \cdot ||u|| \, cos\phi \leq ||z|| \cdot ||u||$
  - 因此，如果感知器算法发现了第 k 个错误，可以证明 $k \leq (D/\gamma)^2$