

-----机器学习-----

1. 监督学习

(1) 决策树（不都是线性分类器，可以表示非线性函数）

- 随机变量的熵（更大的熵意味着更少的信息，更多的不确定性）： $H(V) = -\sum p(v_k) \log_2 p(v_k)$
- 布尔随机变量的熵： $B(q)$ 与上面的类似
- 属性A的信息增益： $\text{Gain}(A) = B(p/p+n) - \text{Reminder}(A)$
- 测试剩余熵： $\text{Reminder}(A) = \sum_{k=0}^n \left(\frac{p_k+n_k}{p+n} \right) B\left(\frac{p_k}{p_k+n_k}\right)$
- 决策树将选择有最大信息增益的结点
- 拓展：随机森林，XGBoost

(2) 线性回归与分类

- 回归：最小化均方误差，略
- 分类：感知器学习规则。对于数据不可分离的情况，采用Sigmoid函数进行概率化表示。对于多分类问题，采用Softmax函数表示出每一个类别的概率。最后都是最大似然估计。

$$\max_w ll(w) = \max_w \sum_i \log P(y^{(i)} | x^{(i)}; w)$$

感知器学习规则

- 如果真实的 $y \neq h_w(x)$ (出错), 调整权重
 - 如果 $w \cdot x < 0$ 但输出应是 $y=1$
 - False negative 假阴性
 - x_i 为正时 w_i 增加 (为了使输出为1, 应使 $w \cdot x$ 更大)
 - x_i 为负时 w_i 减小
 - 如果 $w \cdot x > 0$ 但输出应是 $y=0$
 - False positive 误报
 - x_i 为正时 w_i 减小 (为了使输出为0, 应使 $w \cdot x$ 更小)
 - x_i 为负时 w_i 增加
- 什么样的参数更新规则可以反映上面的思想呢?
- 感知器学习规则:
 - $w \leftarrow w + \alpha (y - h_w(x)) x$

learning rate 学习速率

+1, -1, or 0 (no error)

- 支持向量机SVM、最近邻等（属于非参数模型）。

(3) 训练策略


- 随机梯度 (SGD), 批梯度 (BGD), 小批量梯度 (MBGD)、正则化

2. 神经网络


- 全连接、CNN、RNN、ResNet、Attention、Transformer
- 关于梯度下降的算法参考“深度学习”——“梯度下降”

3. 统计学习和贝叶斯网络

- 极大似然估计
- 拉普拉斯平滑



拉普拉斯平滑

- 假设我们看到三个正面: $\theta_{ML} = 0$ 是一个合理的估计吗? 
- 带强度 α 的拉普拉斯平滑:
 - 假装你在开始之前看到了每一个结果 α 次
 - $\theta_{lap} = (h+\alpha)/[(h+\alpha) + (t+\alpha)]$
 - $= (3+\alpha)/(3+2\alpha)$
 - 通常, 对于一个 K 值变量:
 - $\theta_k = (N_k+\alpha) / \sum_k (N_k+\alpha) = (N_k+\alpha) / (N + K\alpha)$
 - 对于 $\alpha \gg N$, θ_k 趋于 $1/K$ (一致/均匀先验)
 - 对于 $\alpha \ll N$, θ_k 趋向于 N_k/N (ML 估计)

13

- 朴素贝叶斯