No.

Date

CS182 总复习 Decision Tree

Preliminary: Entropy 相关 ($\log x$ 默认为 $\log_2 x$，且 $0\log 0 = 0$)

$$H(X) = -\sum_{x\in X} p(x)\log p(x) = \mathbb{E}\left[\log \frac{1}{p(x)}\right]$$

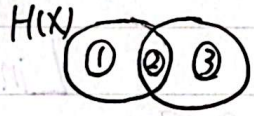
条件熵: $H(Y|X) = \sum_{x\in X} p(x)\, H(Y|X=x)$

$$= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log p(y|x)$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)} = \mathbb{E}\left[\log\frac{1}{p(y|x)}\right]$$

互信息: $I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$

$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Venn Graph: 

⇒ $H(X)$: ①+② $H(Y)$: ②+③ $H(X|Y)$ = ① $H(Y|X)$ = ③ $I(X;Y)$ = ②

(ID3) 在划分属性时，选用 $\arg\max_i I(X_i;Y)$，若假设：样本类型为 $(X_1, X_2, \cdots, Y)$（样本范围 $X_j = x_j$ 为先前划分依据）

互信息为依据外，也可用 Training Error 为划分依据

或：Gini 指数 ← (CART 算法)

Decision Tree 有可解释性，但易 overfit。为缓解 overfit，pruning 是主要手段。分为两种：

预 pruning：结点划分前先估计是否能带来决策树泛化性能提升

后 pruning：对一棵 train 后的树自下而上对非叶结点考察

No.

Date

KNN.

给一个新 datum，欲预测 label，则找 K个最近的 training set 中的点，majority vote 出其 label（依 training points' label）.

为了处理 ties. 可：① 若二分类，2|K ② 距离 → 权重

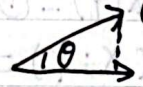
③ 移去k个中最远的 ④：考虑新 distance Metric

Complexity: train: $O(1)$ predict: $O(MN)$

Performance: 有理论保证！ → (≈ best you can do)

error rate $\overset{(<)}{\leq} 2 \times$ Bayes Error Rate

Perceptron Linear Classifier.

Preliminary:  求 $proj_b a$?

$\vec{a}\cdot\vec{b} = a^Tb = \|b\|\|a\|\cos\theta$. 欲：$\|proj_b a\| = \|\vec{a}\|\cos\theta$

则 $proj_b a = \frac{|a^Tb|}{\|b\|^2}\cdot b$ ($\frac{b}{\|b\|}$ 提供 b 方向上单位向量)

(Online) Perceptron 算法：

初：$w=\vec{0}$ $b=0$；每个 $x^{(t)}$，$\hat{y} = sign(w^Tx+b) = \begin{cases} +1, & w^Tx+b\geq 0 \\ -1, & \text{else} \end{cases}$

if $\hat{y}=-1$, $y=1$, $w \leftarrow w+x^{(t)}$, $b \leftarrow b+1$

if $\hat{y}=1$, $y=-1$, $w \leftarrow w-x^{(t)}$, $b \leftarrow b-1$

$\Rightarrow$ $\begin{cases} w \leftarrow w + y^{(t)}x^{(t)} \\ b \leftarrow b+1 \end{cases}$ if prediction 错了

更新原理：如 $\hat{y}=-1, y=1$，欲 w^Tx+b ↑，先 $w=w+x$

则 $(w+x)^Tx+b = (w^Tx+b) + (x^Tx) \geq w^Tx+b$

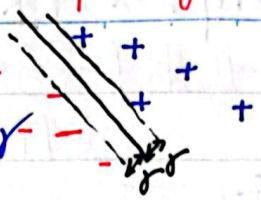
可见：w可视为 $x^{(1)},\cdots,x^{(N)}$ 线性组合.

One trick: $\vec{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$, $\vec{\theta} = \begin{bmatrix} w \\ b \end{bmatrix}$. 则判断 $\theta^Tx \gtrless 0$,

且 $\theta \leftarrow \theta + y^{(t)}x^{(t)}$ 若 $\hat{y}\neq y$

Campus

△: Perceptron 也有 overfitting 问题



Margin: γ: i.e., $\forall i.\ y^{(i)}(\theta^* \cdot x^{(i)}) \geq \gamma$

$\|x^{(i)}\| \leq R$.

，则若数据都在 norm dist 为 R 的"球"中. 则 online perceptron：

的原点为圆心的 !!

最多 $(R/\gamma)^2$ 次错误。

且 $\|\theta^*\|=1$

Proof: 设 θ^* 是完美分类超平面参数，即 $\forall i.\ y^{(i)}(\theta^* \cdot x^{(i)}) \geq \gamma$

则 online 学 θ 时: $\theta^{(k+1)} = \theta^{(k)} + y^{(i)} x^{(i)}$

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

$\theta^{(k+1)}$ 在 θ^* 上投影 $\geq \theta^{(k)} \cdot \theta^* + \gamma$

因 $\theta^{(0)}$ 空初始化，故 $\theta^{(k+1)} \cdot \theta^* \geq k \cdot \gamma$，显然 $\|\theta^{(k+1)}\| \geq k\gamma$

同时: $\|\theta^{(k+1)}\|_2^2 = \|\theta^{(k)} + y^{(i)} x^{(i)}\|_2^2$ ⇒ 余弦定理. 点求.

$= \|\theta^{(k)}\|_2^2 + y^{(i)2}\|x^{(i)}\|_2^2 + 2y^{(i)}(\theta^{(k)} \cdot x^{(i)})$ ≤0, 因为是分错的!

$\leq \|\theta^{(k+1)}\|_2^2 + \|x^{(i)}\|_2^2 \leq \|\theta^{(k+1)}\|_2^2 + R^2$

因 $\theta^{(0)}$ 空初始化，故 $\|\theta^{(k+1)}\|_2^2 \leq kR^2$

$\therefore\ k^2\gamma^2 \leq \|\theta^{(k+1)}\|_2^2 \leq kR^2$, $k \leq R^2/\gamma^2$

SVM & Kernel Method

Goal: 找到最好的划分 hyperplane。"正中间" ⇒ 最鲁棒

可见有三个要求: ① 正确划分正负样本 ② 在正负样本"中间"

③ 离正负样本尽可能远

① : $\begin{cases} w^T x_i + b \geq 0, & y_i = +1 \\ w^T x_i + b \leq 0, & y_i = -1 \end{cases}$

② : 设离 hyper plane 最近的正负样本 x_*^+ & x_*^-，则应，

$$\frac{|w^T x_*^+ + b|}{\|w\|} = \frac{|w^T x_*^- + b|}{\|w\|}$$，且: $\frac{w^T x_i + b}{\|w\|} \geq \frac{w^T x_*^+ + b}{\|w\|}$

$$\frac{w^T x_i + b}{\|w\|} \leq \frac{w^T x_*^- + b}{\|w\|}$$

KOKUYO

$w^T x_+ + b = -1$.

放缩 w 与 b，会让超平面无穷解，故约束：$w^T x_+ + b = 1$.

$$\therefore \begin{cases} w^T x_i + b \geq 1, & y_i = +1 \\ w^T x_i + b \leq -1, & y_i = -1 \end{cases}$$

且取中等号的数据称为支持向量 ☆

这样一来，$r = \frac{1}{\|w\|}$，$d = \frac{2}{\|w\|}$

↑ 优化角度上等价

③ 问题构造，$\min_{w,b} \frac{1}{2}\|w\|^2$，s.t.，$y_i(w^T x_i + b) \geq 1$

解它流程：Lagrange 得对偶问题 (dual problem)

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i^* (1 - y_i(w^T x_i + b))$$

*对每条约束的 Lagrange 乘子

where $\alpha = (\alpha_1; \alpha_2; \cdots; \alpha_m)$，L 对 w, b 偏导：

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i, \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

这个式子代回，并考虑 $\alpha_i \geq 0$ 的约束，对偶问题诞生：

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \cdots, m$$

解出 α 后，$f(x) = w^T x + b$

$$= \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$$

注意到原优化问题的不等式约束，故应满足 KKT 条件：

① $\alpha_i \geq 0$ ② $y_i f(x_i) - 1 \geq 0$ ③ $\alpha_i (y_i f(x_i) - 1) = 0$

对于 (x_i, y_i) 若 $\alpha_i = 0$，则无影响；若 $\alpha_i > 0$，则必有 $y_i(f(x_i)) = 1$，即样本位于最大边界

说明训练完成后，大部分训练样本不需保留，最终模型与支持向量有关

可见，核心便是解对偶问题了！Beyond the scope of 182 (SMO 算法).

No.

Date

在上述讨论中，我们假设样本 linearly seperable. 但现实中 hyperplane 不是'plane'了，如 XOR 问题！

对于这样的问题，要将样本从原始空间映射到一个更高维的特征空间，使得在新空间中线性可分。

Good fact! 若原始空间是有限维，一定存在一个高维特征空间使样本可分。

设 $\phi(x)$ 为 x 映射后的特征向量

$\Rightarrow$ 对偶问题 $\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)$

$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \quad \alpha_i \geq 0$

通常，$\phi(x)$ 空间维度很高，算内积可能很难；可设想：

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \phi(x_i)^T \phi(x_j)$$

即：x_i 与 x_j 在特征空间的内积等于它们在原始样本空间中通过 $K(\cdot,\cdot)$ 计算的结果

则 $f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b$，$K(\cdot,\cdot)$ 为核函数

若知 $\phi(\cdot)$，$K(\cdot,\cdot)$ 不难；但现实中通常不知形式！

怎样的函数能作为核函数？一个对称函数对应的核矩阵半正定。核矩阵：对于 $D = \{x_1, \dots, x_m\}$：

$$K = \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_m) \\ \vdots & \ddots & \vdots \\ K(x_m,x_1) & \cdots & K(x_m,x_m) \end{bmatrix} \quad K_{ij} = K(x_i, x_j).$$

半正定：$\forall x \in \mathbb{R}^n, x \neq \vec{0}$，则均：$x^T A x \geq 0$

常见核函数：$K(x_i, x_j) = x_i^T x_j$；$(x_i^T x_j)^d$（多项式核）；

$\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$（Gauss）；$\exp\left(-\frac{\|x_i - x_j\|}{\sigma}\right)$（Laplace）；$\tanh(\beta x_i^T x_j + \theta)$（Sigmoid kernal）

且若 K_1, K_2 为核 func，$\gamma_1 K_1 + \gamma_2 K_2$ 也是；$\forall g(x)$，则：

$K(x, z) = g(x) K_1(x, z) g(z)$ 也是

Campus

No.

Date

Logistic Regression ⚠ 分类任务

*: $y = \frac{1}{1+e^{-(w^Tx+b)}}$, $\ln \frac{p(y=1|x)}{p(y=0|x)} = w^Tx+b$

y视作为 x 为正例可能性！

通过 MLE 估计 w, b，目标：最大化似然

即：$l(w,b) = \sum_{i=1}^{m} \ln p(y_i|x_i;w,b)$

对于 $\forall i$，二中其一为0

$p(y_i|x_i;w,b) = y_i p_1(\hat{x}_i;\beta) + (1-y_i) p_0(\hat{x}_i;\beta)$

其中：$\beta^T\hat{x} = w^Tx+b$ (trick)，$p_1 = p(y=1|\hat{x};\beta)$

$\Leftrightarrow l(\beta) = \sum_{i=1}^{m} (-y_i\beta^T\hat{x}_i + \ln(1+e^{\beta^T\hat{x}_i}))$

Logistic Regression 无闭式解，可用 GD、SGD等

$\mapsto h_\theta(x) = p(y=1|x)$

*：可见对于 SGD for Logistic Regression，update为：

$\theta_k \leftarrow \theta_k + \lambda(h_\theta(x^{(i)}) - y^{(i)})x_k^{(i)}$

不止它！ Least mean square $h_\theta(x) = \theta^Tx$

Perceptron $h_\theta(x) = \text{sign}(\theta^Tx)$ 更新公式均可以上式表达

KOKUYO

No.

Date

Linear Regression

△: logistics regression 仍是分类

对于线性回归，模型预不再是 $sign(w^T x+b)$ 的 label. 而是实数值。

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_i=(x_{i1};\cdots;x_{id})$. $y_i\in\mathbb{R}$, $x_i\in\mathbb{R}^d$。衡量 $f(x_i)$ 与 y_i 差别：

$$(w^*,b^*)=\underset{w,b}{\operatorname{argmin}}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

$$=\underset{w,b}{\operatorname{argmin}}\sum_{i=1}^{m}(y_i-wx_i-b)^2=\underset{w,b}{\operatorname{argmin}}\,E(w,b)$$

则 $\frac{\partial E(w,b)}{\partial w}=2\left(w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)x_i\right)$

$$\frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)$$

闭式解：$w=\dfrac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}(\sum_{i=1}^{m}x_i)^2}$，$b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)$

更一般的，D 表示为 $X\in\mathbb{R}^{m\times(d+1)}$（1是'1'，trick!），$w\in\mathbb{R}^{d+1}$

欲：$w^*=\underset{w}{\operatorname{argmin}}\,[(\vec{y}-Xw)^T(\vec{y}-Xw)]\;Ew$

$\frac{\partial Ew}{\partial w}=2X^T(Xw-\vec{y})$，则 $X^TXw=X^T\vec{y}$

$w^*=(X^TX)^{-1}X^Ty$

KOKUYO

No.

Date

梯度下降与BP

减号 学习率

GD: $\theta \leftarrow \theta^{(0)}$, while not converged do: $\theta \leftarrow \theta - \gamma \nabla_\theta J(\theta)$

Eg: 一次更新的复杂度: $O(ND)$, N: D中N个样本, D: 九维度为D

这是对于 GD 使用所有样本的 loss 结果作回传的结果

SGD: $\theta \leftarrow \theta^{(0)}$, while not converged do:

for $i \in$ shuffle $([1, 2, \cdots, N])$ do: $\theta \leftarrow \theta - \gamma \nabla_\theta J^{(i)}(\theta)$

Mini-batch: 每一次取k个样本的子集, 作GD

* GD中的 $\nabla_\theta J(\theta)$ 要考虑"平均"

GD learning rate ↑, 则曲线上轨迹越波折。SGD比GD更易波动。SGD比GD要训更多步, 但每步都比GD cheaper.

对于 BP, 要知道一个 Basic fact: 一个神经网络由一组operation 组成:

$f_L(w_L, f_{L-1}(w_{L-1}, f_{L-2} \cdots, f_1(w_1, x)) \cdots)$

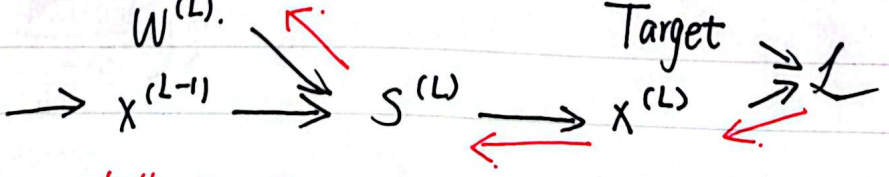
(Basically) All f functions are linear + (simple) non-linear operator, 如 $\varphi(z) = \frac{1}{1+e^{-z}}$, 非常好的性质为:

$$\varphi'(z) = -\frac{-e^{-z}}{(1+e^{-z})^2} = \varphi(z) \cdot (1 - \varphi(z))$$, 非常好算!

Forward:

1st layer: $s^{(1)} = W^{(1)T} x^{(0)}$, $x^{(1)} = \sigma(s^{(1)})$

2nd layer: $s^{(2)} = W^{(2)T} x^{(1)}$, $x^{(2)} = \sigma(s^{(2)})$ ... loss L



若更新 $\frac{\partial L}{\partial W^{(L)}}$: follow "→"

$$\frac{\partial L}{\partial W^{(L)}} = \frac{\partial L}{\partial x^{(L)}} \cdot \frac{\partial x^{(L)}}{\partial s^{(L)}} \cdot \frac{\partial s^{(L)}}{\partial W^{(L)}}$$

要具体数值, 则 forward pass 中的 $x^{(L)}$, $s^{(L)}$, $W^{(L)}$ 结果都记好:

$$\frac{\partial L}{\partial W^{(L)}}\Big|_{W^{(L)}} = \frac{\partial L}{\partial x^{(L)}}\Big|_{x^{(L)}} \cdot \frac{\partial x^{(L)}}{\partial s^{(L)}}\Big|_{s^{(L)}} \cdot \frac{\partial s^{(L)}}{\partial W^{(L)}}\Big|_{W^{(L)}}$$