Final Review

Decision Tree

△ H & I

$H(Y) = \sum P(Y_i) \log P(Y_i)$

$H(Y|X) = \sum P(X_i) H(Y|X_i)$

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

x_1	x_2	y
1	0	0
1	0	0
1	0	1
1	0	1
1	1	1
1	1	1
1	1	1
1	1	1

$I(Y;X_1) = H(Y) - H(Y|X_1)$

$H(Y) = \frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4} = -\frac{1}{2} + \frac{3}{4}\log_2\frac{3}{4}$

$H(Y|X_1) = 1 \cdot H(Y|x_1=1) = 1\log_2 1 = 0$

$H(Y|X_2) = \frac{1}{2}H(Y|x_2=0) + \frac{1}{2}H(Y|X_2=1)$

$= \frac{1}{2}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{2}\cdot(1\cdot\log_2 1)$

$= -\frac{1}{2}$

$\therefore I(Y;X_2) = \frac{3}{4}\log_2\frac{3}{4} > I(Y;X_1) = -\frac{1}{2} + \frac{3}{4}\log_2\frac{3}{4}$

决策树分裂：每次分裂 $I(X_i;Y)$ 最大之 X_i

5. [20 points] [Decision Tree] A dataset is given below. Now we want to discover the relationship between the features and the target variable by using a Decision Tree.

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Play Tennis? (Y)
sunny	hot	high	no
overcast	hot	high	yes
rain	mild	high	yes
rain	cool	normal	yes
sunny	mild	high	no
sunny	mild	normal	yes
rain	mild	normal	yes
overcast	hot	normal	yes

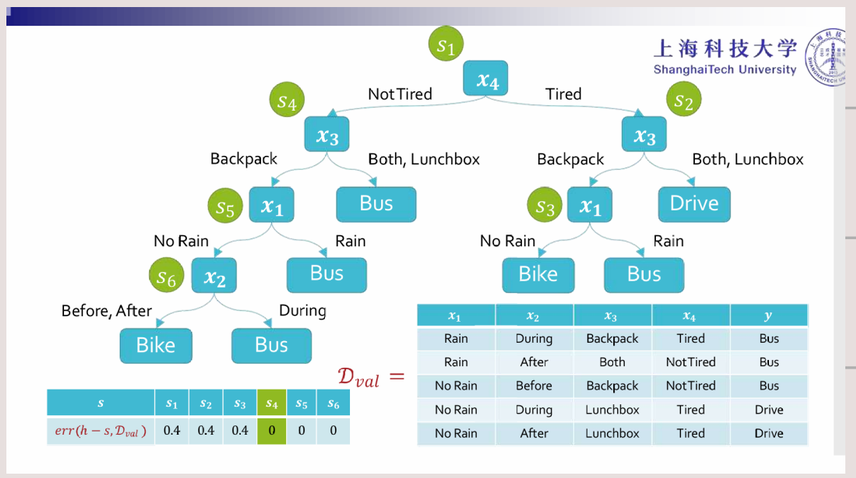
(a) Using the dataset above, calculate the mutual information for each feature (X_1, X_2, X_3) to determine the root node for a Decision Tree trained on the above data.
- What is $I(Y;X_1)$? [3 points]
- What is $I(Y;X_2)$? [3 points]
- What is $I(Y;X_3)$? [3 points]
- What feature should be split on at the root node? [1 points]

(b) Calculate what the next split should be. [5 points]

(c) Draw the resulting tree. [5 points]

剪枝：所有非叶节点编号 S_i

$err[S_i]$ 表示 S_i 子树缩为一点后训练误差



$\mathcal{D}_{val} =$

x_1	x_2	x_3	x_4	y
Rain	During	Backpack	Tired	Bus
Rain	After	Both	NotTired	Bus
No Rain	Before	Backpack	NotTired	Bus
No Rain	During	Lunchbox	Tired	Drive
No Rain	After	Lunchbox	Tired	Drive

s	s_1	s_2	s_3	s_4	s_5	s_6
$err(h-s, \mathcal{D}_{val})$	0.4	0.4	0.4	0	0	0

KNN

训练集：已有点集（有分类标签）

新测试点：寻找最近 k 个训练点 ⇒ 投票决定标签

$N \to +\infty$：

$error(h) < 2 \times$ Bayes Error rate

平局处理：加点 / 移点 / 加权 / 换距离函数

Perceptron

参数：$\vec{w}$、b → bias 量

预测：$sign(\vec{w}^T\vec{x} + b) = sign\left(\begin{bmatrix}\vec{w}\\ b\end{bmatrix}^T\begin{bmatrix}\vec{x}\\ 1\end{bmatrix}\right) = h^{(t)}$

若 $h^{(t)} = y^{(t)}$：参数不变

$h^{(t)} \neq y^{(t)}$：$\begin{cases}\vec{w} += y^{(t)}\vec{x}^{(t)} \\ b += y^{(t)}\end{cases} \Rightarrow \begin{bmatrix}\vec{w}\\ b\end{bmatrix} += y^{(t)}\begin{bmatrix}\vec{x}^{(t)}\\ 1\end{bmatrix}$

x_1	x_2	$\hat{y}$	y	Mistake?
−1	2	+	−	Yes
1	0	+	+	No
1	1	−	+	Yes
−1	0	−	−	No
−1	−2	+	−	Yes
1	−1	+	+	No

（初始 $\vec{w} = \begin{bmatrix}0\\0\end{bmatrix}$，又不考虑 b）

△ Margin：可分问题中最大之点-边界距离

最小值 → Margin = γ

点-边界距离 → $\vec{w}^T\vec{x} + b = 0$

算点-边界距离：

边界上一点为 $\vec{x}'$，当前点为 $\vec{x}''$

距离 $= \left|(\vec{x}' - \vec{x}'')^T \frac{\vec{w}}{\|\vec{w}\|}\right|$ → 单位法向量

$= \left|\vec{x}'^T\vec{w} - \vec{x}''^T\vec{w}\right|\frac{1}{\|\vec{w}\|} = \left|\vec{w}^T\vec{x}'' + b\right|\frac{1}{\|\vec{w}\|}$

△ Mistake Bound

$R \cdot \gamma \Rightarrow$ 犯错次数 $= \left(\frac{R}{\gamma}\right)^2$

证：$\vec{\theta}^{(k)}$ 在 $\vec{x}^{(i)}, y^{(i)}$ 处犯错：

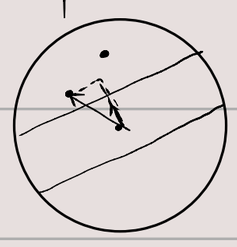
$$y^{(i)} \vec{x}^{(i)T} \vec{\theta}^{(k)} \leq 0$$

$$\Rightarrow \vec{\theta}^{(k+1)} = \vec{\theta}^{(k)} + y^{(i)} \vec{x}^{(i)}$$

∵ 边界 $= \gamma$

∴ 存 $\vec{w}$，$\|\vec{w}\| = 1$，使得 $\forall$ 训练点：

$$y^{(i)} \vec{x}^{(i)T} \vec{w} \geq \gamma$$



$$\vec{\theta}^{(k+1)T} \vec{w} = \vec{\theta}^{(k)} \vec{w} + y^{(i)} \vec{x}^{(i)T} \vec{w}$$

$$\geq \vec{\theta}^{(k)} \vec{w} + \gamma$$

$$\therefore \vec{\theta}^{(k+1)T} \vec{w} \geq k\gamma \quad ①$$

又∵ $\|\vec{x}^{(i)}\| \leq R$

算模长：$\|\vec{\theta}^{(k+1)}\|^2 = \left(\vec{\theta}^{(k)} + y^{(i)} \vec{x}^{(i)}\right)^2$

$$= \vec{\theta}^{(k)2} + \vec{x}^2 + 2\left(y^{(i)} \vec{x}^{(i)T} \vec{\theta}^{(k)}\right)$$

$\vec{x}^2 \leq R$；$2\left(y^{(i)} \vec{x}^{(i)T} \vec{\theta}^{(k)}\right) \leq 0$

$$\leq \|\vec{\theta}^{(k)}\|^2 + R^2$$

$$\Rightarrow \|\vec{\theta}^{(k+1)}\|^2 \leq kR^2 \Rightarrow \|\vec{\theta}^{(k+1)}\| \leq \sqrt{k} R \quad ②$$

① ② $\Rightarrow \sqrt{k} R \geq \|\vec{\theta}^{(k+1)}\| \geq \vec{\theta}^{(k+1)T} \vec{w} \geq k\gamma$

$$\Rightarrow k \leq \left(\frac{R}{\gamma}\right)^2$$

Kernel

$x^{(i)} \rightarrow \phi(x^{(i)}) \Rightarrow x^{(i)}$ 各特征一系列组合映射

$\phi(x_1)\phi(x_2) = K(x_1, x_2)$

SVM

目标：找出分类器 $\vec{w}$：

满足约束 $\begin{cases} \|\vec{w}\|_2 = 1 \\ \forall (\vec{x}_i, y_i)，有\ y_i(\vec{w}^T \vec{x}_i + b) \geq \gamma \end{cases}$ （边界计算）

一情况下，使 γ 最大

$$\Rightarrow \vec{w} = \arg\max_{\vec{w}} \gamma$$

For：$\begin{cases} \|\vec{w}\| = 1 \\ y_i(\vec{w}^T \vec{x}_i + b) \geq \gamma \quad (\forall i) \end{cases}$

变为：$\|\vec{w}\|$ 固定 γ 变化 $\Rightarrow \gamma$ 固定 $\|\vec{w}\|$ 变化

约束：$y_i\left(\frac{\vec{w}^T}{\gamma} \vec{x}_i + b\right) \geq 1 \Rightarrow y_i(\vec{w}' \vec{x}_i + b) \geq 1$（$\vec{w}'$ 与变化量）

目标：γ 最大 $\Rightarrow \vec{w}'$ 最小

问题变为：

$$\vec{w}' = \arg\min_{\vec{w}'} \frac{1}{2}\|\vec{w}'\|^2$$

使得 $\forall i$：$y_i(\vec{w}'^T \vec{x}_i + b) \geq 1$

即：$\min \frac{1}{2} \vec{w}'^2$

使 $\forall i$：$y_i(\vec{w}^T \vec{x}_i + b) \geq 1$

$\rightarrow -y_i(\vec{w}^T \vec{x}_i + b) + 1 \leq 0$

即为：$\min \max L(w, b, \alpha)$

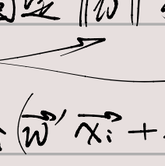
Lagrange 对偶：求 $\max \min L(w, b, \alpha)$

此时可消去 γ，只保留 $\vec{w}$ 来间接记录 γ 变化情况和 $\vec{w}$ 原值

可设为一对变化值：$\vec{w}' = \frac{\vec{w}}{\gamma}$，则 $\|\vec{w}'\|$ 变化只取决于 γ 变化

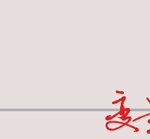
$\vec{w}'$ 意义：集成了 $\vec{w}$ 原值和 γ 变化规律：原本为 $\vec{w}$ 为自变量，γ 为 $\vec{w}$ 直接响应量（特定 $\vec{w}$ 必对应特定 γ）和 $\vec{w}$ 关键特征 $\|\vec{w}\|$ 固定 $\Rightarrow \vec{w}'$ 则用此特征记录下了 γ 响应情况，∴ 将 $\vec{w}$ 变化与 γ 响应集成。

图示：



$\angle \vec{w} = \theta_1$ 时：有 $\gamma_{(\theta_1)}$

$\angle \vec{w} = \theta_2$ 时，有 $\gamma_{(\theta_2)}$

$\Rightarrow$ 

θ 变化时，$\|\vec{w}\|$ 可用于记录 γ，θ 记录原本 $\vec{w}$ 变化情况！

↳（可用特征二利用）

$$\therefore L \overset{(\vec{w},b,\vec{\alpha})}{=} \frac{1}{2}\|\vec{w}\|^2 + \sum_{i=1}^{m}\alpha_i\left[-y_i(\vec{w}^T\vec{x_i}+b)+1\right]$$

$$\nabla_w L(\vec{w},b,\vec{\alpha}) = \vec{w} + \sum_{i=1}^{m}\alpha_i(-y_i\vec{x_i}) = 0$$

$$\Rightarrow \vec{w} = \sum_{i=1}^{m}\alpha_i y_i \vec{x_i}$$

$$\nabla_b L(\vec{w},b,\vec{\alpha}) = \sum_{i=1}^{m}-(\alpha_i y_i) = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i = 0$$

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

$$= \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \vec{x_i}^T\vec{x_j} - \sum_{i,j=1}^{m}(\alpha_i y_i \vec{x_i})^T(\alpha_j y_j \vec{x_j}) + \alpha_i y_i b + \sum_{i=1}^{m}\alpha_i$$

$$= -\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \vec{x_i}^T\vec{x_j} - b\underbrace{\sum_{i=1}^{m}\alpha_i y_i}_{=0} + \sum_{i=1}^{m}\alpha_i$$

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

↳ L 自变量为 $\vec{\alpha}$ $(\alpha_i \geq 0)$

对偶问题：

找 $\arg\max_{\alpha} \sum\alpha_i - \frac{1}{2}\sum\alpha_i\alpha_j y_i y_j \vec{x_i}^T\vec{x_j}$

$$s.t.\begin{cases}\forall i\ \alpha_i \geq 0\\ \forall i\ \sum\alpha_i y^{(i)} = 0\end{cases}$$

⇒ 可优化 ✓

找出 $\vec{\alpha}$ 后：$\vec{w} = \sum\alpha_i y_i \vec{x_i}$ （b 也可得出）

由 KKT 条件：若 $\vec{x_i}$ 在边界上，$\alpha_i \neq 0$

否则 $\alpha_i = 0$ ↳ 支持向量

△ 软边界分类

目标：$\arg\min_{\vec{w},\xi} \frac{1}{2}\|\vec{w}\|^2 + C\sum_i\xi_i$

$$s.t.\ \forall i,\begin{cases} y_i\vec{w}^T\cdot\vec{x_i} \geq 1-\xi_i \quad \text{↳ 允许误差}\\ \xi \geq 0\end{cases}$$

$$L = \frac{1}{2}\|\vec{w}\|^2 + C\sum\xi_i - \sum\alpha_i[y_i(\vec{x}^T\vec{w}+b)-1+\xi_i] - \sum\beta_i\xi_i$$

↳ 参数：$\vec{w}$、b、ξ_i

↳ 求导：$C-\alpha_i-\beta_i = 0$

$$\Rightarrow 0 \leq \alpha_i \leq C$$

∴ 对偶问题：

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

$$s.t.\begin{cases}\forall i\ \underline{0 \leq \alpha_i \leq C}\ \text{（唯一一改变）}\\ \forall i\ \sum\alpha_i y_i = 0\end{cases}$$

Linear Regression

$$J(\vec{w},b) = \frac{1}{N}\sum_{i=1}^{N}\left|y^{(i)} - (\vec{w}^T\vec{x}^{(i)}+b)\right|^2$$

优化：$\vec{\theta} = \arg\min_{\theta} J(\vec{\theta})$ $\left(\vec{\theta} = [b\ \vec{w}]\right)$

梯度：$= \arg\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=1}^{m}\theta_j x_j^{(i)}\right)^2$ → 标量

$$\vec{\theta} -= \eta\frac{\partial J(\theta)}{\partial\vec{\theta}} \Rightarrow \theta_i -= \frac{\partial J(\theta)}{\partial\theta_i}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\vec{\theta}^T\vec{x}^{(i)} - y^{(i)})\cdot x_j^{(i)}$$

↳ 批量梯度

随机梯度：$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}J^{(i)}(\theta)$

$$J^{(i)}(\theta) = \frac{1}{2}(\vec{\theta}^T\vec{x}^{(i)} - y^{(i)})^2$$

$$\frac{\partial J^{(i)}(\theta)}{\partial\theta_i} = (\vec{\theta}^T\vec{x}^{(i)} - y^{(i)})\cdot\vec{x}^{(i)}$$

闭式解：$\nabla_\theta J(\vec{\theta}) = \frac{1}{N}\frac{\partial}{\partial\vec{\theta}}\sum\frac{1}{2}(\vec{\theta}^T\vec{x}^{(i)} - y^{(i)})^2$

$$= \frac{1}{2N}\frac{\partial}{\partial\vec{\theta}}(X\vec{\theta}-\vec{y})^T(X\vec{\theta}-\vec{y})$$

↳ $(\vec{\theta}^TX^TX\vec{\theta} - 2\vec{\theta}^TX^T\vec{y} + \vec{y}^T\vec{y})$

$$= \frac{1}{2N}(2X^TX\vec{\theta} - 2X^T\vec{y}) = 0$$

$$\Rightarrow \vec{\theta} = (X^TX)^{-1}X^T\vec{y}$$

↳ 奇可逆

MLE & MAP (Lec 9)

等待被2

Logistic Regression

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\vec{\theta}^T\vec{x}}} = h_{\vec{\theta}}(\vec{x})$$

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)\left(1-\sigma(z)\right)$$

$$\begin{cases} P(Y=1 \mid \vec{x};\vec{\theta}) = h_\theta(\vec{x}) \\ P(Y=0 \mid \vec{x};\vec{\theta}) = 1-h_\theta(\vec{x}) \end{cases}$$

似然: $L(\vec{\theta}) = \prod_{i=1}^{n} \left(h_{\vec{\theta}}(\vec{x}^{(i)})\right)^{Y^{(i)}} \left(1-h_{\vec{\theta}}(\vec{x}^{(i)})\right)^{(1-Y^{(i)})}$

$$l(\vec{\theta}) = \sum_{i=1}^{n}\left(Y^{(i)} \log h_{\vec{\theta}}(\vec{x}^{(i)}) + (1-Y^{(i)}) \log\left(1-h_{\vec{\theta}}(\vec{x}^{(i)})\right)\right)$$

$$= \sum_{i=1}^{n} Y^{(i)} \log \frac{1}{1+e^{-\vec{\theta}^T\vec{x}}} + (1-Y^{(i)}) \log \frac{e^{-\vec{\theta}^T\vec{x}}}{1+e^{-\vec{\theta}^T\vec{x}}} \quad \to \frac{1}{1+e^{\vec{\theta}^T\vec{x}}}$$

$$= \sum_{i=1}^{n} Y^{(i)}\vec{\theta}^T\vec{x} - \log\left(1+e^{\vec{\theta}^T\vec{x}}\right) + (1-Y^{(i)})\left(-\log\left(1+e^{\vec{\theta}^T\vec{x}}\right)\right)$$

$$= \sum_{i=1}^{n} Y^{(i)}\vec{\theta}^T\vec{x}^{(i)} - \log\left(1+e^{\vec{\theta}^T\vec{x}}\right) \quad ☆$$

$$J(\vec{\theta}) = \frac{-1}{n} l(\vec{\theta})$$

优化: 梯度下降.

$$\nabla_\theta J(\vec{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\left(Y^{(i)}\vec{x}^{(i)} - \frac{e^{\vec{\theta}^T\vec{x}}\cdot\vec{x}^{(i)}}{1+e^{\vec{\theta}^T\vec{x}}}\right)$$

($\to g(\vec{\theta}^T x)$; $\to P(Y^{(i)}=1 \mid x^{(i)};\vec{\theta})$)

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(Y^{(i)} - g(\vec{\theta}^T\vec{x}^{(i)})\right)\vec{x}^{(i)}$$

$$\Rightarrow J(\vec{\theta}) -= \eta \nabla_\theta J(\vec{\theta})$$

若用 SGD: $\nabla_\theta J^{(i)}(\vec{\theta}) = -\frac{1}{n}\left(Y^{(i)} - g(\vec{\theta}^T\vec{x}^{(i)})\right)\vec{x}^{(i)}$

Mini SGD: $\nabla_\theta J^{(i_s)}(\vec{\theta}) = -\frac{1}{n}\sum_{s=1}^{S}\left(Y^{(s)} - g(\vec{\theta}^T\vec{x}^{(s)})\right)\vec{x}^{(s)}$

Regularization

$$J(\vec{\theta}) \to J(\vec{\theta}) + \lambda R(\vec{\theta})$$

$$R(\vec{\theta}) = \|\vec{\theta}\|_2^2 \text{ (L2)} \quad \text{or} \quad \|\vec{\theta}\|_1 \text{ (L1)}$$

$\to \sum \theta_m^2$ $\quad$ $\to \sum |\theta_m|$ (零点处不可微)

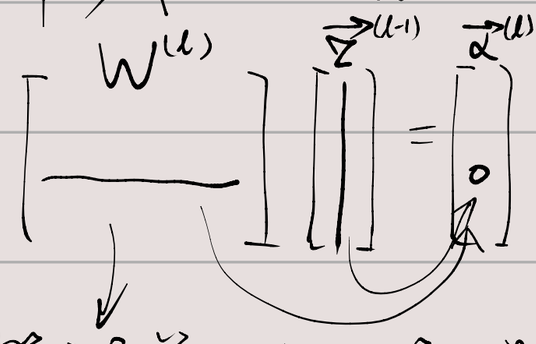
Neural Network

△ 每层传播: $a_j^{(l)}$ 与 $z_j^{(l)}$ (线性 & 激发)

$$a_j^{(l)} = \sum_{i=0}^{D} w_{j,i}^{(l)} z_i^{(l-1)}, \quad z_j^{(l)} = f(a_j^{(l)})$$

↳ 枚举 l 层节点 (含 1)

推导形式: $\vec{a}^{(l)} = W^{(l)}\vec{z}^{(l-1)}$, $\vec{z}^{(l)} = [1, f(\vec{a}^{(l)})]^T$



第 j 行第 i 列 ⇒ 行向量为 l-1 层后一权重向量

↳ 行数 = 下一层向量维数

若多分类: 最后一层 Output 激发为 Softmax

△ 反向传播

$$\begin{cases} L(W) = \frac{1}{N}\sum L_i + R(W) \\ L_i(x_i, y_i, W) = -\log \underbrace{P(Y=y_i \mid X=x_i)}_{\text{即: 预测成功二概率}} \end{cases}$$

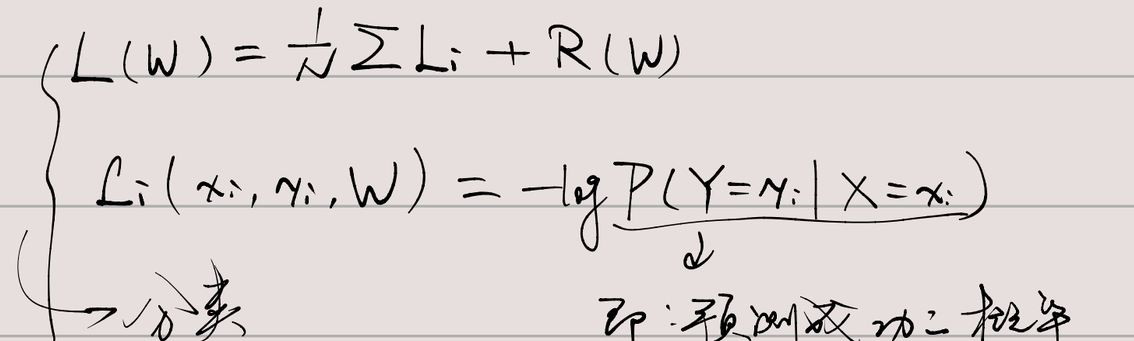
↳ 分类

回归: 最小二乘 or 交叉熵

$$W^{(t+1)} = W^{(t)} - \eta \frac{\partial E}{\partial W^{(t)}}$$

↳ 根据计算图逐层求各参数梯度

$$\begin{cases} \text{末层: 激发梯度} \to \text{对 } W \text{ 梯度} \\ \text{余层: 对 } W \text{ 改为对 } W^{(L)} \end{cases}$$



$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(L)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}^{(L)}}\frac{\partial \mathbf{x}^{(L)}}{\partial \mathbf{s}^{(L)}}\frac{\partial \mathbf{s}^{(L)}}{\partial \mathbf{x}^{(L-1)}}\frac{\partial \mathbf{x}^{(L-1)}}{\partial \mathbf{s}^{(L-1)}}\frac{\partial \mathbf{s}^{(L-1)}}{\partial \mathbf{W}^{(L-1)}}$$

The order needs to be reversed for Jacobians!

Computing the loss:

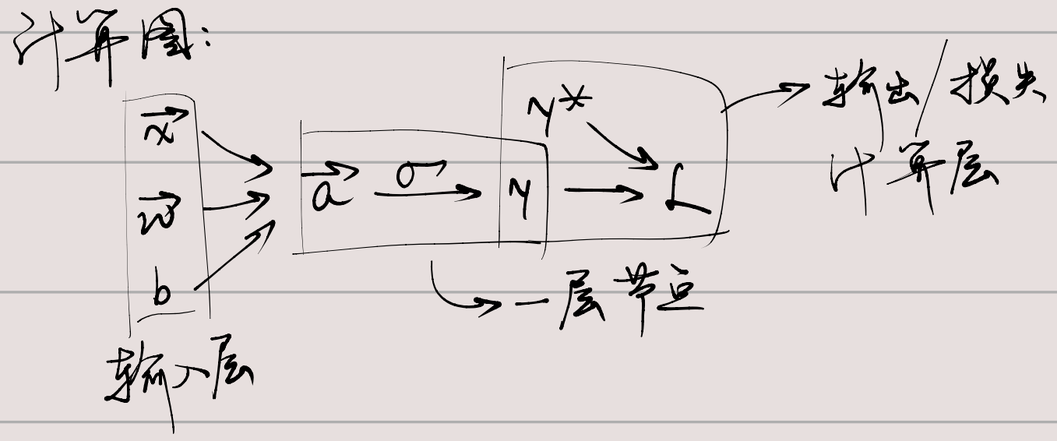
$$s = wx + b$$
$$y = \sigma(s)$$
$$\mathcal{L} = \frac{1}{2}(y - t)^2$$

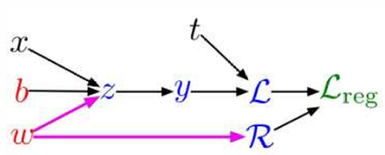
Computing the derivatives:

$$\frac{d\mathcal{L}}{dy} = y - t$$
$$\frac{d\mathcal{L}}{ds} = \frac{d\mathcal{L}}{dy}\sigma'(s)$$
$$\frac{d\mathcal{L}}{dw} = \frac{d\mathcal{L}}{ds}x$$
$$\frac{d\mathcal{L}}{db} = \frac{d\mathcal{L}}{ds}$$

$\vec{x} \rightarrow \vec{a} \rightarrow \vec{y} \rightarrow \mathcal{L}$

计算图：





Forward pass:

$$z = wx + b$$
$$y = \sigma(z)$$
$$\mathcal{L} = \frac{1}{2}(y - t)^2$$
$$\mathcal{R} = \frac{1}{2}w^2$$
$$\mathcal{L}_{\text{reg}} = \mathcal{L} + \lambda\mathcal{R}$$

Backward pass:

$$\delta_{\mathcal{L}_{\text{reg}}} = 1$$
$$\delta_{\mathcal{R}} = \delta_{\mathcal{L}_{\text{reg}}}\frac{d\mathcal{L}_{\text{reg}}}{d\mathcal{R}} = \delta_{\mathcal{L}_{\text{reg}}}\lambda$$
$$\delta_{\mathcal{L}} = \delta_{\mathcal{L}_{\text{reg}}}\frac{d\mathcal{L}_{\text{reg}}}{d\mathcal{L}} = \delta_{\mathcal{L}_{\text{reg}}}$$
$$\delta y = \delta_{\mathcal{L}}\frac{d\mathcal{L}}{dy} = \delta_{\mathcal{L}}(y - t)$$
$$\delta_z = \delta_y\frac{dy}{dz} = \delta_y\sigma'(z)$$
$$\delta_w = \delta_z\frac{dz}{dw} + \delta_{\mathcal{R}}\frac{d\mathcal{R}}{dw} = \delta_z x + \delta_{\mathcal{R}}w$$
$$\delta_b = \delta_z\frac{dz}{db} = \delta_z$$

其他部分如节等待放之

Matrix Calculus

等待放之

CNN

等待放之

Recommender Systems

等待放之

Boosting / Adaboosting

初始数据：权重 $D_1(i) = \frac{1}{N}$（等重）

数据记为 $(\vec{x}_1, y_1) \cdots (\vec{x}_n, y_n)$，$y_i \in \{-1, 1\}$

△每轮循环：

* 根据当前数据训练一个新的分类为 h_t
* 计算 h_t 误差 ε_t：$\varepsilon_t = \sum_{i=1}^{N} 1\{h_t(\vec{x}_i) \neq y_i\} \cdot D_t(i)$ → 误分类的权重和
* 计算该分类器权重 α_t：$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$

⇒ ε_t 高：分类不好 ⇒ α_t 低

ε_t 低：分类效果好 ⇒ α_t 高

* 更新数据的权重：

Scoring 缩放值：$S_i = \exp(-\alpha_t \cdot y_i h_t(\vec{x}_i))$ （正分类则 $y_i h_t(\vec{x}_i) = 1$，误分类则 $= -1$）

$S = \exp(-\alpha_t)$，下轮关注性下降

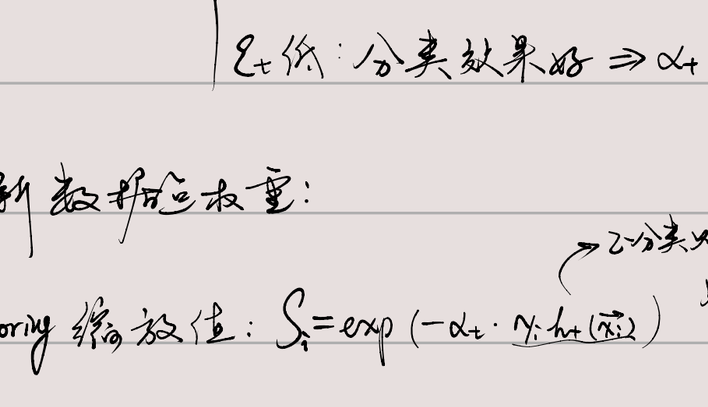
$S = \exp(\alpha_t)$，下轮关注更关注

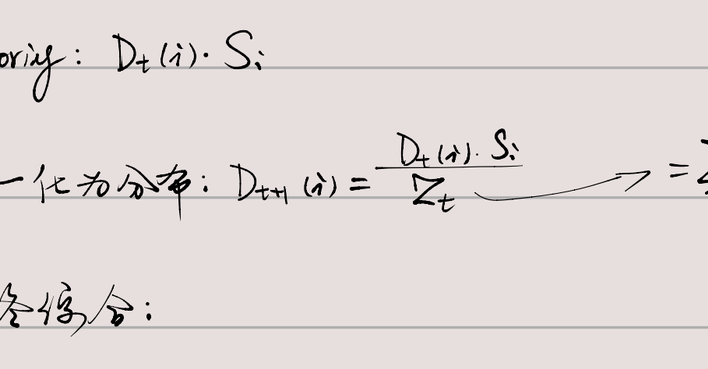
Scoring：$D_t(i) \cdot S_i$

归一化为分布：$D_{t+1}(i) = \frac{D_t(i) \cdot S_i}{Z_t}$，$Z_t = \sum_i D_t(i) \cdot S_i$

△最终结合：

$$H(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(\vec{x})\right)$$ （分类结合）





AdaBoosting 可最小化指数损失原因证明：

等价于2

Bagging

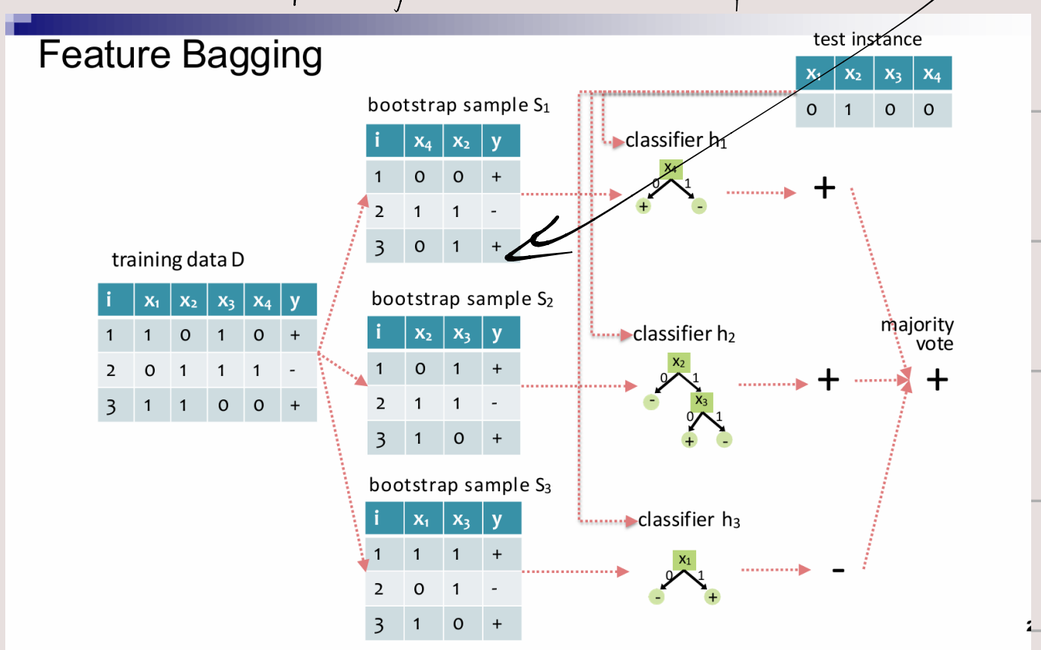
△ Sample Bagging

每组可放回抽样若干数据点，每组训练模型，最终结果为模型投票（分类）或取平均（回归）产生

△ Feature Bagging

将"取样数据点"改为"取样特征"

→ 同样 with Replacement



△ Random Forests

Sample Bagging + Feature Bagging

↓ 子数据集构建 → 分裂时只考虑部分特征

⇒ 自助数据集 + 分裂特征随机化

⇒ 训练出若干个决策树，投票或取平均

（参数为 B、p）

超参数选择：等价于2

EM

△ Jensen 不等式

凸函数 ⌣：$E(f(X)) \geq f(EX)$

凹函数 ⌢（log）：$E(f(X)) \leq f(EX)$

△ 数据：$\vec{x}_1, \vec{x}_2, \cdots$ 无标签

目标：找一个模型（大前提）

找出最佳参数 θ，使 $l(\theta)=\sum_{i=1}^{m} \log p(\vec{x}_i;\theta)$

设 $z^{(i)}$ 为 $x^{(i)}$ 之潜在变量（如 GMM 中，$z^{(i)}$ 表示 $x^{(i)}$ 属于哪个高斯，$z^{(i)} \sim$ 多项式分布）→ 值域为 $\{1,2,3,\cdots,k\}$

$\Rightarrow l(\theta)=\sum_{i=1}^{m} \log \sum_{z_i} p(\vec{x}_i, z_i;\theta)$

即：对潜在变量 z 为中介，找出最优 θ

△ 优化推导

对于 $l(\theta)=\sum_{i=1}^{m} \log \sum_{z_i} p(\vec{x}_i, z_i;\theta)$：

$$l(\theta)=\sum_{i=1}^{m} \log \sum_{z_i} Q_i(z_i) \frac{p(\vec{x}_i, z_i;\theta)}{Q_i(z_i)}$$

令随机变量 $X=\frac{p(\vec{x}_i, z_i;\theta)}{Q_i(z_i)}=F(z_i)$，$f(x)=\log x$

则 $l(\theta)=\sum_{i=1}^{m} f(EX) \geq \sum_{i=1}^{m} E(f(X))$（Jensen 不等式）

$$=\sum_{i=1}^{m} \sum_{z_i} Q_i(z_i) \log \frac{p(\vec{x}_i, z_i;\theta)}{Q_i(z_i)} \quad (1)$$

意义推导：(1)式为：固定 θ 时，$l(\theta)$ 为一个下界函数，该函数自变量为 $Q_i(z_i)$，即：取不同 $Q_i(z_i)$（即 z_i 分布），有不同下界值。

⇒ 期望：找到一个 $Q_i(z_i)$，使得下界如下确界使的未最紧。⇒ 即：$\forall$ 随机变量 X，Jensen 不等式都取等。

X 构造得为：$X=c$（常数）

$\Rightarrow \frac{p(\vec{x}_i, z_i; \vec{\theta})}{Q_i(z_i)} = c$

求解 $Q_i(z_i)$：

$Q_i(z_i) \propto p(\vec{x}_i, z_i; \vec{\theta})$

Q 为分布 $\Rightarrow Q_i(z_i) = \frac{p(\vec{x}_i, z_i; \vec{\theta})}{\sum_{z_i} p(\vec{x}_i, z_i; \vec{\theta})}$ → 所有 z_i 求和

$= \frac{p(\vec{x}_i, z_i; \vec{\theta})}{p(\vec{x}_i; \vec{\theta})}$

$= p(z_i | \vec{x}_i; \vec{\theta})$

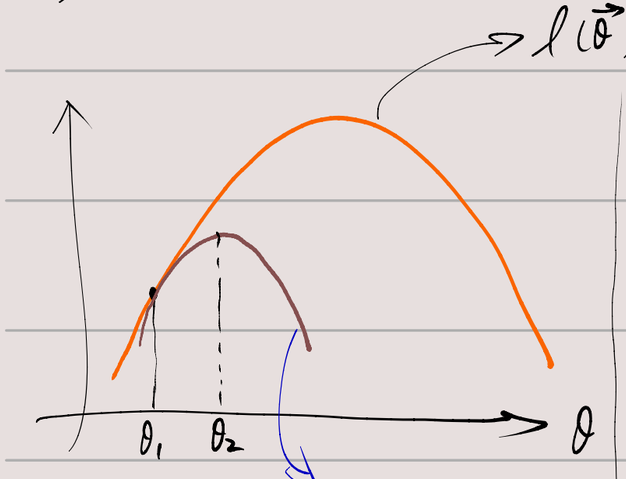
$\Rightarrow$ 此为 E 步骤二：在固定 $\vec{\theta}$ 情况下，找到最合适 "潜在辅助变量"

数学本质为利用 Jensen 不等式右端二函数（左端为 $l(\vec{\theta})$）找到 $l(\vec{\theta})$ 此时最大下界，即最好一优化结果（$= l(\vec{\theta})$）

M 步骤：$\vec{\theta} = \arg\max_{\vec{\theta}} \sum_i \sum_{z_i} Q_i(z_i) \log \frac{p(\vec{x}_i, z_i; \vec{\theta})}{Q_i(z_i)}$ 找出 Jensen 右侧取最大值时 $\vec{\theta}$.

例：

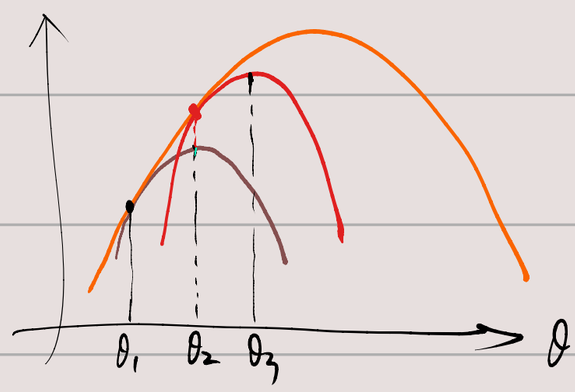
轮次 1



E 步骤：在当前 θ_1 下，找到 Jensen 右侧函数，使其紧贴合 Jensen 右侧函数

M 步骤：沿 Jensen 右侧函数优化，找到其取最优时对应 θ_2

轮次 2：



E 步骤：在当前 θ_2 情况下，找到最优 Jensen 右侧函数，使其紧贴合

Jensen 右侧函数 $= l(\vec{\theta})$

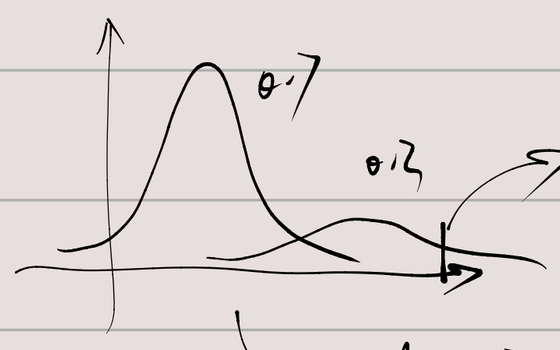
M 步骤：优化 Jensen 右侧函数，迭代新 $\vec{\theta}$.

其余轮次依此类推

GMM → 每点 $\in$ 任一特定高斯（相应）

Problem：已有观测点集. π, μ, Σ

学习 K 个高斯分布（Eg：GM 分布），在此分布中，每点有 K 个概率（分属于哪个高斯） 期望：使观测到该点集概率最大



$p(x) = \sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \Sigma_k)$ ← γ_{ik}

→ 多个高斯加权合成分布.

$\sum_i p(z_i = k) N(x_i | \mu_k, \Sigma_k)$

γ_{ik}

E $M \longrightarrow \mu_k, \Sigma_k, \underline{\pi_k}$ → 权重

↳ 初始?

$p(z_i = k \mid x_i) = \gamma_{ik}$

E: $\gamma_{ik} = p(z^{(i)} = k \mid x^{(i)}; \phi, \mu, \Sigma)$

$\propto$ (贝叶斯)

M: $\ell = \log \prod_{i=1}^{N} \sum_{k=1}^{k} \gamma_{ik} (\pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k))$

$= \sum\sum \gamma_{ik} [\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)]$

↳ $f(\pi_k, \mu_k, \Sigma_k^{-1})$

$$\begin{cases} \frac{\partial \ell}{\partial \pi_k} = 0 \\ \frac{\partial \ell}{\partial \mu_k} = 0 \\ \frac{\partial \ell}{\partial \Sigma_k^{-1}} = 0 \end{cases}$$

$\Rightarrow \pi_k = \frac{1}{n} \sum \gamma_{ik}$

$\mu_k = \frac{\sum \gamma_{ik} x_i}{\sum \gamma_{ik}}$

$\Sigma_k = \frac{\sum \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum \gamma_{ik}}$