期末复习

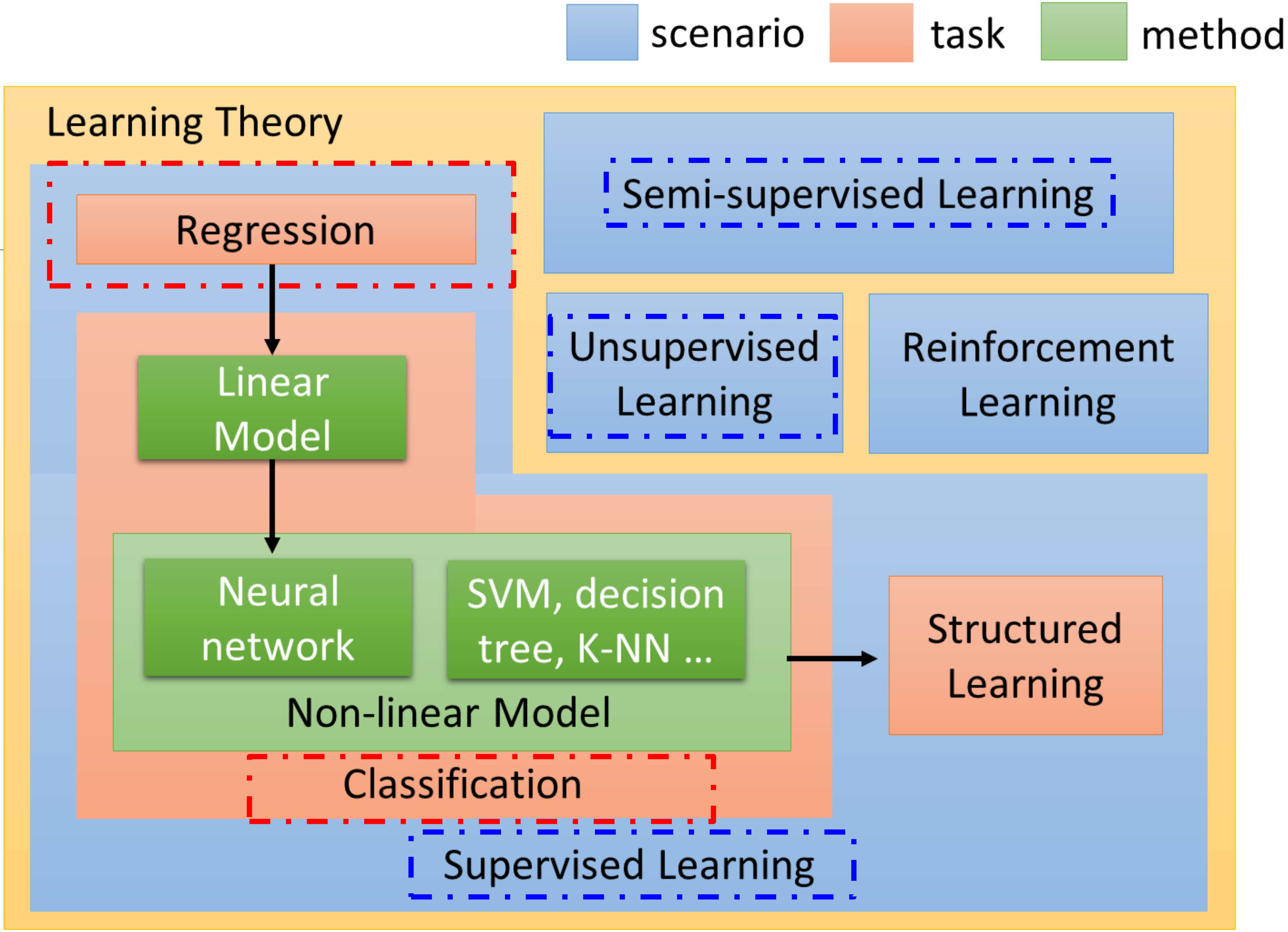
期末考试复习要点

- 选择题
- 简答题
- 计算题
- 综合分析题

- **范围：所有章节，没有编程题，理解每个方法的原理**

总览

- 两大任务：分类&预测
- 数据情况：监督，半监督，无监督



基本概念

- 机器学习三要素
  - **模型**：学习什么样的模型？ 机器学习中模型是指所要学习的（条件概率分布或决策函数）。
  - **策略**：按照什么准则学习或选择最优的模型？
    - Loss function 策略是构造损失函数，其目标是为了（从假设空间中选择最优的模型）。
      - 对数损失函数 或对数似然损失函数 $L(Y,f(X)) = -\log P(Y|X)$
      - 平方损失函数 $L(Y,f(X)) = (Y-f(X))^2$
  - **算法**：学习模型的具体计算方法

    机器学习算法的目标是最小化（学习误差）。

基本概念

- 梯度下降 $w^*, b^* = arg\min_{w,b} L(w,b)$

梯度下降算法中的梯度，其含义是（ 寻找最优方向）。

- (Randomly) Pick an initial value w^0, b^0
- Compute $\frac{\partial L}{\partial w}|_{w=w^0,b=b^0}, \frac{\partial L}{\partial b}|_{w=w^0,b=b^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w}|_{w=w^0,b=b^0} \qquad b^1 \leftarrow b^0 - \eta \frac{\partial L}{\partial b}|_{w=w^0,b=b^0}$$

- Compute $\frac{\partial L}{\partial w}|_{w=w^1,b=b^1}, \frac{\partial L}{\partial b}|_{w=w^1,b=b^1}$

$$w^2 \leftarrow w^1 - \eta \frac{\partial L}{\partial w}|_{w=w^1,b=b^1} \qquad b^2 \leftarrow b^1 - \eta \frac{\partial L}{\partial b}|_{w=w^1,b=b^1}$$

基本概念

◆ 梯度下降——优化

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^t - \eta \boldsymbol{g}_i^t$$

$$\frac{\eta}{\sigma_i^t}$$

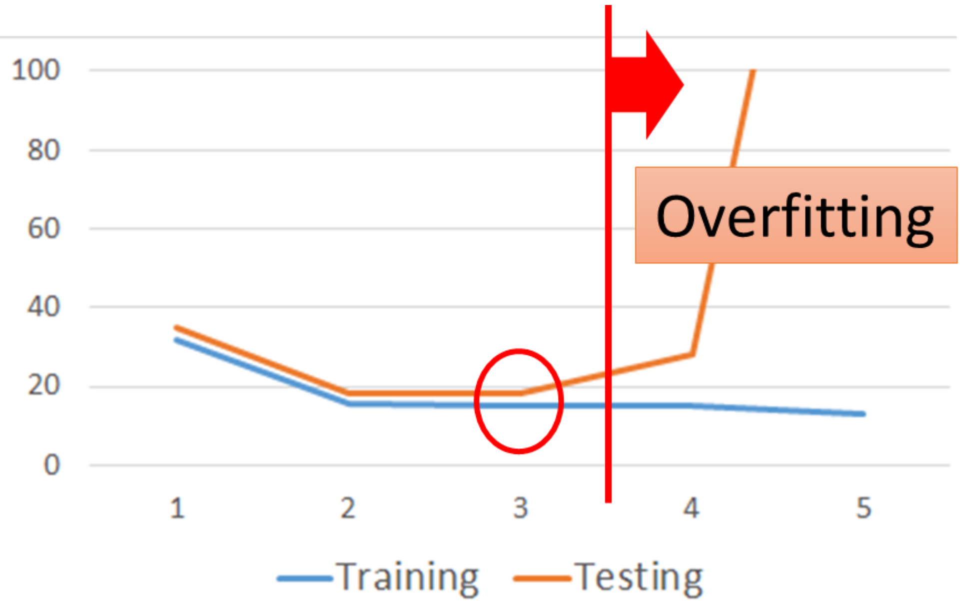
$$\boldsymbol{m}^t = \lambda \boldsymbol{m}^{t-1} - \eta \boldsymbol{g}^{t-1}$$

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} + \boldsymbol{m}^t$$

基本概念

◆**过拟合**：模型对已知数据预测很好，对未知数据预测很差

- 增加训练集数量
- 控制模型复杂度
  - 正则化
  - 剪枝
  - 模型轻量化策略
- 多模型融合
  - 集成



基本概念

- 评估策略
  - 预测：MSE（mean square error）
    - $E = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$
  - 分类：F1

表 2.1　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- TP-将正类预测为正类数
- FN-将正类预测为负类数
- FP-将负类预测为正类数
- TN-将负类预测为负类数

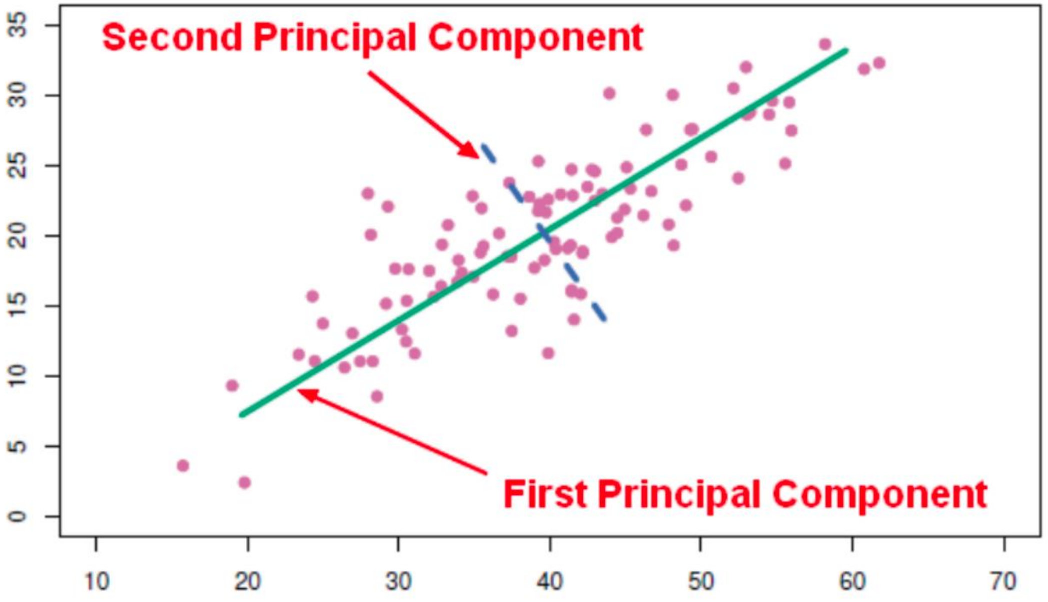
$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

PCA

◆ 数据分析方法，将原始数据变换为一组各维度线性无关的数据表示方法，用于提取数据的主要特征分量，常用于高维数据的降维



主成分理解

最能表现每个样本点特征的维度投影

⇩

方差最大

PCA

- 最大可分性——方差最大
  - 希望投影后的值在一个方向上分散，而这种分散程度，用数学上的方差来表示
  - $$\mathrm{var}(X)=\frac{1}{m}\sum_{i=1}^{m}(X_i-\bar{X})^2$$
    $$\bar{X}=mean(X)$$

□ 目标：方差最大化
□ 通过特征分解，找到特征向量
□ 每一个特征向量均为一个一维基，数据在这个基上的坐标为转换后的坐标

算法

$$S=\frac{1}{N}\sum(x-\bar{x})(x-\bar{x})^T$$

输入：样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
低维空间维数 d'.

过程:

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i-\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$;

2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;

3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;

4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}$.

输出：投影矩阵 $\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})$.

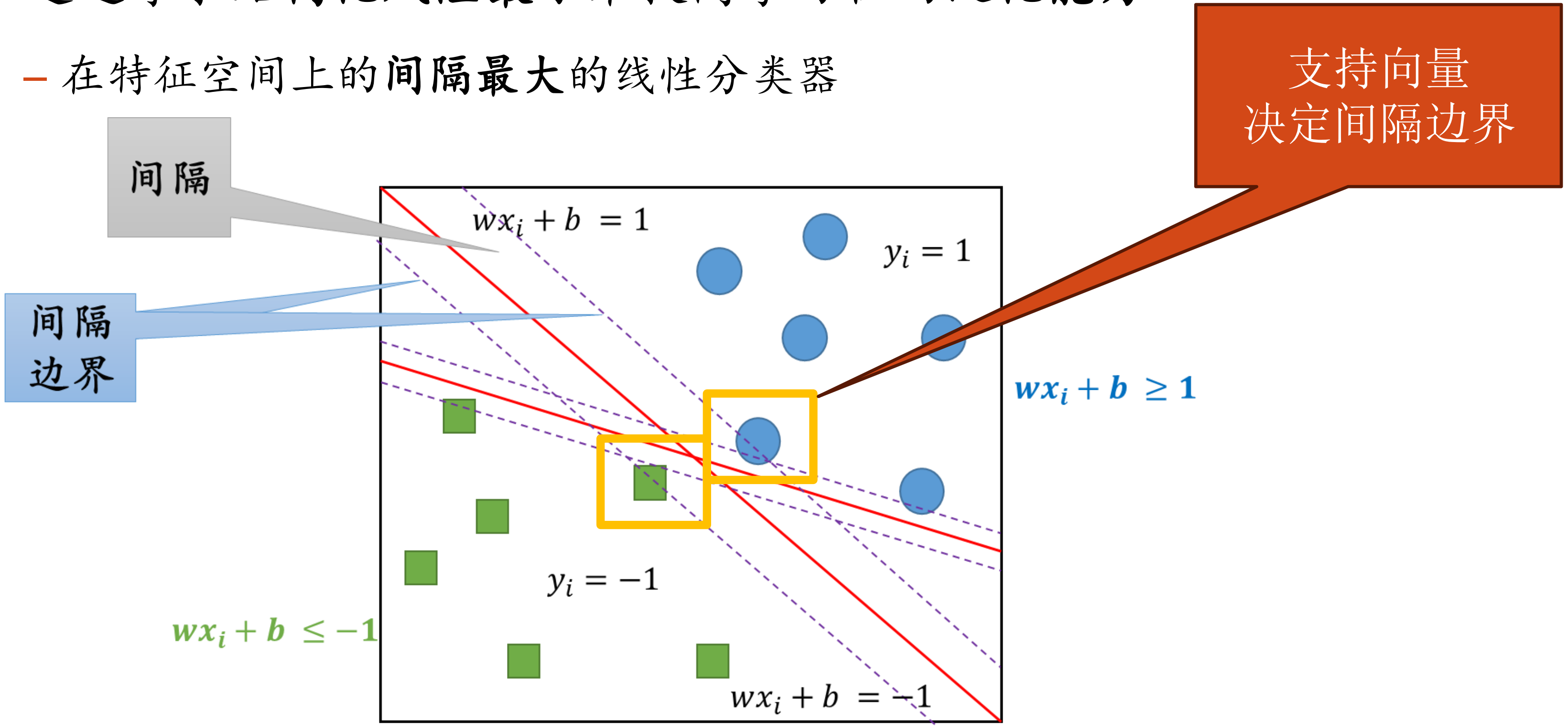
线性分类模型

- 感知机
  - 决策函数：$f(x) = sign(w^T x + b)$
  - **输入**：特征向量；**输出**：实例的类别，取+1和−1二值
  - 中心思想：误分类点驱动
  - Loss function：$L(w,b) = -\sum_{x_i \in M} y_i(w^T x_i + b)$ 误分类点到超平面的距离
  - 算法：随机梯度下降；选择误分类点更新参数

线性分类模型

◆支持向量机SVM

- 通过寻求**结构化风险最小**来提高学习机的**泛化能力**
  - 在特征空间上的**间隔最大**的线性分类器



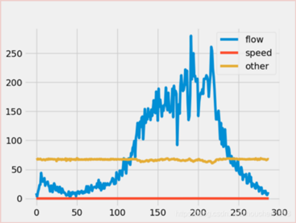
线性分类模型

- 逻辑回归
  - 为了解决连续的线性函数不适合进行分类的问题，引入非线性函数g 来预测类别标签的条件概率$p(y=c|x)$
  - 二分类：$\boldsymbol{p(y=1|x)=g\big(f(x;w)\big)}$
    - 函数f：线性函数
    - 函数g：把线性函数的值域从实数区间“挤压”到了(0,1)之间，可以用来表示概率。

线性分类模型

◆ 逻辑回归

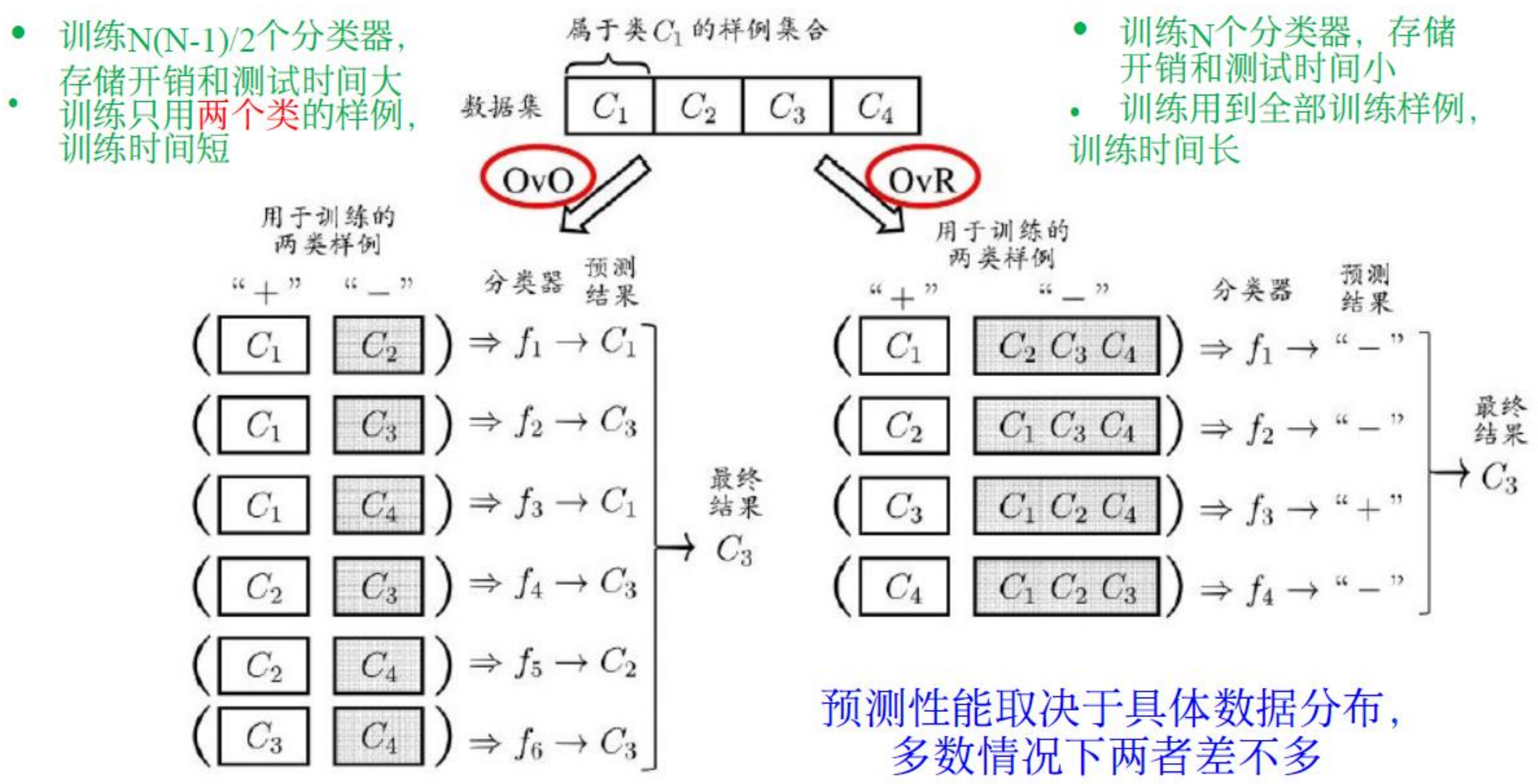
- 决策函数：$\boldsymbol{p(y=1|x)=\frac{1}{1+exp(-(wx+b))}}$
- 目标：最大似然函数 ➡ 交叉熵损失函数

	感知机	逻辑回归	线性回归
训练数据格式			
决策函数	$f=sign(wx+b)$	$f=\frac{1}{1+e^{-(wx+b)}}$	$y=wx+b$
损失函数	$l=-\sum_{x_i\in M} y_i(wx_i+b)$	$l=-\sum_{i=1}^{N}[y_i\ln f(x_i)+(1-y_i)\ln(1-f(x_i))]$	$l=\sum_{i=1}^{N}[f(x_i)-y_i]^2$
训练结果	模型输出为+1/-1 直接给出类别标签	模型输出为(0,1) 根据阈值0.5进行判断，大于0.5属于类别1，小于0.5属于类别0	模型输出预测值

线性分类模型

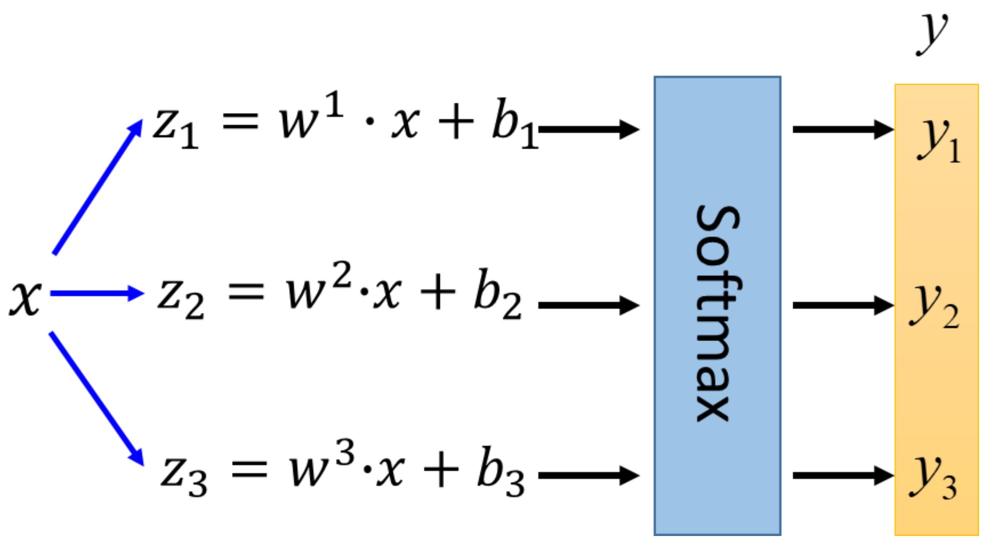
- 逻辑回归
  - 引申到多分类情况
    - 拆解法：
      - 若干个二分类
        - OvO; OvR

拆解法：将一个多分类任务拆分为若干个二分类任务求解



线性分类模型

- 逻辑回归
  - 引申到多分类情况
    - Softmax
      - “一对其余”方式的改进，仍需要C个判别函数



线性分类模型

- 线性可分支持向量机
  - 数据为线性可分数据集（一定能够完全分隔开）

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t.\ \ y_i(wx_i + b) - 1 \geq 0, \qquad i = 1,2,\cdots,N$$

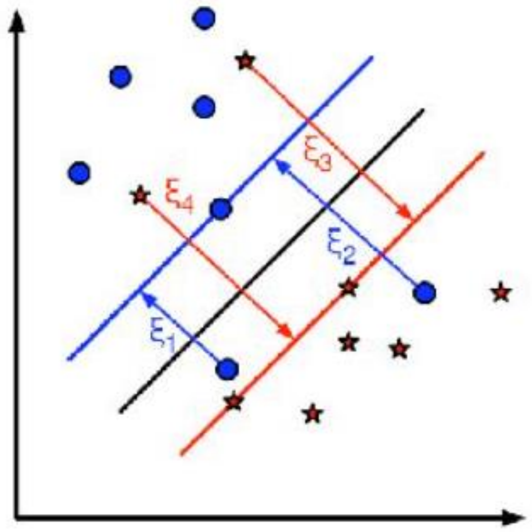
$$\min_{\alpha} L = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y^i y^j (\boldsymbol{x}^i \cdot \boldsymbol{x}^j) - \sum_{i=1}^{N} \alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y^i = 0，\ \ \alpha_i \geq 0$$

线性分类模型

- 线性支持向量机
  - 场景：允许数据集中存在噪声点，即不能完全线性分开的数据

在线性支持向量机的目标式子中，存在一个超参数C，其含义为（错误分类点的容忍性）。



$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t.\ \ y_i(wx_i + b) \geq 1 - \xi_i\,,\quad i = 1,2,\cdots,N$$

C非常大： ξ_i-> 0，意味不允许有噪声点（错误分类的点）
C非常小：允许有一些噪声点

非线性分类模型

- 非线性支持向量机
  - 低维→高维
  - 引入核函数

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i$$

s.t. $\sum_{i=1}^{N}\alpha_i y_i = 0$

$0 \le \alpha_i \le C$

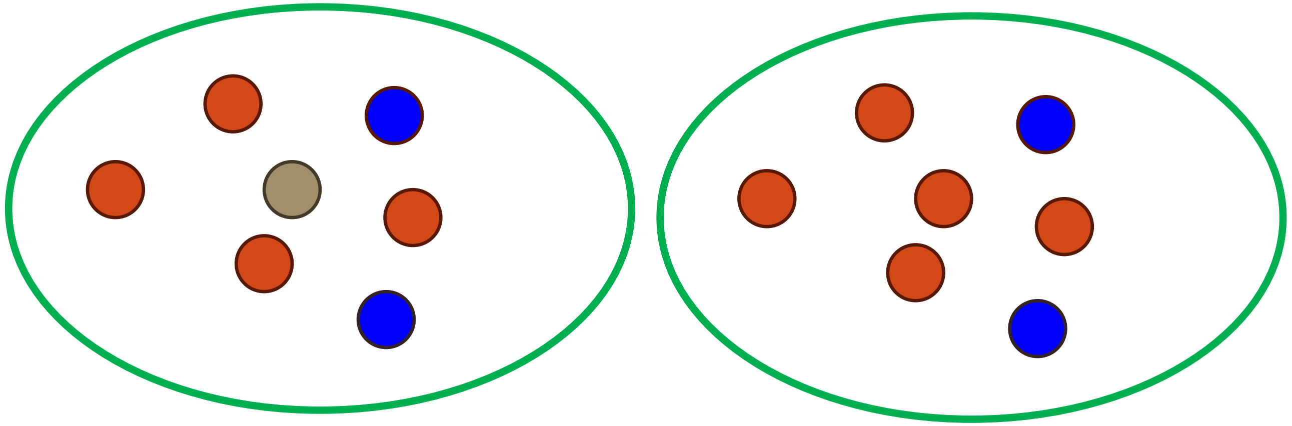
决策函数

$$f(x) = sign(w^T\Phi(x) + b)$$

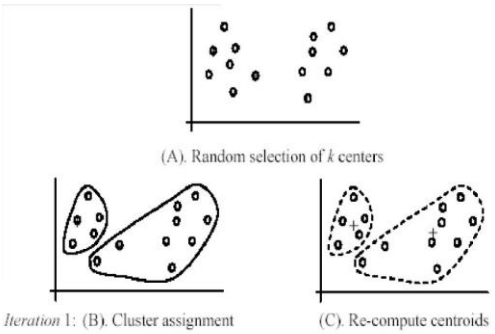
$$f(x) = sign\left(\sum_{i=1}^{N}\alpha^* y_i K(x, x_i) + b^*\right)$$

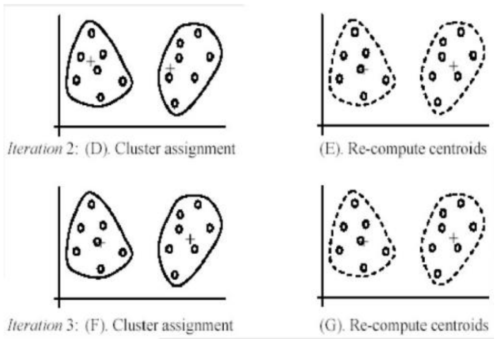
非线性分类模型

- KNN
  - 有监督学习
  - 依据距离函数
    - 欧式距离 $d(A,B)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
  - 投票原则



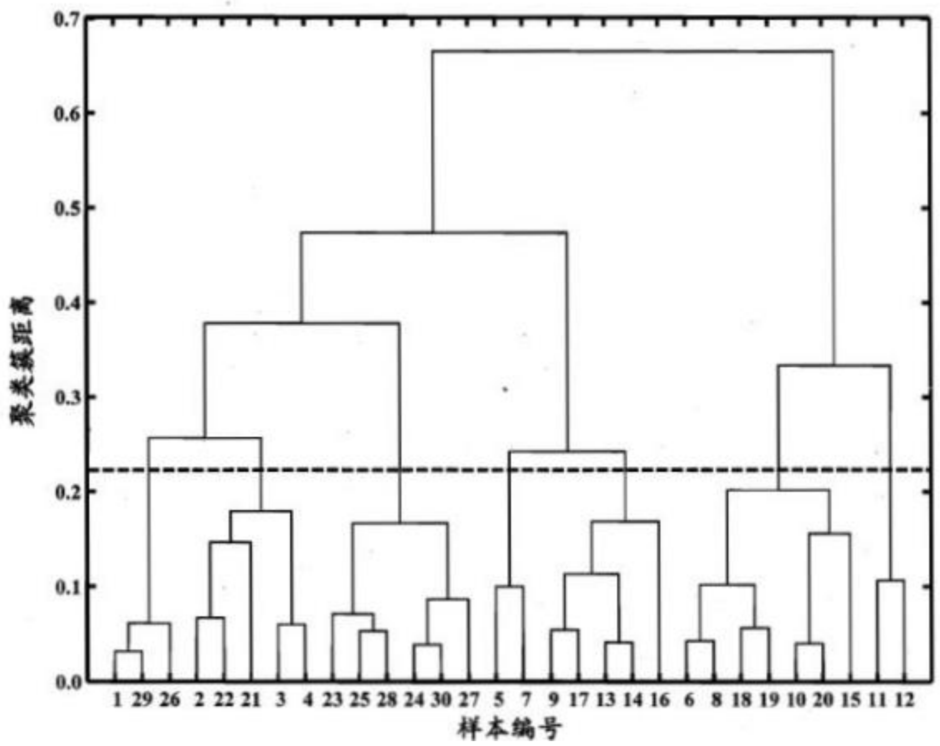
- k-means
  - 无监督学习
  - 迭代的方式选出每个簇中心点
  - 依据距离函数
    - 欧式距离





层次聚类

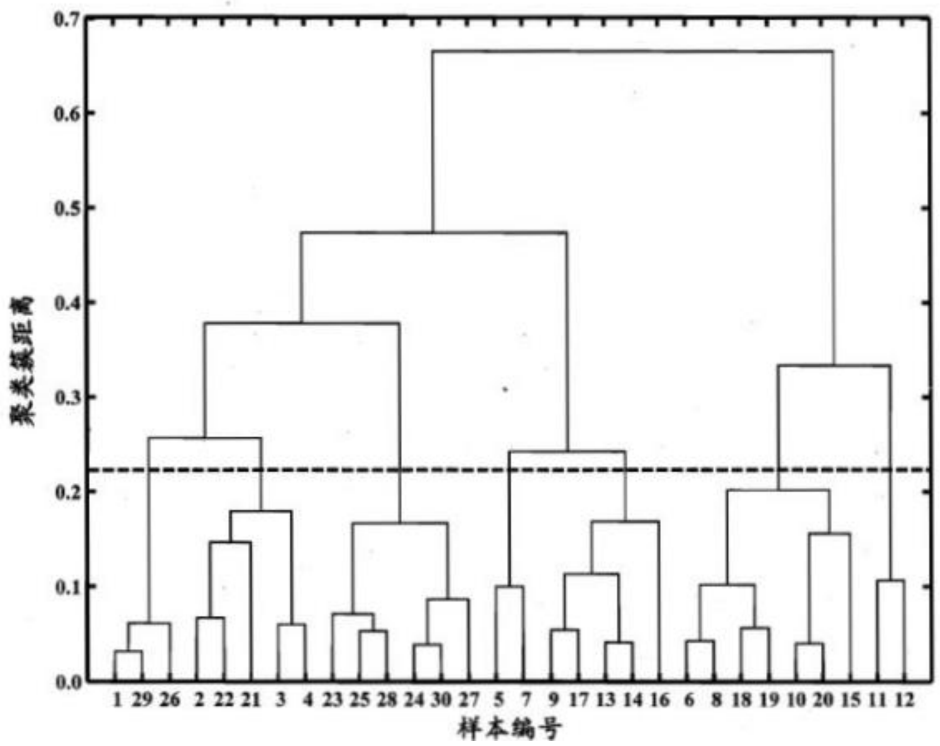
- 从不同层次对数据集进行划分，从而形成树形的聚类结构
  - 聚合：自下而上
  - 分裂：自上而下



层次聚类

- 从不同层次对数据集进行划分，从而形成树形的聚类结构
  - 聚合：自下而上
  - 分裂：自上而下

- 和k-means不同，无需预先指定簇数，可以灵活的确定最终的簇数（仍需人工定义）
- 树状的结构，具有可视化效果
- 需要计算所有关系距离，计算量大，复杂度高
- 对初始数据敏感，一旦某个对象被合并到某个簇中，它就不能再被分配到其他簇中，这可能导致对初始数据的微小变化非常敏感，产生不同的聚类结果



贝叶斯学习

◆**后验概率计算：** $P(B_i|A) = \frac{P(A|B_i)|P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{C} P(A|B_j)(P(B_j)}$

朴素贝叶斯

贝叶斯定理和特征条件独立

$$f(x) = argmaxP(c)\prod_{i=1}^{d} P(x_i|c)$$

主要用于分类，效率较高

贝叶斯网络

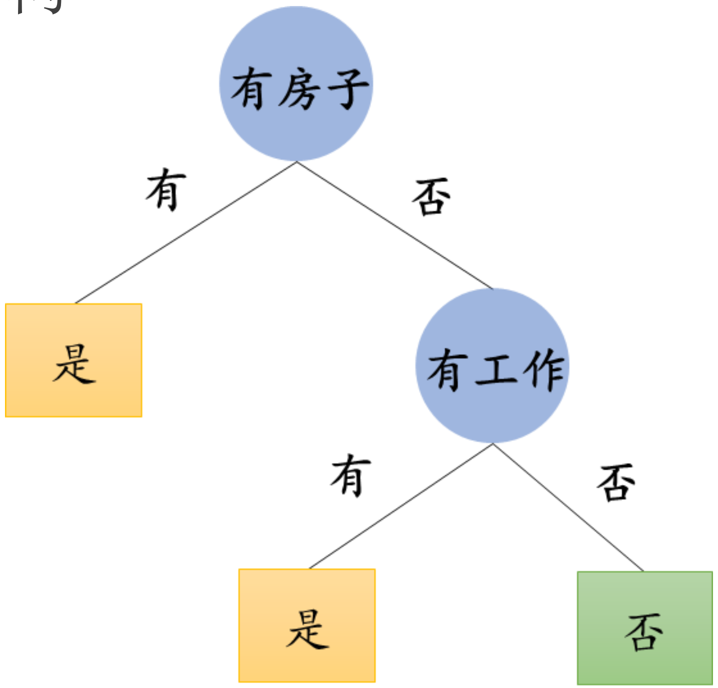
有向无环图+条件概率表

预测：全概率公式推导

推理：后验概率公式推导

决策树

- ◆分类决策树模型是一种描述对实例进行分类的树形结构
- ◆内部结点：特征或者属性
- ◆叶节点：类
- ◆关键：在当前状态下选择哪个属性作为分类依据
- ◆ID3算法——信息增益衡量节点“纯度”
  - 表示得知特征X的信息而使得类Y的信息的不确定性减少的程度



决策树

ID3算法

- 信息论中熵：随机变量不确定性
- $H(D) = -\sum_{i=1}^{k} p_i log p_i$
- $H(D|A) = \sum_{i=1}^{n} p_i H(D|A = x_i), p_i = P(A = x_i)$
- $g(D, A) = H(D) - H(D|A)$

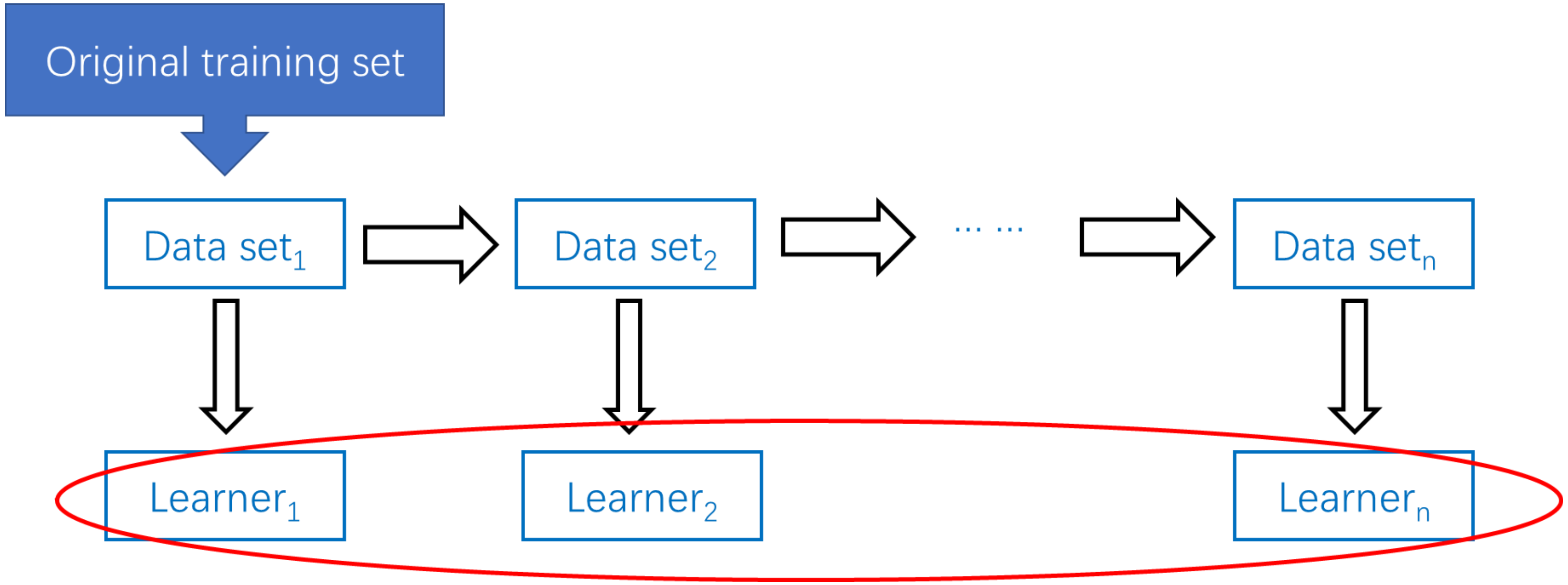
集成学习

- 集成学习通过构建并结合多个学习器来完成学习任务
- 个体学习器可以相同模型也可以不同模型，但必须“好而不同”
  - 好：性能好
  - 不同：结果具有不同的分布!!
- 集成学习经典的两种
  - Boosting: 学习器之间存在着强依赖关系，串行
  - Bagging：个体学习器之间不存在强依赖，并行

集成学习

◆Boosting策略

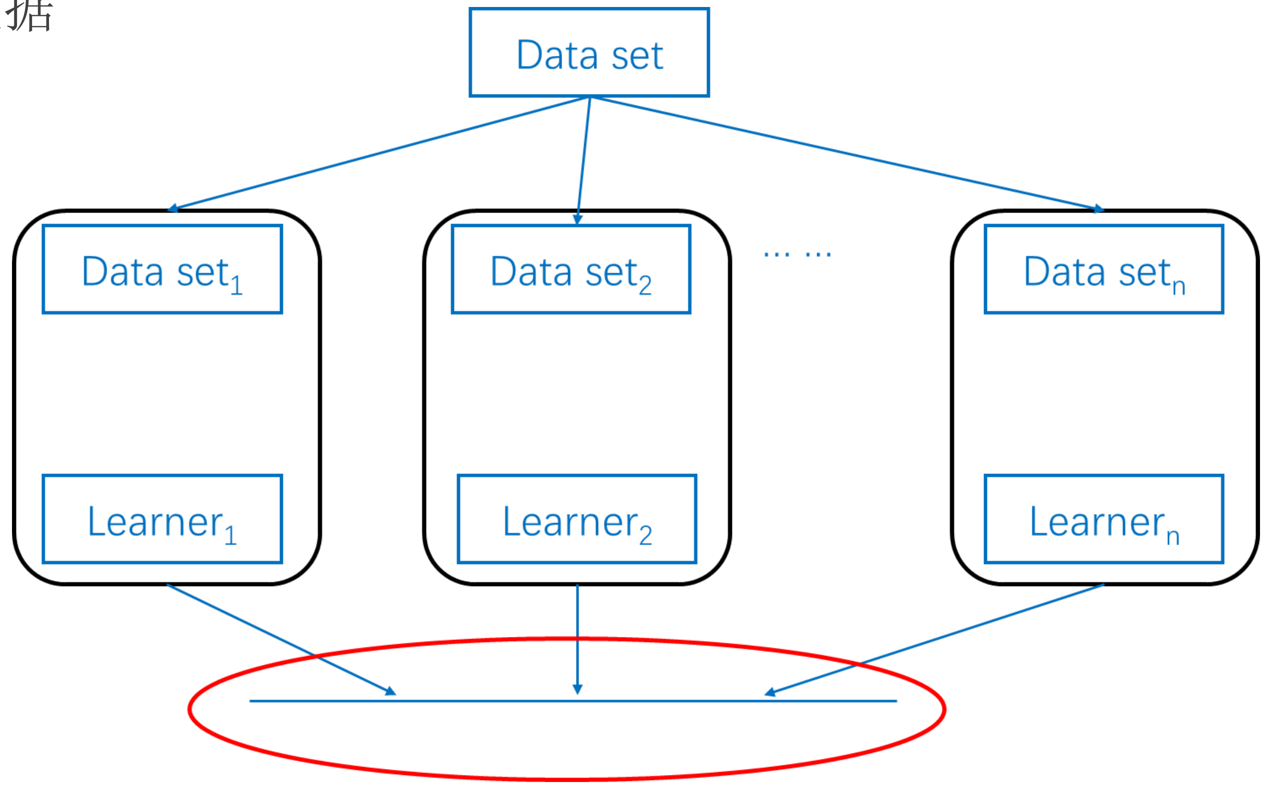
- 后一个学习器基于前一个学习器的结果，进行数据分布的调整
- 最后的结果基于学习器性能的好坏作为权重进行融合



集成学习

Bagging

- 每个学习器独立学习部分数据
- 平均或者投票

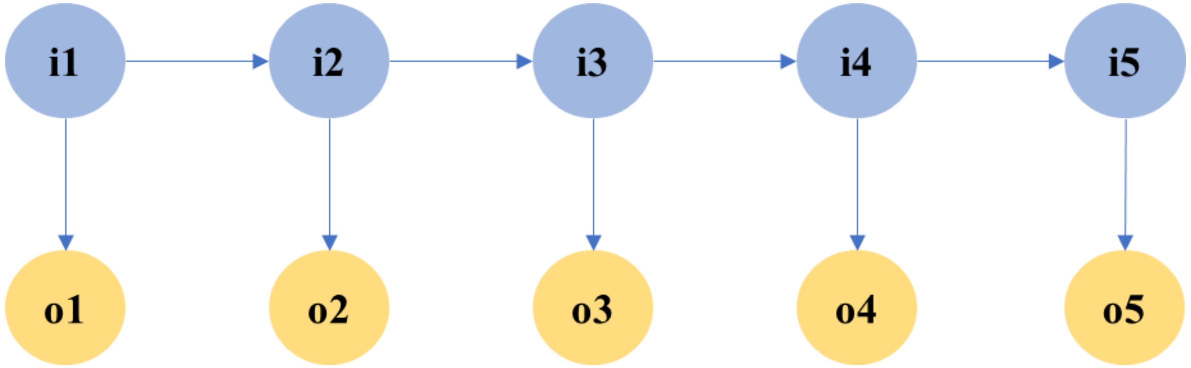


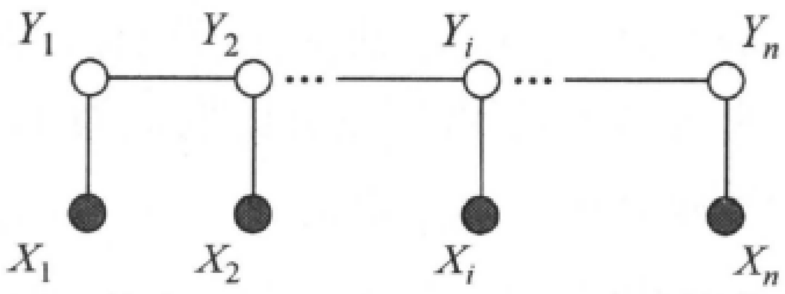
EM算法

- 一种求解参数的算法，无监督学习
- 含有隐变量情况，能够利用观测数据来估计隐变量和模型参数
- 迭代算法，在概率模型中寻找参数极大似然估计的算法
- **（E-step）如果参数已知，根据训练数据推断出最优隐变量的值**
  - $Q(\theta,\theta^i)=E_Z\left[\log P(Y,Z|\theta)\,|Y,\theta^i\right]=\sum_Z P(Z|Y,\theta^i)\log P(Y,Z|\theta)$
- **（M-step）如果隐变量已知，对参数做极大似然估计**
  - $\theta=\underset{\theta}{\operatorname{argmax}}\,Q(\theta,\theta^i)$

隐马尔科夫模型&条件随机场

- 时序模型
- 有向图vs无向图
- $P(Y_i|Y_{i-1})$ vs $P(Y_i|X, Y_{i-1}, Y_{i+1})$

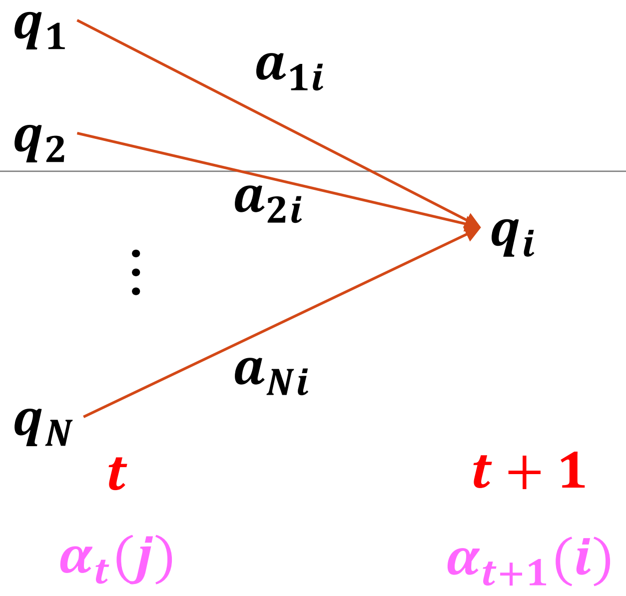




隐马尔科夫模型&条件随机场

- 隐马尔科夫
  - $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$
- 概率计算

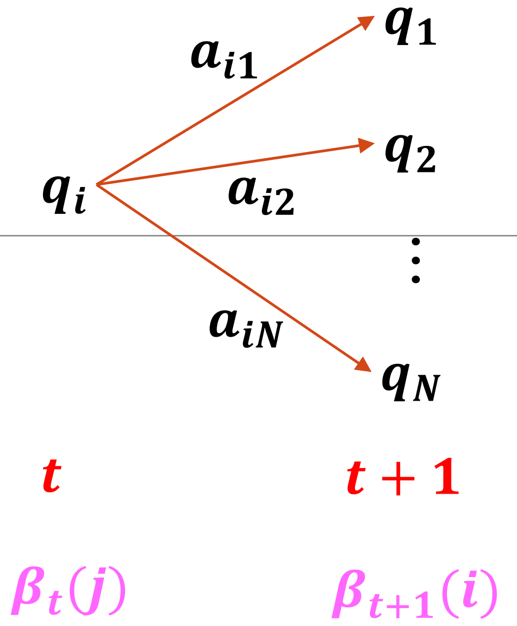
HMM——概率计算



◆前向/后向算法

- 前向概率：$\alpha_t(i) = P(o_1, o_2, \cdots, o_t, i_t = q_i|\lambda)$
  - 初始：$\alpha_1(i) = \pi_i b_i(o_1), i = 1, 2, \cdots, N$
  - 递推：$\alpha_{t+1}(i) = [\sum_{j=1}^{N} \alpha_t(j) a_{ji}] b_i(o_{t+1}), i = 1, 2, \cdots, N$
  - 终止：$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

HMM——概率计算



◆前向/后向算法

- 后向概率：$\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T | i_t = q_i, \lambda)$
  - 初始： $\boldsymbol{\beta_T(i) = 1, i = 1, 2, \cdots, N}$
  - 递推： $\boldsymbol{\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\ , i = 1, 2, \cdots, N}$
  - 终止： $\boldsymbol{P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)}$

HMM-维特比算法

◆思想：动态规划求概率最大路径

◆维特比算法

- 在时刻t状态为i的所有单个路径$(i_1, i_2, \cdots, i_t)$中概率最大值为：
  - $\delta_t(i) = \max_{i_1, i_2, \cdots, i_{t-1}} P(i_t = i, i_{t-1}, \cdots, i_1, o_t, \cdots, o_1 | \lambda), \quad i = 1, 2, \cdots, N$
  - $= \max_{1 \le j \le N} [\delta_{t-1}(j) a_{ji}] b_i(o_t), \; i = 1, \cdots, N; t = 1, \cdots, T-1$
- 定义在时刻t状态为i的所有单个路径中概率最大的路径的第t-1个结点为:
  - $\Psi_t(i) = \arg\max_{1 \le j \le N} [\delta_{t-1}(j) a_{ji}], \quad i = 1, 2, \cdots, N$

隐马尔科夫模型&条件随机场

- 条件随机场试图对多个变量在给定观测之后的条件概率进行建模
  - $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 观测序列
  - $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$ 为对应的标记序列
- 目标：构建条件概率$\boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x})$

- 参数化形式
  - $P(y|x) = \frac{1}{Z}\exp\left(\boxed{\sum_j}\sum_1^{n-1}\lambda_j t_j(y_{i+1}, y_i, x, i) + \boxed{\sum_k}\sum_{i=1}^{n}\mu_k s_k(y_i, x, i)\right)$
  - $Z = \sum_y \exp\left(\sum_j \sum_1^{n-1}\lambda_j t_j(y_{i+1}, y_i, x, i) + \sum_k \sum_{i=1}^{n}\mu_k s_k(y_i, x, i)\right)$

y	$\{y_1$	y_2	y_3	y_4	y_5	$y_6\}$
	D	**N**	**V**	**P**	**D**	**N**
x	$\{x_1$	x_2	x_3	x_4	x_5	$x_6\}$

$$t_j(y_{i+1}, y_i, x, i) = \begin{cases} 1, if\ y_{i+1} = P, y_i = V\ and\ \ x_i = "knock" \\ 0, otherwise \end{cases}$$

$$s_k(y_i, x, i) = \begin{cases} 1, if\ \ y_i = V\ and\ \ x_i = "knock" \\ 0, otherwise \end{cases}$$

隐马尔科夫模型&条件随机场

- 条件随机场试图对多个变量在给定观测之后的条件概率进行建模
  - $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 观测序列
  - $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$ 为对应的标记序列
- 目标：构建条件概率$\boldsymbol{P(y|x)}$
  - $\max_{w} P_w(y|x) = \frac{1}{Z_w}\exp(\sum_{i=1}^{n} w_i f_i(x,y))$

- **维特比算法**：求**非规范化**概率最大的最优路径问题$\max\limits_{y}(w \cdot F(y,x))$
- 输入：特征向量$F(y,x)$，权值向量w，观测序列：$x=(x_1,x_2,\cdots,x_n)$
- 输出：最优路径$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$
- 初始
  - $\delta_1(j)=w \cdot F_1(y_0=start,y_1=j,x),\ \ j=1,2,\cdots,m$
- 递推
  - $\delta_i(l)=\max\limits_{1 \le j \le m}\{\delta_{i-1}(j)+w \cdot F_i(y_{i-1}=j,y_i=l,x)\},\ \ l=1,2,\cdots,m$
  - $\psi_i(l)=\arg\max\limits_{1 \le j \le m}\{\delta_{i-1}(j)+w \cdot F_i(y_{i-1}=j,y_i=l,x)\},\ \ l=1,2,\cdots,m$

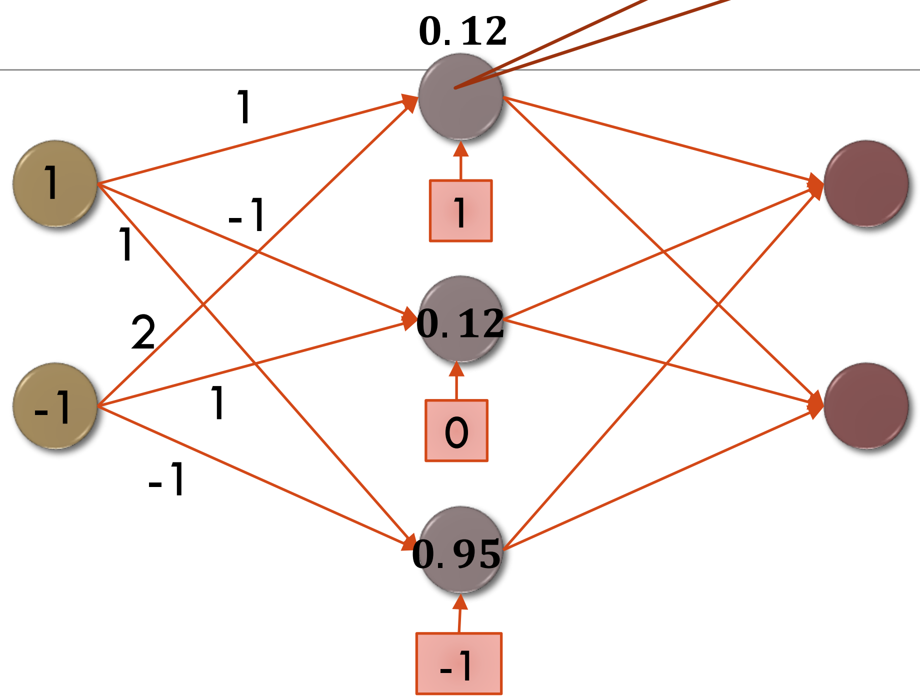
神经网络

- **全连接神经网络**
  - 通用
- **卷积神经网络**
  - 卷积层&池化
  - 适合图像视频数据
  - 减少参数（减少连接，共享参数）
- **循环神经网络**
  - 存储记忆单元
  - 适合时序数据

BP算法

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

◆前向传播



$$1 \times 1 + (-1) \times (-1) + 1 = 3$$

$$\frac{1}{1+e^{-3}} = 0.95$$

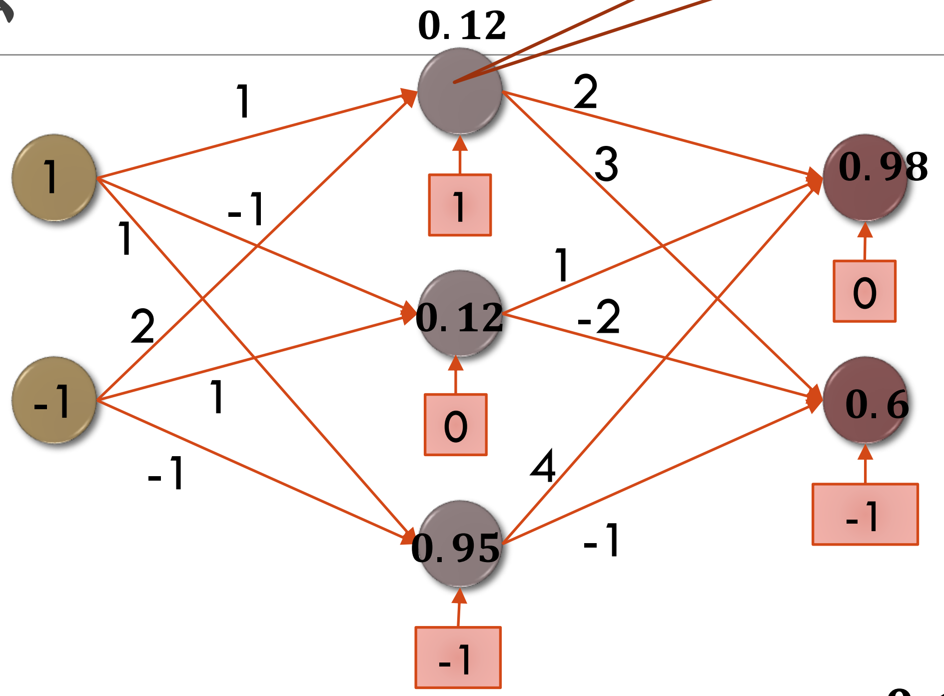
$$1 \times 1 + (-1) \times 2 - 1 = -2 \qquad 1 \times (-1) + (-1) \times 1 - 0 = -2$$

$$\frac{1}{1+e^{2}} = 0.12 \qquad \frac{1}{1+e^{2}} = 0.12$$

BP算法

- 前向传播

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



$$0.12 \times 2 + 0.12 \times 1 + 0.95 \times 4 - 0 = 4.16$$

$$\frac{1}{1+e^{-4.16}} = 0.98$$

$$0.12 \times 3 + 0.12 \times (-2) + 0.73 \times (-1) + 1 = 0.39$$

$$\frac{1}{1+e^{-0.39}} = 0.6$$

BP算法

上一层

下一层

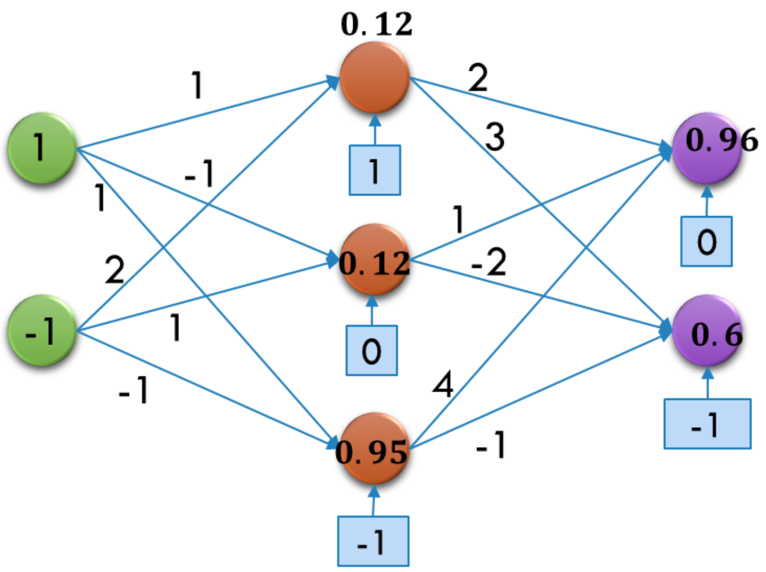
- 设$\boldsymbol{y_1} = \mathbf{1}, \boldsymbol{y_2} = \mathbf{1}, \boldsymbol{\eta} = \mathbf{1}$
  - $\Delta w_{ij} = \delta_i h_j$
  - $\delta_i = (\hat{y}_i - y_i) \cdot f_i'$

损失函数

激活函数

  - $\Delta\theta_i = -\delta_i$
  - $\delta_1 = (\hat{y}_1 - y_1) \cdot f_1' \quad = (0.96 - 1) \times 0.96 \times (1 - 0.96)$

    $= 0.001536$
  - $\Delta w_{11} = \delta_1 h_1 = 0.001536 \times 0.12 = 0.00018432$
  - $w_{11} = 2 -$ 步长 $\times 0.00018432 = 1.9998$

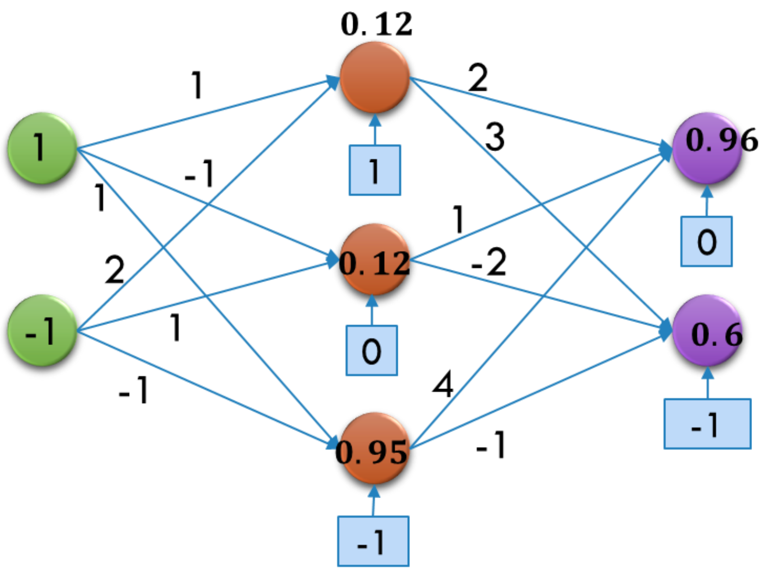


BP算法

- $\Delta w_{ij}^{k} = \delta_i^k h_j^{k-1}$
- $\delta_i^k = (\hat{h}_i - h_i) \cdot f_i'$
- $\delta_i^k = \left(\sum_l \delta_l^{k+1} w_{li}^{k+1}\right) \cdot f_i'$
- $\Delta\theta_i = -\delta_i$
- $\delta_1^2 = \left(\sum_{l=1}^{2} \delta_l^3 w_{l1}^3\right) \cdot f_1'$

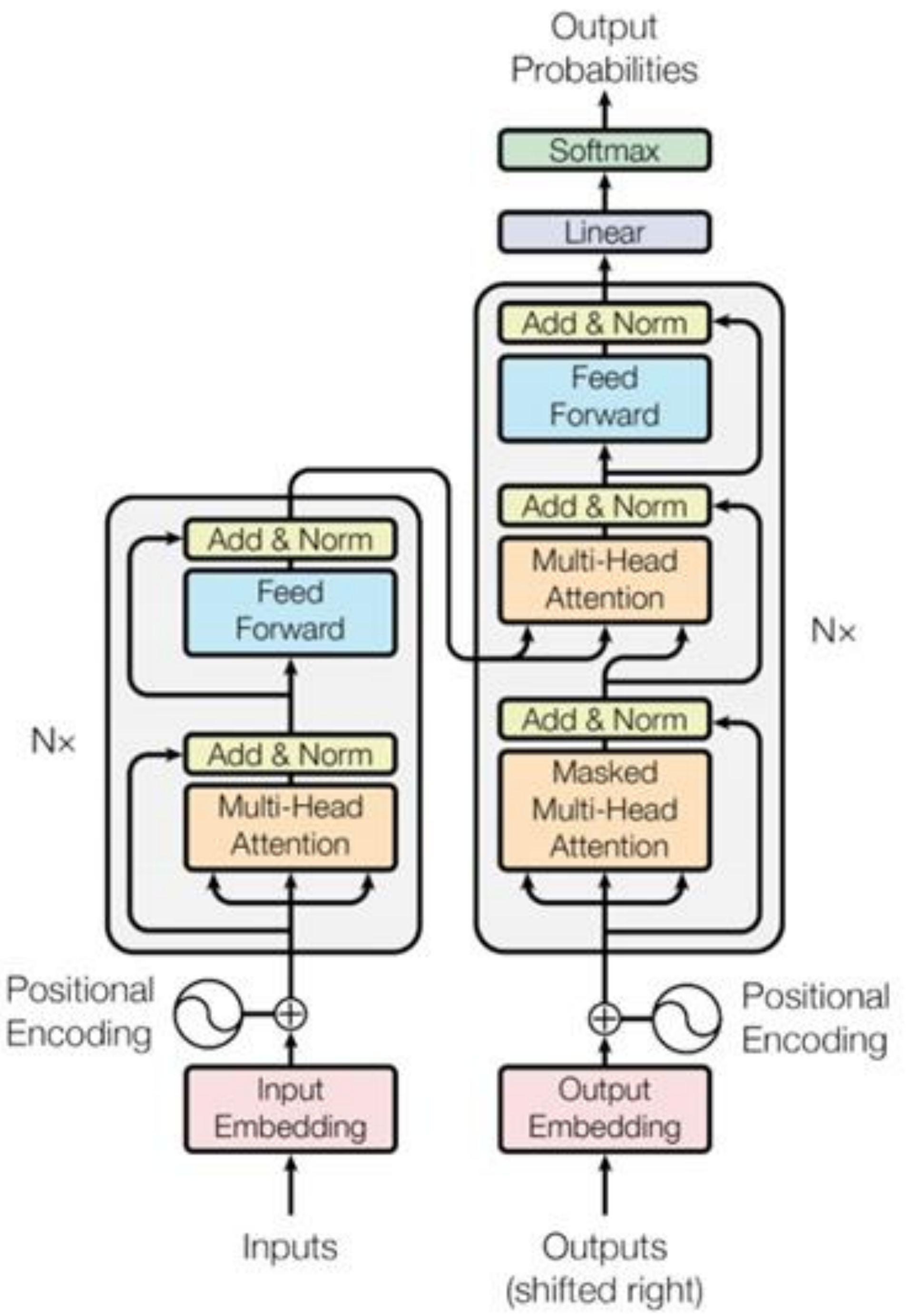
  $= [(0.96 \times 2) + (0.6 \times 3)] \times 0.12 \times (1 - 0.12)$

  $= 0.3928$
- $\Delta w_{11}^2 = \delta_1^2 h_1^1 = 0.3928 \times 1 = 0.3928$
- $w_{11}^2 = 1 - \text{步长} \times 0.3928 = 0.6072$

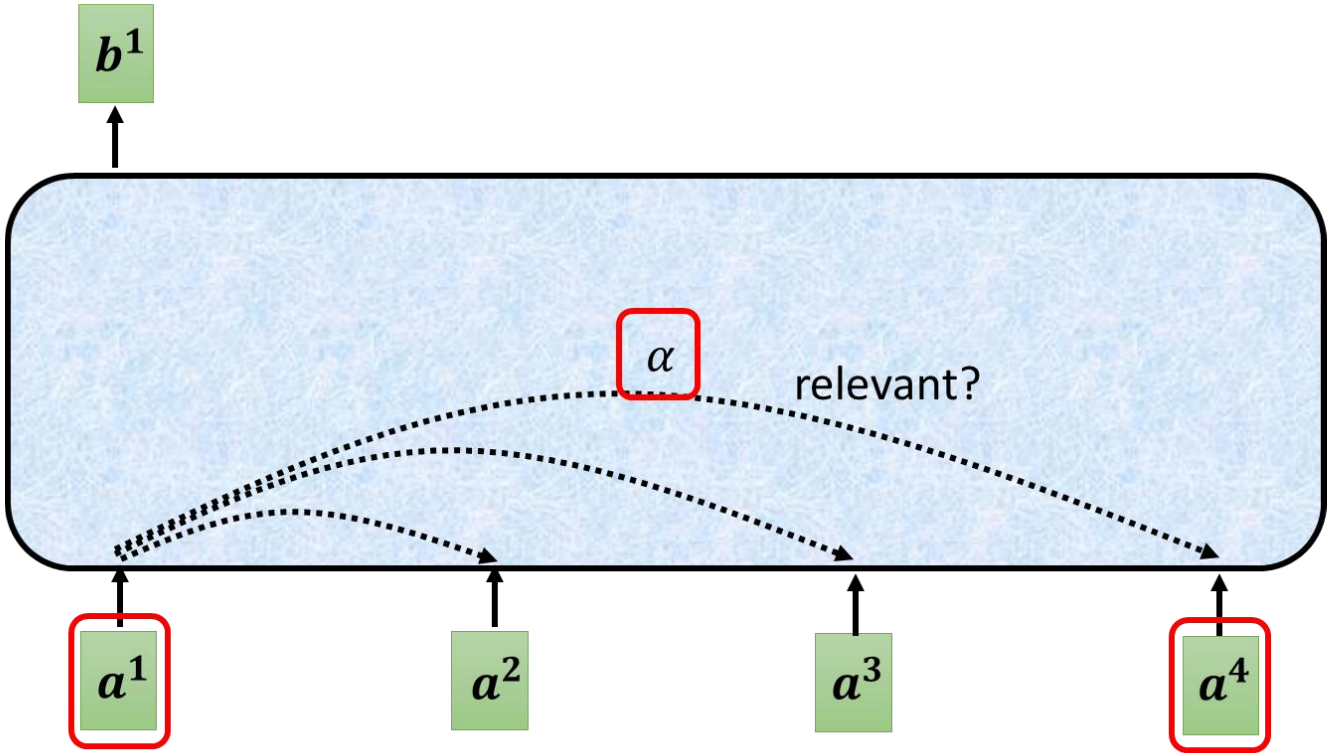


Transformer

- 自注意力机制：捕获长距离信息
  - 多头：多个角度
- 残差连接
  - 稳定训练，避免退化问题
- 编码解码层之间的交互
  - 交叉注意力



自注意力机制



处理时序数据，考虑数据之间的联系

自注意力机制

$$Q = W^q I$$

$$K = W^k I$$

$$V = W^v I$$

$$A' \leftarrow A = K^T Q$$

$$O = V A'$$

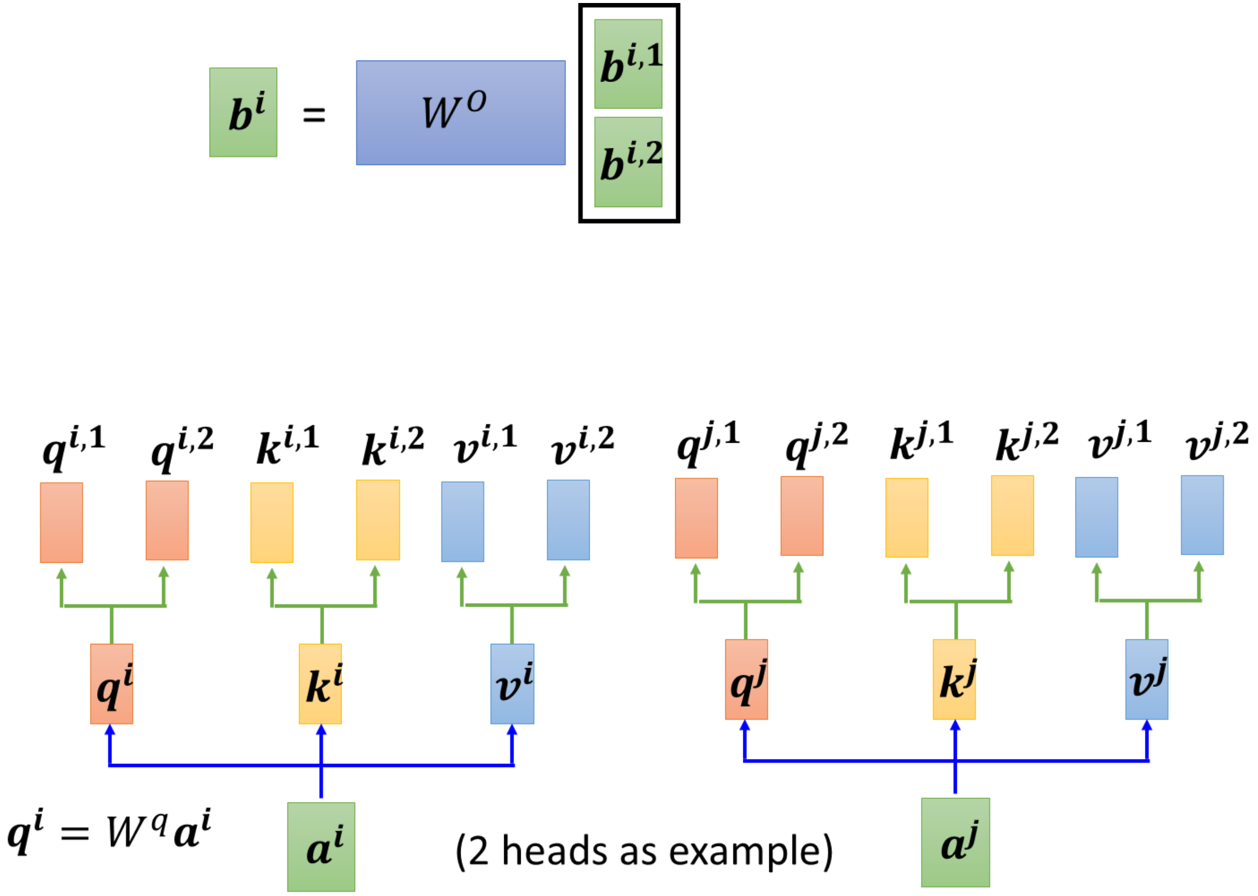
处理时序数据，考虑数据之间的联系

通过查询Q和键K计算内容之间的关联得分

基于关联得分，从值V中获取信息

并行操作

多头注意力机制



处理时序数据，考虑数据之间的联系

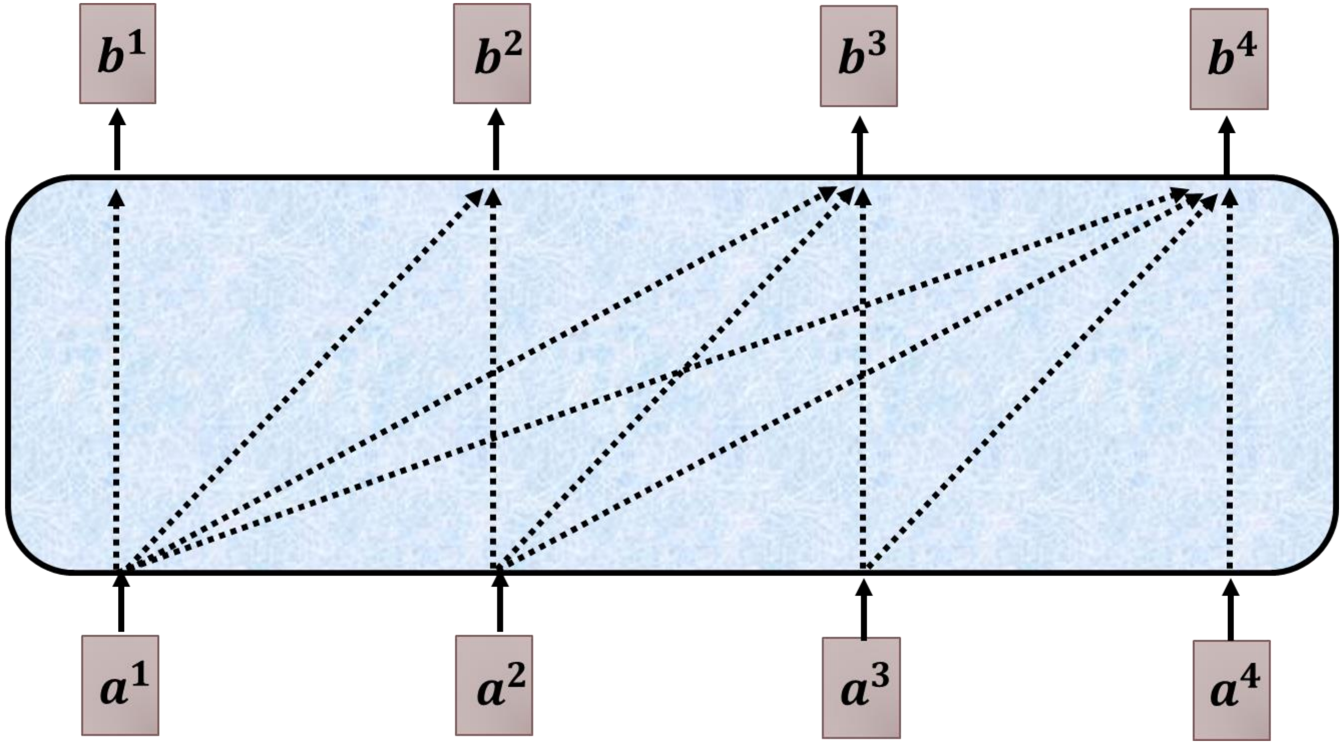
通过查询Q和键K计算内容之间的关联得分

基于关联得分，从值V中获取信息

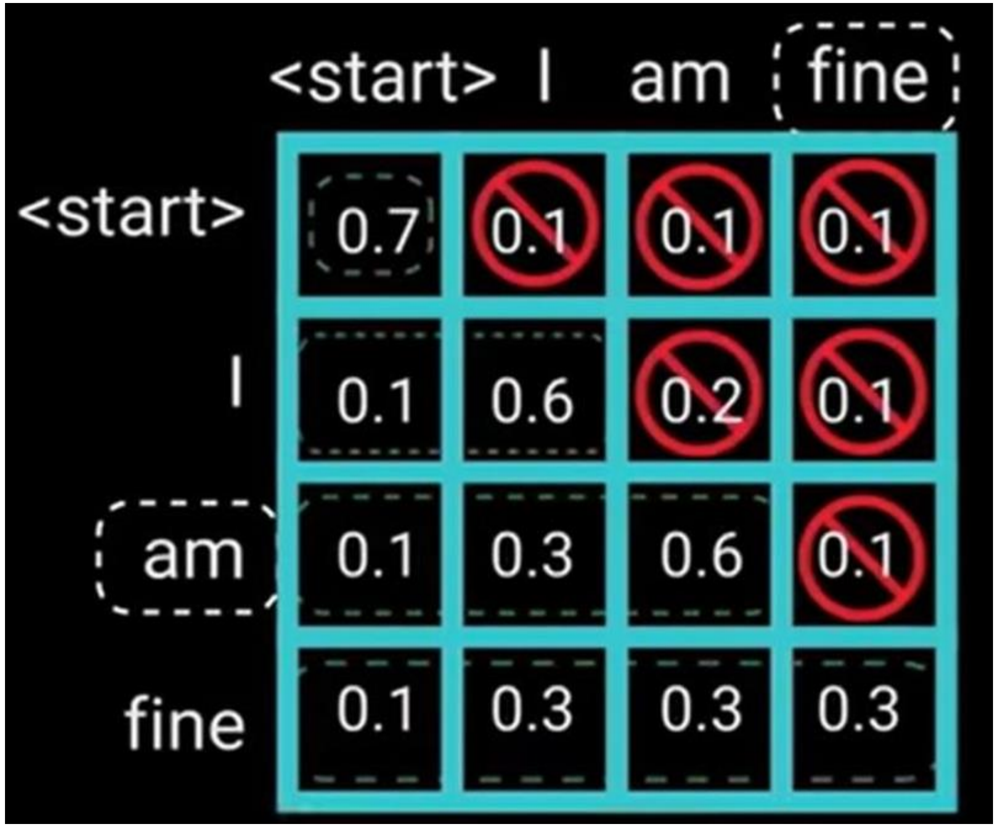
并行操作

多头机制，每个头独立计算，增强模型表达

掩码机制

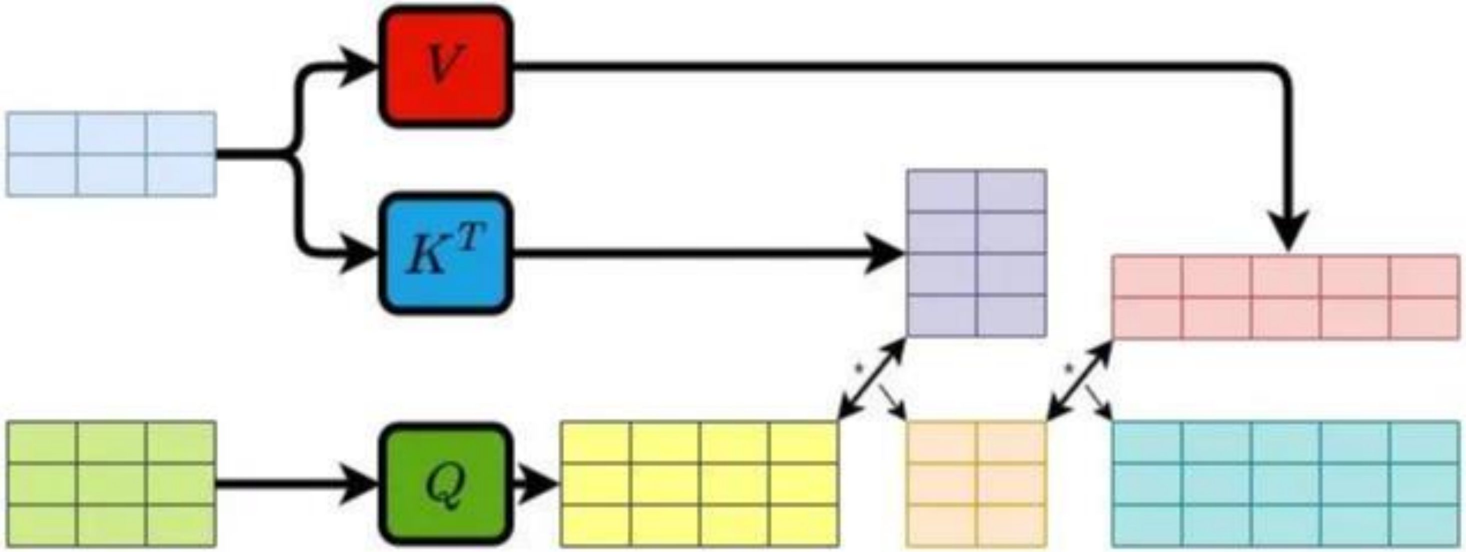


b^1
b^2
b^3
b^4
a^1
a^2
a^3
a^4



<start> I am fine
<start> 0.7 0.1 0.1 0.1
I 0.1 0.6 0.2 0.1
am 0.1 0.3 0.6 0.1
fine 0.1 0.3 0.3 0.3

交叉注意力机制



附录——作业

◆1.对下面分类模型结果进行评估，计算精度，召回率，和F1值

图片识别任务，识别图片是否是一只小狗。假设有100张图片，60张是小狗，40张是其他图像。有位同学构建图片识别模型，得到如下结果：**模型查找出了50张小狗图片，但其中只有40张是真正的小狗图片**，请对其模型进行评估

真实情况	预测结果	
	正例	负例
正例	40	20
负例	10	30

$$精确率p = \frac{40}{50} = 80\%$$

$$召回率r = \frac{40}{60} = 66.7\%$$

$$F1=\frac{2*0.8*0.67}{0.8+0.67} = 72.9\%$$

附录——作业

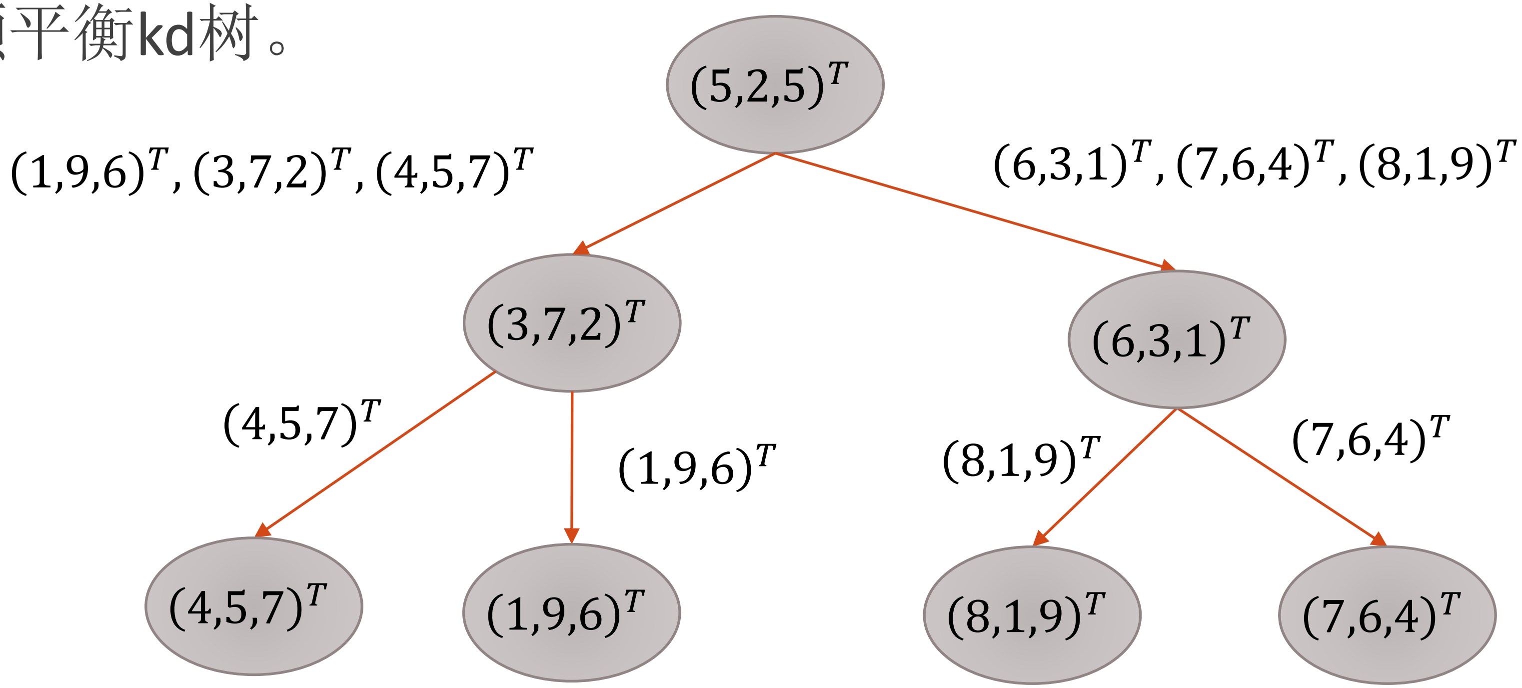
- 2.已知训练数据集，正实例点是$x_1=(3,3)^T$, $x_2=(4,2)^T$, 负实例点是$x_3=(1,1)^T$, $x_4=(2,0)^T$，用原始形式求出感知机模型决策函数$f=\text{sign}(wx+b)$，按照例题形式给出过程

$\eta=1$

x_1	$(3,3)$	1	$3x_1+3x_2+1$
x_3	$(2,2)$	0	$2x_1+2x_2+0$
x_3	$(1,1)$	-1	x_1+x_2-1
x_3	$(0,0)$	-2	-2
x_1	$(3,3)$	-1	$3x_1+3x_2-1$
x_3	$(2,2)$	-2	$2x_1+2x_2-2$
x_3	$(1,1)$	-3	x_1+x_2-3

附录——作业

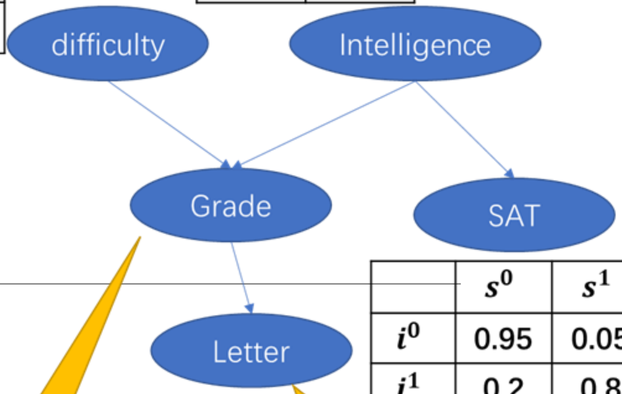
◆ 3.给定一个三维空间的数据集：*T*={(1,9,6), (4,5,7), (8,1,9), (3,7,2), (7,6,4), (5,2,5), (6,3,1)}构造一颗平衡kd树。



附录——作业4

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3



	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- （1）学生George有多大可能从课程Econ101教授那里获得一封好的推荐信($\boldsymbol{l^1}$)

$$\boldsymbol{p(l^1)} = p(l^1|g^1)\boxed{p(g^1)} + p(l^1|g^2)p(g^2) + p(l^1|g^3)p(g^3) = \mathbf{0.5023}$$

$$p(g^1) = p(g^1|i^0,d^0)p(i^0)p(d^0) + p(g^1|i^0,d^1)p(i^0)p(d^1)$$
$$+ p(g^1|i^1,d^0)p(i^1)p(d^0) + p(g^1|i^1,d^1)p(i^1)p(d^1) = 0.3620$$

$$p(g^2) = 0.2884 \qquad p(g^3) = 0.3496$$

- （2）若George不聪明($\boldsymbol{i^0}$)，概率会降到多少

$$\boldsymbol{p(l^1|i^0)} = \frac{p(l^1,i^0)}{p(i^0)} = \frac{p(l^1|g^1)\boxed{p(g^1|i^0)}p(i^0) + p(l^1|g^2)p(g^2|i^0)p(i^0) + p(l^1|g^3)p(g^3|i^0)p(i^0)}{p(i^0)} = \mathbf{0.3886}$$

$$p(g^1|i^0) = p(g^1|i^0,d^0)p(d^0) + p(g^1|i^0,d^1)p(d^1)$$

- （3）George去招聘，招聘官相信George有30%的高智商，但看到Econ101的课程得了$\boldsymbol{g^3}$，判断George具有高智商的概率

$$p(g^3|i^1) = p(g^3|i^1,d^0)p(d^0) + p(g^3|i^1,d^1)p(d^1)$$

$$\boldsymbol{p(i^1|g^3)} = \frac{p(g^3,i^1)}{p(g^3)} = \frac{\boxed{p(g^3|i^1)}p(i^1)}{p(g^3)} = \mathbf{0.0789}$$

附录——作业

- 5. 利用ID3算法构建一颗决策树

$$H(D) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0.9403$$

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
属性1	A	A	A	A	A	B	B	B	B	C	C	C	C	C
属性2	真	真	假	假	假	真	假	真	假	真	真	假	假	假
类	1	2	2	2	1	1	1	1	1	2	2	1	1	1

$$G(D,1) = 0.9403 - 0.6936 = 0.2467$$

$$G(D,2) = 0.9403 - 0.8922 = 0.0481$$

$$H(D|1) = -\frac{5}{14}\left[\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right] - \frac{4}{14}\left[\frac{4}{4}\log\frac{4}{4}\right] - \frac{5}{14}\left[\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right] = 0.3468 + 0 + 0.3468 = 0.6936$$

$$H(D|2) = -\frac{6}{14}\left[\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right] - \frac{8}{14}\left[\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}\right] = 0.4286 + 0.4636 = 0.8922$$

附录——作业

- 6. 利用隐马尔科夫模型进行句子标注工作，假设词性共有：代词、动词、名词和介词，词汇表中共有7个词语[苏州大学，开创，坐落，教育，于，江苏省，苏州市]。经过统计，可得如下参数数据：每个词性状态出现在第一位的概率为(0.3,0.2,0.3,0.2), 词性转移概率表如下（横向看，即代词向其他词转移概率为[0.3,0.25,0.25,0.2]）：

	代词	动词	名词	介词
代词	0.3	0.25	0.25	0.2
动词	0.16	0.12	0.28	0.44
名词	0.14	0.43	0.27	0.16
介词	0.2	0.2	0.5	0.1

	苏州大学	开创	坐落	教育	于	江苏省	苏州市
代词	0.3	0.1	0.1	0.1	0.1	0.1	0.2
动词	0.1	0.2	0.3	0.1	0.1	0.1	0.1
名词	0.2	0.1	0.05	0.2	0.05	0.2	0.2
介词	0.05	0.05	0.05	0.05	0.7	0.05	0.05

- 请对“苏州大学/坐落/于/苏州市”进行词性标注

苏州大学/坐落/于/苏州市

	代词	动词	名词	介词
代词	0.3	0.25	0.25	0.2
动词	0.16	0.12	0.28	0.44
名词	0.14	0.43	0.27	0.16
介词	0.2	0.2	0.5	0.1

	苏州大学	开创	坐落	教育	于	江苏省	苏州市
代词	0.3	0.1	0.1	0.1	0.1	0.1	0.2
动词	0.1	0.2	0.3	0.1	0.1	0.1	0.1
名词	0.2	0.1	0.05	0.2	0.05	0.2	0.2
介词	0.05	0.05	0.05	0.05	0.7	0.05	0.05

t=1 o1=苏州大学

$\delta_1(1) = 0.3 \times 0.3 = 0.09$

$\delta_1(2) = 0.2 \times 0.1 = 0.02$

$\delta_1(3) = 0.3 \times 0.2 = 0.06$

$\delta_1(4) = 0.2 \times 0.05 = 0.01$

t=2 o2=坐落

$\delta_2(1) = \max(0.09 \times 0.3, 0.02 \times 0.16, 0.06 \times 0.14, 0.01 \times 0.2) \times 0.1 =$

$\max(0.027, 0.0032, 0.0084, 0.002) \times 0.1 = 0.0027$ $\quad \boldsymbol{\Psi_2(1) = 1}$

$\delta_2(2) = \max(0.09 \times 0.25, 0.02 \times 0.12, 0.06 \times 0.43, 0.01 \times 0.2) \times 0.3$

$= \max(0.0225, 0.0024, 0.0258, 0.002) \times 0.3 = 0.00774$ $\quad \boldsymbol{\Psi_2(2) = 3}$

$\delta_2(3) = \max(0.0225, 0.0056, 0.0162, 0.005) \times 0.05 = 0.001125$ $\quad \boldsymbol{\Psi_2(3) = 1}$

$\delta_2(4) = \max(0.018, 0.0088, 0.0096, 0.001) \times 0.05 = 0.0009$ $\quad \boldsymbol{\Psi_2(4) = 1}$

苏州大学/坐落/于/苏州市

	代词	动词	名词	介词
代词	0.3	0.25	0.25	0.2
动词	0.16	0.12	0.28	0.44
名词	0.14	0.43	0.27	0.16
介词	0.2	0.2	0.5	0.1

	苏州大学	开创	坐落	教育	于	江苏省	苏州市
代词	0.3	0.1	0.1	0.1	0.1	0.1	0.2
动词	0.1	0.2	0.3	0.1	0.1	0.1	0.1
名词	0.2	0.1	0.05	0.2	0.05	0.2	0.2
介词	0.05	0.05	0.05	0.05	0.7	0.05	0.05

t=3 o3=于

$\delta_3(1)$=max(8.1e−4, 12.384e−4, 1.575e−4, 1.8e−4)∗0.1 =12.384e−5 $\quad \boldsymbol{\Psi_3(1)=2}$

$\delta_3(2)$ =max(6.75e-4, 9.288e-4, 4.8375e-4, 1.8e-4)*0.1 = 9.288e-5 $\quad \boldsymbol{\Psi_3(2)=2}$

$\delta_3(3)$ =max(6.75e-4, 21.672e-4, 3.0375e-4, 4.5e-4)*0.05 = 10.836e-5 $\quad \boldsymbol{\Psi_3(3)=2}$

$\delta_3(4)$ =max(5.4e-4, 34.056e-4, 1.8e-4, 0.9e-4)*0.7 = 238.392e-5 $\quad \boldsymbol{\Psi_3(4)=2}$

苏州大学/坐落/于/苏州市

	代词	动词	名词	介词
代词	0.3	0.25	0.25	0.2
动词	0.16	0.12	0.28	0.44
名词	0.14	0.43	0.27	0.16
介词	0.2	0.2	0.5	0.1

	苏州大学	开创	坐落	教育	于	江苏省	苏州市
代词	0.3	0.1	0.1	0.1	0.1	0.1	0.2
动词	0.1	0.2	0.3	0.1	0.1	0.1	0.1
名词	0.2	0.1	0.05	0.2	0.05	0.2	0.2
介词	0.05	0.05	0.05	0.05	0.7	0.05	0.05

t=4 o4=苏州市

$\delta_4(1)$ =max(3.7152e-5, 1.48608e-5, 1.51704e-5, 4.76784e-4)*0.2 = 9.53568e-5 $\boldsymbol{\Psi_4(1)=4}$

$\delta_4(2)$ =max(3.096e-5, 1.11456e-5, 4.65948e-5, 4.76784e-4)*0.1 = 4.76784e-5 $\boldsymbol{\Psi_4(2)=4}$

$\delta_4(3)$ =max(3.096e-5, 2.60064e-5, 2.92572e-5, 11.9196e-4)*0.2 = $\boxed{23.8392\text{e-}5}$ $\boldsymbol{\Psi_4(3)=4}$

$\delta_4(4)$ =max(2.4768e-5, 4.08672e-5, 1.73376e-5, 2.38392e-4)*0.05 = 1.19196e-5 $\boldsymbol{\Psi_4(4)=4}$

3-2-4-1: 名词，动词，介词，名词

附录——作业

◆7.利用条件随机场进行词性标注。假设词性有名词n，动词v，代词p，定义特征函数及相应权重如下：

	函数条件	权重
t1	=1 （$y_{t-1}=n$，$y_t=v$） =0 其它	0.6
t2	=1 （$y_{t-1}=p$，$y_t=n$） =0 其它	0.8
t3	=1 （$y_{t-1}=v$，$y_t=n$） =0 其它	0.5
s1	=1 （$y_t=n$，$x_t=$ 人名） =0 其它	0.9
s2	=1 （$y_t=n$，$x_t=$ 地点） =0 其它	0.9
s3	=1 （$y_t=p$，$x_t=at$） =0 其它	0.7

◆请对“Bob drank coffee at Starbucks”进行词性标注（给出具体计算过程）

		start	n	v	p
Bob **n**	n	0.9	—	—	—
	v	0	—	—	—
	p	0	—	—	—
drank **v**	n	—	0.9	0+0.5=0.5	0+0.8=0.8
	v	—	0.9+0.6=1.5	0	0
	p	—	0.9	0	0
coffee **n**	n	—	0.9	1.5+0.5=2	0.9+0.8=1.7
	v	—	0.9+0.6=1.5	1.5	0.9
	p	—	0.9	1.5	0.9
at **p**	n	—	2	1.5+0.5=2	1.5+0.8=2.3
	v	—	2+0.6=2.6	1.5	1.5
	p	—	2+0.7=2.7	1.5+0.7=2.2	1.5+0.7=2.2
starbucks **n**	n	—	2.3+0.9=3.2	2.6+0.5+0.9 =4	$\boxed{2.7+0.8+0.9 =4.4}$
	v	—	2.3+0.6=2.9	2.6	2.7
	p	—	2.3	2.6	2.7

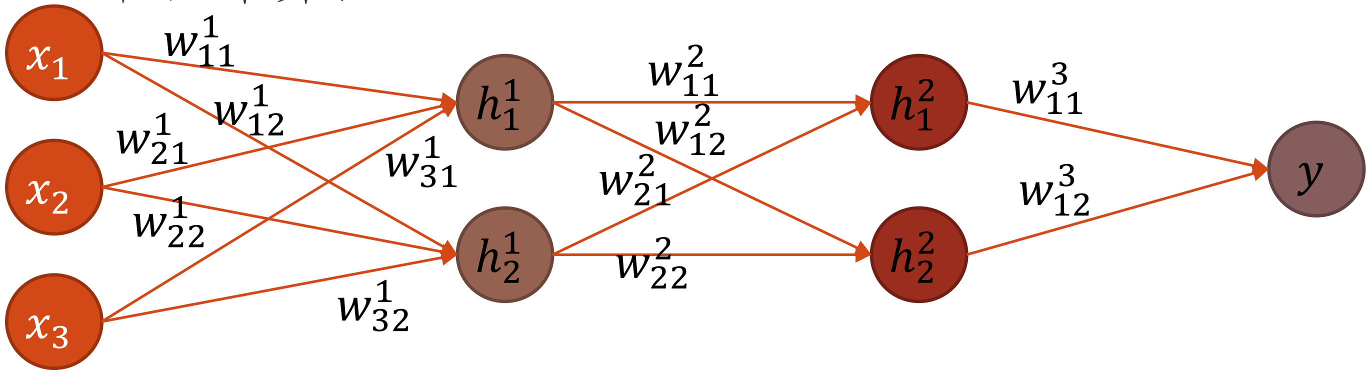
	函数条件	权重
t1	=1 （$y_{t-1}=n$, $y_t=v$） =0 其它	0.6
t2	=1 （$y_{t-1}=p$, $y_t=n$） =0 其它	0.8
t3	=1 （$y_{t-1}=v$, $y_t=n$） =0 其它	0.5
s1	=1 （ $y_t=n$, $x_t=$ 人名 ） =0 其它	0.9
s2	=1 （ $y_t=n$, $x_t=$ 地点 ） =0 其它	0.9
s3	=1 （$y_t=p$, $x_t=at$） =0 其它	0.7

Bob drank coffee at Starbucks

$$\delta_i(l) = \max_{1\le j\le m}\{\delta_{i-1}(j) + w\cdot F_i(y_{i-1}=j, y_i=l, x)\}$$

附录——作业

8.如下的BP神经网络，学习步长$\eta=1$，各点的阈值$\theta=0$, 输入层到隐层的激活函数为$f(x)=\max(1,x)$, 隐层到隐层以及隐层到输出层的激活函数为$f(x)=\frac{1}{1+e^{-x}}$，设输入样本$x_1=1, x_2=0$，$x_3=1$，输出节点的期望输出$y=0.98$，利用预测误差$E=\frac{1}{2}(\hat{y}-y)^2$对**连接权进行调整**（只调整一轮, 阈值更新不计算）

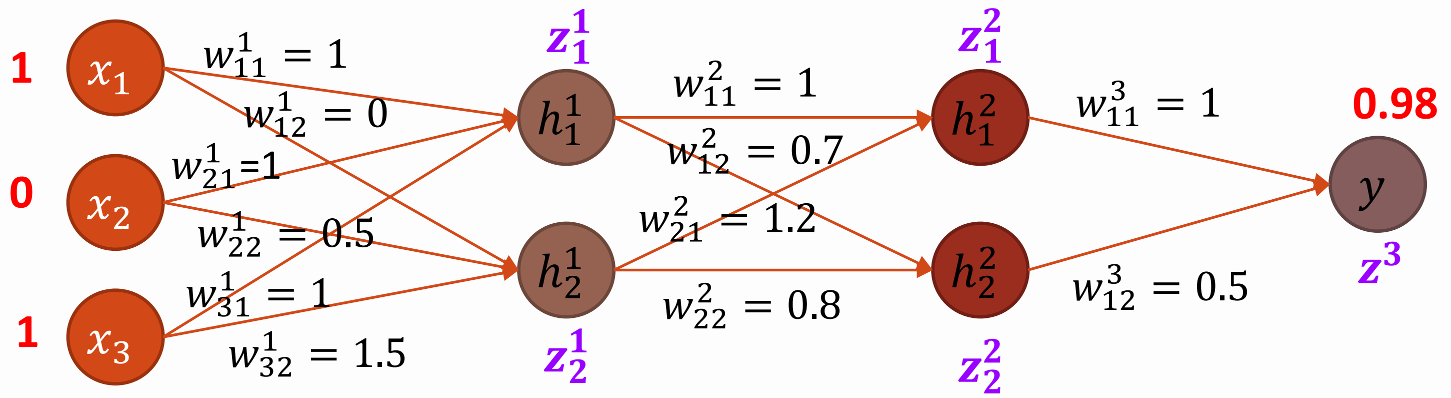


$w_{11}^1=1$
$w_{12}^1=0$
$w_{21}^1=1$
$w_{22}^1=0.5$
$w_{31}^1=1$
$w_{32}^1=1.5$

$w_{11}^2=1$
$w_{12}^2=0.7$
$w_{21}^2=1.2$
$w_{22}^2=0.8$

$w_{11}^3=1$
$w_{12}^3=0.5$

附录——作业



$w_{11}^1 = 1$
$w_{12}^1 = 0$
$w_{21}^1 = 1$
$w_{22}^1 = 0.5$
$w_{31}^1 = 1$
$w_{32}^1 = 1.5$

$w_{11}^2 = 1$
$w_{12}^2 = 0.7$
$w_{21}^2 = 1.2$
$w_{22}^2 = 0.8$

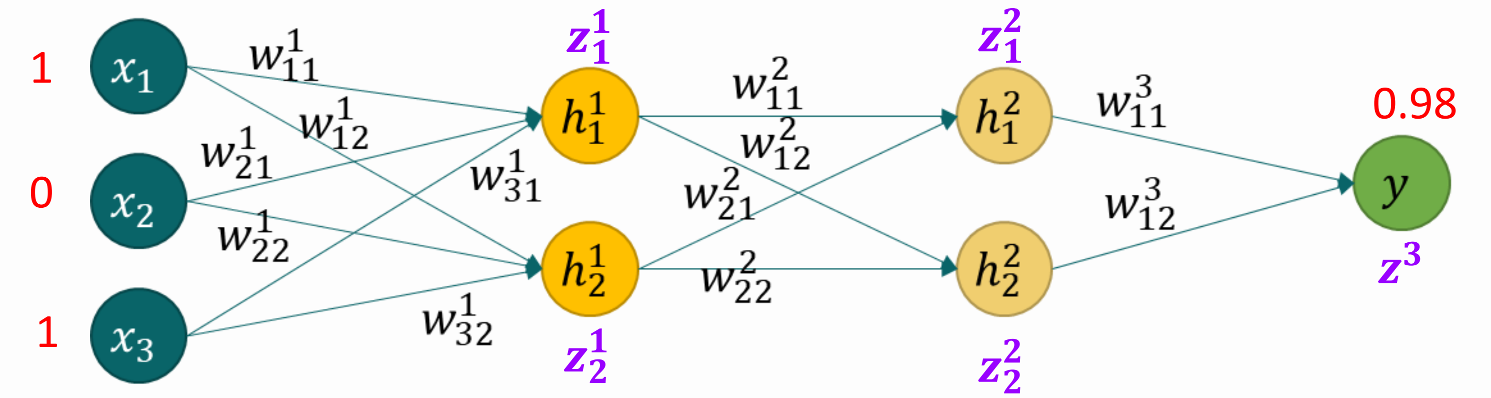
$w_{11}^3 = 1$
$w_{12}^3 = 0.5$

- 前向传播
  - $z_1^1 = 1 * 1 + 0 * 1 + 1 = 2, \quad z_2^1 = 1 * 0 + 0 * 0.5 + 1 * 1.5 = 1.5$
  - $h_1^1 = \max(1,2) = 2, \quad h_2^1 = \max(1,1.5) = 1.5$
  - $z_1^2 = 2 * 1 + 1.5 * 1.2 = 3.8, \quad z_2^2 = 2 * 0.7 + 1.5 * 0.8 = 2.6$
  - $h_1^1 = \frac{1}{1+e^{-3.8}} = 0.9781, \quad h_2^1 = \frac{1}{1+e^{-2.6}} = 0.9309$

$z^3 = 0.9781 * 1 + 0.9309 * 0.5 = 1.4436$

$\hat{y} = \frac{1}{1 + e^{-1.4436}} = 0.809$

附录——作业



$w_{11}^1 = 1$
$w_{12}^1 = 0$
$w_{21}^1 = 1$
$w_{22}^1 = 0.5$
$w_{31}^1 = 1$
$w_{32}^1 = 1.5$

$w_{11}^2 = 1$
$w_{12}^2 = 0.7$
$w_{21}^2 = 1.2$
$w_{22}^2 = 0.8$

$w_{11}^3 = 1$
$w_{12}^3 = 0.5$

- 反向传播

$$f = \frac{1}{1+e^{-z}}$$

$$f \cdot (1-f)$$

$$\delta_1^3 = (\hat{y} - y) * \boxed{f'(z^3)} = (0.809 - 0.98) * \boxed{0.809 * (1 - 0.809)} = -0.0264$$

- $\delta_j^L = \frac{\partial L}{\partial z_j^L} = \frac{\partial L}{\partial y_j}\frac{\partial y_j}{\partial z_j^L}$

$$\delta_1^2 = \delta_1^3 * w_{11}^3 * f'(z_1^2) = -0.0264 * 1 * 0.9781 * (1 - 0.9781) = -0.0006$$

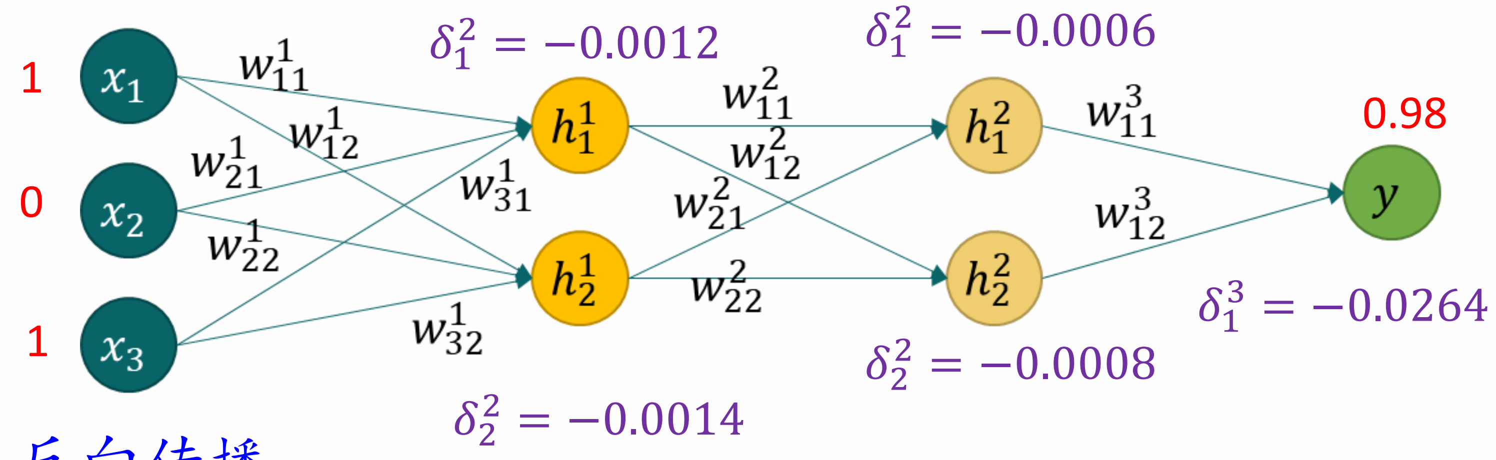
$$\delta_2^2 = \delta_1^3 * w_{12}^3 * f'(z_2^2) = -0.0264 * 0.5 * 0.9309 * (1 - 0.9309) = -0.0008$$

- $\delta_j^l = \frac{\partial L}{\partial z_j^l}$

$$\delta_1^2 = (\delta_1^2 * w_{11}^2 + \delta_2^2 * w_{12}^2) * f'(z_1^1) = (-0.0006 * 1 - 0.0008 * 0.7) * 1 = -0.0012$$

$$\delta_2^2 = (\delta_1^2 * w_{21}^2 + \delta_2^2 * w_{22}^2) * f'(z_2^1) = (-0.0006 * 1.2 - 0.0008 * 0.8) * 1 = -0.0014$$

附录——作业



$$w_{11}^1 = 1$$
$$w_{12}^1 = 0$$
$$w_{21}^1 = 1$$
$$w_{22}^1 = 0.5$$
$$w_{31}^1 = 1$$
$$w_{32}^1 = 1.5$$

$$w_{11}^2 = 1$$
$$w_{12}^2 = 0.7$$
$$w_{21}^2 = 1.2$$
$$w_{22}^2 = 0.8$$

$$w_{11}^3 = 1$$
$$w_{12}^3 = 0.5$$

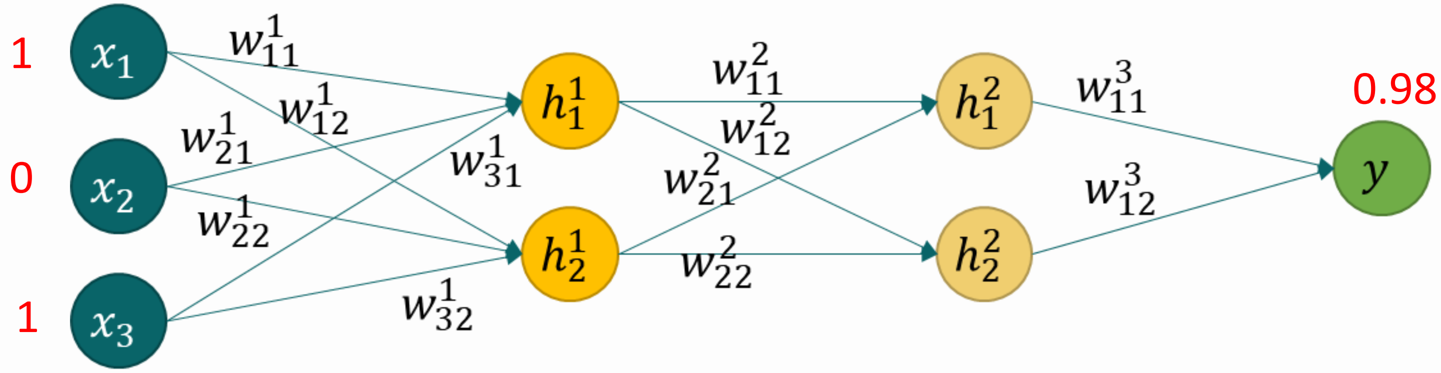
- 反向传播

$$\Delta w_{11}^3 = \delta_1^3 * h_1^2 = -0.0264 * 0.9781 = -0.0258$$
$$\Delta w_{12}^3 = \delta_1^3 * h_2^2 = -0.0264 * 0.9309 = -0.0246$$

$$\Delta w_{11}^2 = \delta_1^2 * h_1^1 = -0.0006 * 2 = -0.0012$$
$$\Delta w_{12}^2 = \delta_2^2 * h_1^1 = -0.0008 * 2 = -0.0016$$
$$\Delta w_{21}^2 = \delta_1^2 * h_1^1 = -0.0006 * 1.5 = -0.0009$$
$$\Delta w_{22}^2 = \delta_2^2 * h_2^1 = -0.0008 * 1.5 = -0.0012$$

$$\Delta w_{11}^1 = \delta_1^1 * x_1 = -0.0012 * 1 = -0.0012$$
$$\Delta w_{12}^1 = \delta_2^1 * x_1 = -0.0014 * 1 = -0.0014$$
$$\Delta w_{21}^1 = \delta_1^1 * x_2 = -0.0012 * 0 = -0.0012$$
$$\Delta w_{22}^1 = \delta_2^1 * x_2 = -0.0014 * 0. = -0.0014$$
$$\Delta w_{31}^1 = \delta_1^1 * x_3 = -0.0012 * 1 = -0.0012$$
$$\Delta w_{32}^1 = \delta_2^1 * x_3 = -0.0014 * 1 = -0.0014$$

附录——作业



$w_{11}^1 = 1$
$w_{12}^1 = 0$
$w_{21}^1 = 1$
$w_{22}^1 = 0.5$
$w_{31}^1 = 1$
$w_{32}^1 = 1.5$

$w_{11}^2 = 1$
$w_{12}^2 = 0.7$
$w_{21}^2 = 1.2$
$w_{22}^2 = 0.8$

$w_{11}^3 = 1$
$w_{12}^3 = 0.5$

- 反向传播

$w_{11}^1 = 1 + 1 * 0.0012 = 1.0012$
$w_{12}^1 = 0 + 0.0014 = 0.0014$
$w_{21}^1 = 1 + 0 = 1$
$w_{22}^1 = 0.5 + 0 = 0.5$
$w_{31}^1 = 1 + 0.0012 = 1.0012$
$w_{32}^1 = 1.5 + 0.0014 = 1.5014$

$w_{11}^2 = 1 + 0.0012 = 1.0012$
$w_{12}^2 = 0.7 + 0.0016 = 0.7016$
$w_{21}^2 = 1.2 + 0.0009 = 1.2009$
$w_{22}^2 = 0.8 + 0.0012 = 0.8012$

$w_{11}^3 = 1 + 0.0258 = 1.0258$
$w_{12}^3 = 0.5 + 0.0246 = 0.5246$