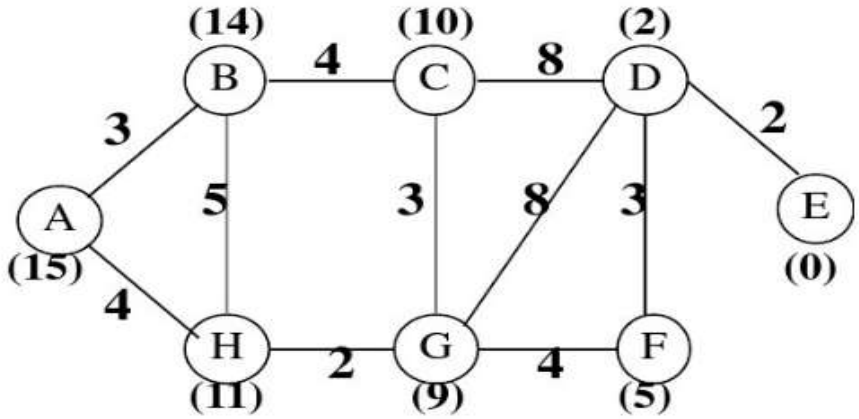
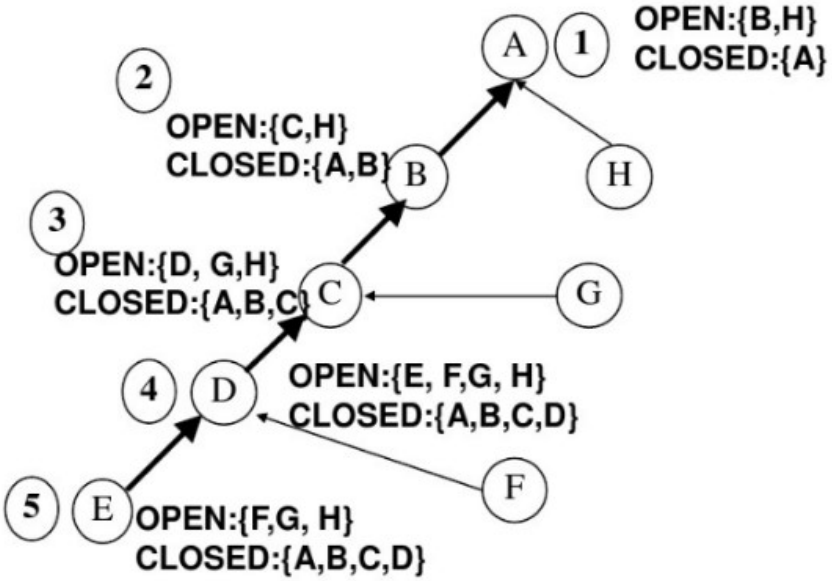
补充:

用 5 种搜索方法分别求解其搜索路径，并给出对应的 OPEN:{ }表和 CLOSED:{ }表

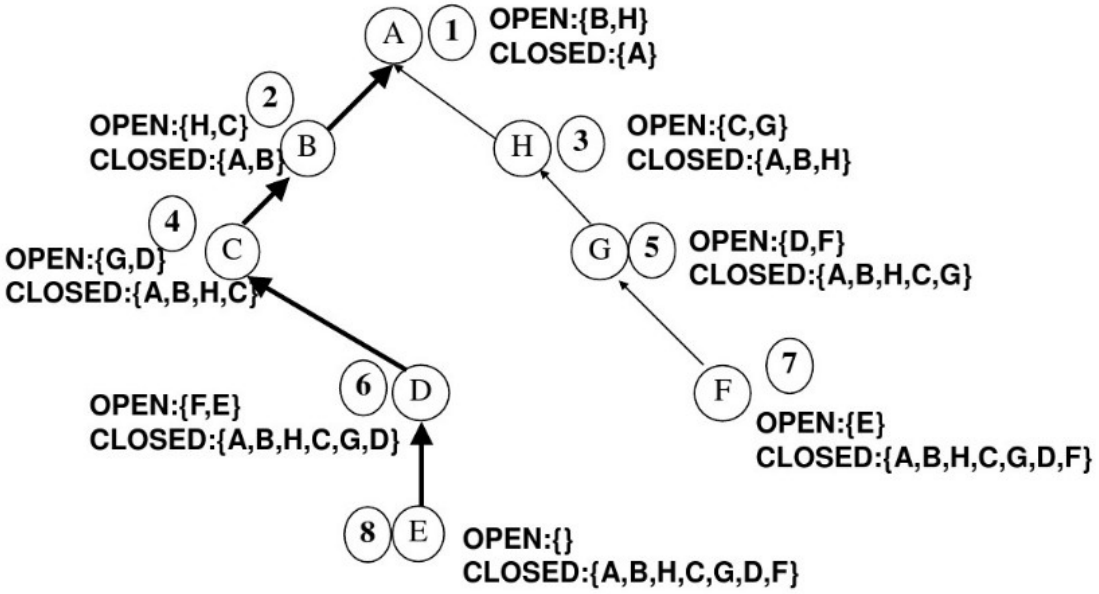


1) 深度优先搜索算法



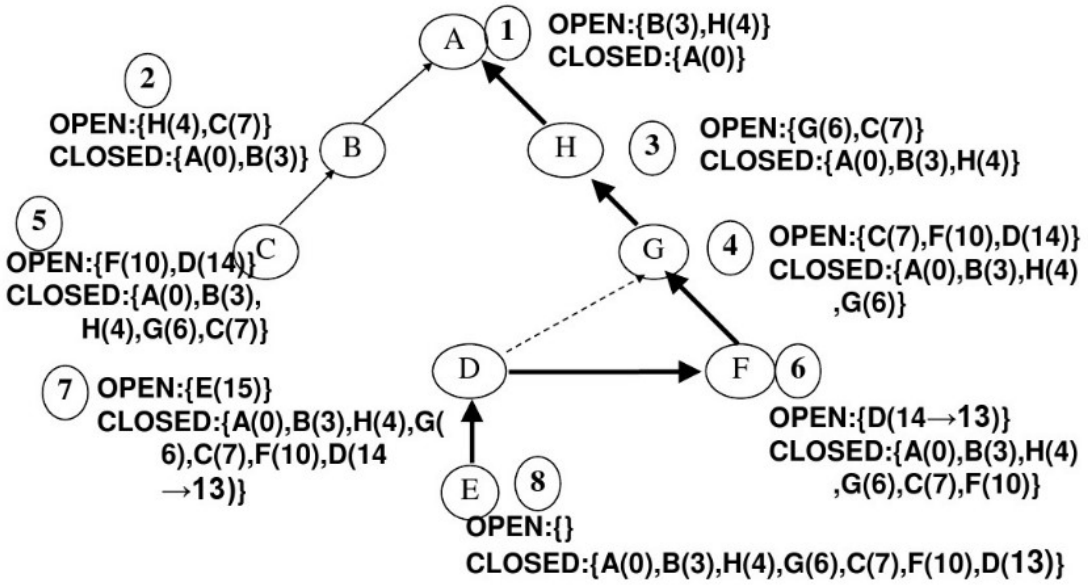
搜索出的路径为： A→B→C→D→E

2) 宽度优先搜索算法



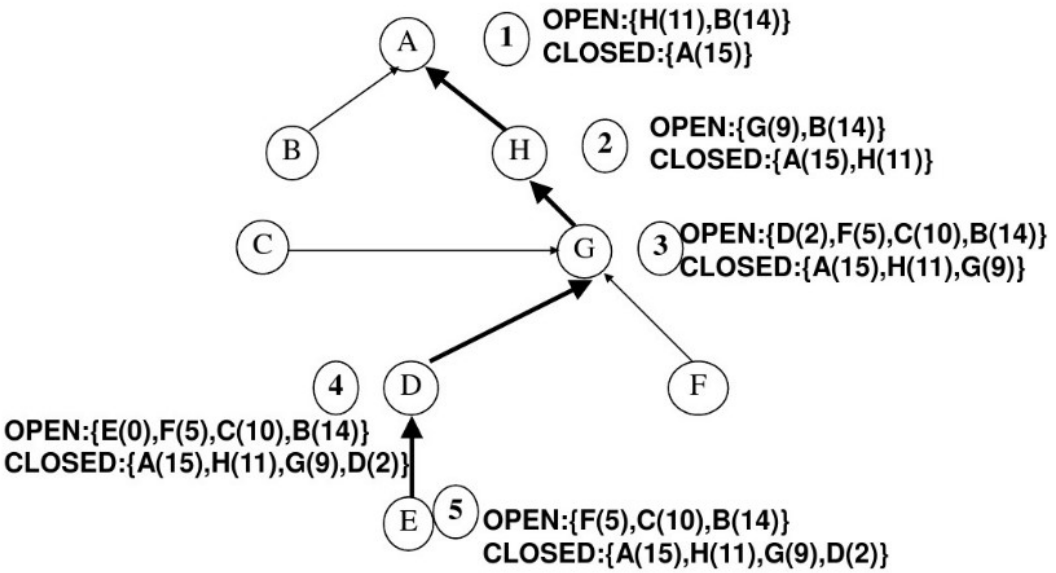
搜索到的路径为：A→B→C→ D→E

3) 动态规划(均一代价)搜索算法



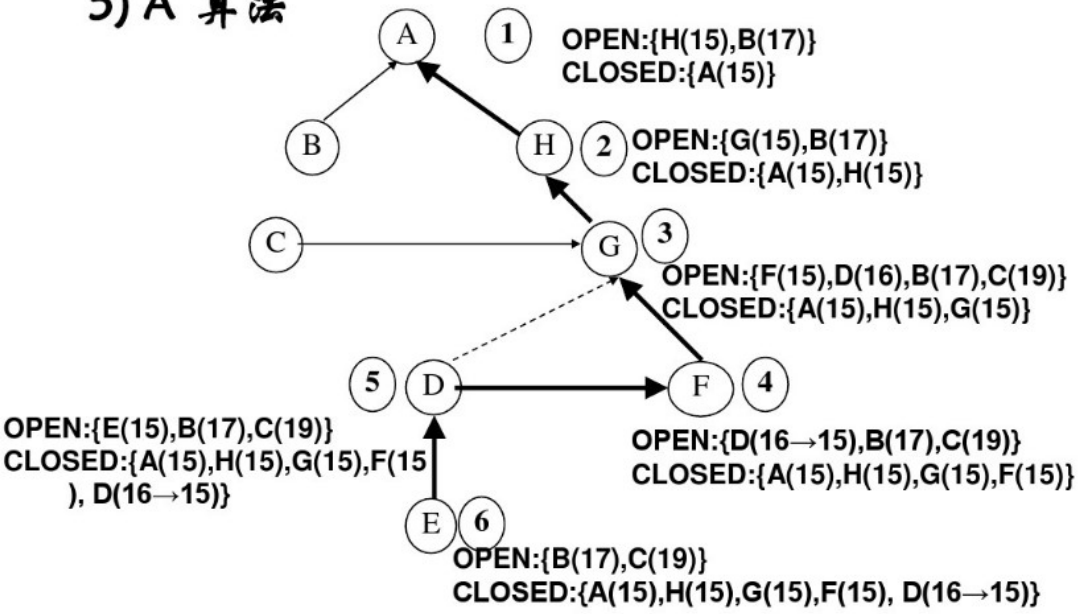
搜索出的路径为：A→H→G→F→D→E，整条路径的代价和为15。

4) 最佳优先搜索算法



搜索出的路径为： A→H→G→D→E，整条路径的代价和为16。

5) A*算法



搜索出的路径为： A→H→G→D→E，整条路径的代价和为15。

第七章

1. 按学习的方式分类，机器学习可分为哪 3 种?

基于学习方式的分类

（1）有导师学习（监督学习(x_i, y_i)）：输入数据中有导师信号，以概率函数、代数函数或人工神经网络为基函数模型，采用迭代计算方法，学习结果为函数。

（2）无导师学习（非监督学习(x_i) ）：输入数据中无导师信号，采用聚类方法，学习结果为类别。典型的无导师学习有发现学习、聚类、竞争学习等。

（3）强化学习（增强学习(能形成奖励函数的数据) ）：以环境反馈（奖/惩信号）作为输入，以统计和动态规划技术为指导的一种学习方法。(伴随蒙特卡洛随机过程）

2. 机器学习、表示学习和深度学习三者之间的联系和区别?

机器学习是对能通过经验自动改进的计算机算法的研究。

为了提高机器学习系统的准确率，我们就需要将输入信息转换为有效的特征，或者更一般性地称为表示(Representation)。如果有一种算法可以自动地学习出有效的特征， 并提高最终机器学习模型的性能，那么这种学习就可以叫作表示学习(Representation Learning)。

为了学习一种好的表示，需要构建具有一定“深度”的模型，并通过学习算法来让模型自动学习出好的特征表示(从底层特征，到中层特征，再到高层特征)，从而最终提升预测模型的准确率。我们就需要一种学习方法可以从数据中学习一个“深度模型”， 这就是深度学习。

3. 简述什么是欠拟合、过拟合。

欠拟合：模型不能在训练集上获得足够低的误差。

过拟合：模型在训练集上误差小，但在测试集上误差大（或远大于训练误差）。

4. 简述 Logistic 回归模型的三要素。

- 模型
  - 线性方法：$f(\mathbf{x},\theta)=\mathbf{w}^{\mathrm{T}}\mathbf{x}+b$
  - 广义线性方法：$f(\mathbf{x},\theta)=\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})+b$
    - 如果$\phi(\mathrm{x})$为可学习的非线性基函数，$f(\mathbf{x},\theta)$就等价于神经网络。

- 学习准则
  - 期望风险

$$\mathcal{R}(f)=\mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}),y)],$$

- 优化
  - 梯度下降

1. PPT 例 7.1 用 ID3 算法完成下述学生选课的例子

2. ID3算法(5/11)

例7.1 用ID3算法完成下述学生选课的例子

假设将决策y分为以下 3 类：

y_1：必修AI

y_2：选修AI

y_3：不修AI

做出这些决策的依据有以下3个属性：

x_1：学历层次 $x_1=1$ 研究生，$x_1=2$ 本科

x_2：专业类别 $x_2=1$ 电信类，$x_2=2$ 机电类

x_3：学习基础 $x_3=1$ 修过AI，$x_3=2$ 未修AI

表7.1给出了一个关于选课决策的训练例子集S。

2. ID3算法(6/11)

表7-1　关于选课决策的训练例子集

序号	属性值			决策方案
	x_1	x_2	x_3	y_i
1	1	1	1	y_3
2	1	1	2	y_1
3	1	2	1	y_3
4	1	2	2	y_2
5	2	1	1	y_3
6	2	1	2	y_2
7	2	2	1	y_3
8	2	2	2	y_3

在该表中，训练例子集S的大小为 8 。ID3算法是依据这些训练例子，以S为根节点，按照信息熵下降最大的原则来构造决策树的。

2. ID3算法(7/11)

序号	属性值			决策方案
	x_1	x_2	x_3	y_i
1	1	1	1	y_3
2	1	1	2	y_1
3	1	2	1	y_3
4	1	2	2	y_2
5	2	1	1	y_3
6	2	1	2	y_2
7	2	2	1	y_3
8	2	2	2	y_3

解：　首先对根节点，其信息熵为：

$$H(S) = -\sum_{}^{3} P(y_i)\log_2 P(y_i)$$

其中，3 为可选的决策方案数，且有

$P(y_1)=1/8$，$P(y_2)=2/8$，$P(y_3)=5/8$

即有：

$$H(S)= -(1/8)\log_2(1/8)- (2/8)\log_2(2/8)- (5/8)\log_2(5/8) =1.2988$$

按照ID3算法，用属性x_i对S进行划分，选一个使S的期望熵最小的属性进行扩展，因此需要先计算S关于每个属性x_i的条件熵：

$$H(S/x_i) = \sum_{t} \frac{|S_t|}{|S|} H(S_t)$$

其中，t为属性x_i的属性值，S_t为x_i=t时的例子集，|S|和$|S_i|$分别是例子集S和S_i的大小。

信息增益：$Gain(S, x_i)=H(S)-H(S/ x_i)$

2. ID3算法(8/11)

序号	属性值			决策方案
	x_1	x_2	x_3	y_i
1	1	1	1	y_3
2	1	1	2	y_1
3	1	2	1	y_3
4	1	2	2	y_2
5	2	1	1	y_3
6	2	1	2	y_2
7	2	2	1	y_3
8	2	2	2	y_3

下面先计算S关于属性x_1的条件熵：

在表7-1中，x_1的属性值可以为1或2。

当x_1=1时，t=1时，有：S_1={1，2，3，4}

当x_1=2时，t=2时，有：S_2={5，6，7，8}

其中，S_1和S_2中的数字均为例子集S中的各个例子的序号，

且有|S|=8，$|S_1|=|S_2|=4$。

由S_1可知：

$Ps_1(y_1)=1/4$，　$Ps_1(y_2)=1/4$，　$Ps_1(y_3)=2/4$

则有：

$$H(S_1)= -Ps_1(y_1)\log_2 Ps_1(y_1)-Ps_1(y_2)\log_2 Ps_1(y_2)-Ps_1(y_3)\log_2 Ps_1(y_3)$$
$$= -(1/4)\log_2(1/4)-(1/4)\log_2(1/4)-(2/4)\log_2(2/4) =1.5$$

2. ID3算法(9/10)

序号	属性值			决策方案
	x_1	x_2	x_3	y_i
1	1	1	1	y_3
2	1	1	2	y_1
3	1	2	1	y_3
4	1	2	2	y_2
5	2	1	1	y_3
6	2	1	2	y_2
7	2	2	1	y_3
8	2	2	2	y_3

再由S_2可知：

$Ps_2(y_1)=0/4$，　$Ps_2(y_2)=1/4$，　$Ps_2(y_3)=3/4$

则有：$H(S_2)=-Ps_2(y_2)\log_2 Ps_2(y_2)-Ps_2(y_3)\log_2 Ps2(y_3)$

$$=-(1/4)\log_2(1/4)-(3/4)\log_2(3/4) =0.8113$$

将$H(S_1)$和$H(S_2)$代入条件熵公式，有：

$$H(S/x_1)=(|S_1|/|S|)H(S_1)+(|S_2|/|S|)H(S_2)$$
$$=(4/8)*1.5+(4/8)*0.8113 =1.1557$$

同理，可以求得：

$H(S/x_2)=1.1557$

$H(S/x_3)=0.75$

根据最小熵原理，应该选择属性x_3对根节点进行扩展。

用x_3对S扩展后所得到的部分决策树如图7.5所示。

2. ID3算法(10/11)

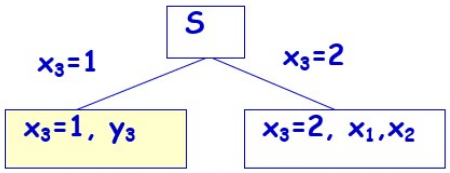


图7.5 部分决策树

在该树中，节点“$x_3=1$，y_3”表示当x_3的属性值为1时，得到决策方案y_3。由于y_3已是具体的决策方案，故该节点的信息熵为0，已经为叶节点。

节点“$x_3=2$，x_1,x_2”的含义是“当x_3的属性值为2时，还需要考虑属性x_1,x_2”，它是一个中间节点，还需要继续扩展。

至于节点“$x_3=2$，x_1,x_2”，其扩展方法与上面的过程类似。通过计算可知，该节点对属性x_1和x_2，其条件熵均为1。因它对属性x_1和x_2的条件熵相同，因此可先选x_1，也可先选x_2，本例是先选择x_2。

依此进行下去，可得到如图7.6所示的最终的决策树。在该决策树中，各节点的含义与图7.5类似。

16

2. ID3算法(11/11)

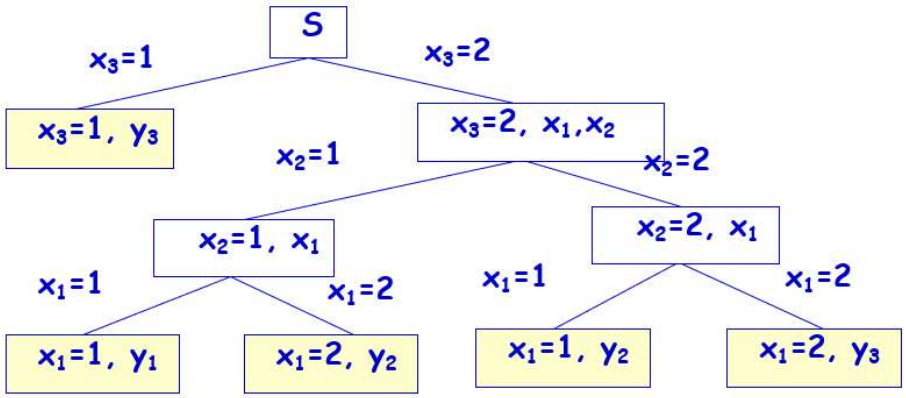


图7.6 最终的决策树

2. 对离散随机变量 X 而言，信息论原理中自信息、信息熵、交叉熵、KL 散度的数学定义分别是什么？

自信息：对于一个可能取值为 x_1,x_2,...,x_n 的离散随机变量 X，它的概率质量函数 P（X），以及任何正的取值在 0 到 1 之间的单调递减函数 I(p_i) 都可以作为信息的度量.我们使用负对数表示自信息：

$$I(p_i) = -log(p_i).$$

信息熵: 是 X 的所有可能结果的自信息期望值：

$$H(X) = E_{x\sim P}[I(x)] = \sum_{i=1}^{n} p(x_i)\, I(x_i) = -\sum_{i=1}^{n} p(x_i)\, log_b(p(x_i)),$$

交叉熵是一个用来比较两个概率分布 p 和 q 的数学工具,与熵类似，计算 log(q) 在概率 p 下的期望：

$$H(p,q) = -\sum_{x} p(x)\, \log q(x)$$

KL 散度是一个在机器学习中用来衡量两个概率分布 P 和 Q 相似度的量：从 Q 到 P 的 KL 散度如下: D_KL(P||Q)

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{i} P(i)\, \log\left(\frac{Q(i)}{P(i)}\right),$$

3. 对离散随机变量 X 而言，信息论原理中信息熵、交叉熵、KL 散度三者的数学关系是什么？
信息熵: H(p)，H(p,p)
交叉熵：H(p, q)，
KL 散度：KL(p||q)
KL(p||q) = H(p, q) - H(p)

1. 误差反向传播(Error Back Propagation)网络的网络拓扑结构是什么？BP 网络的学习过程是由哪两种传播组成。

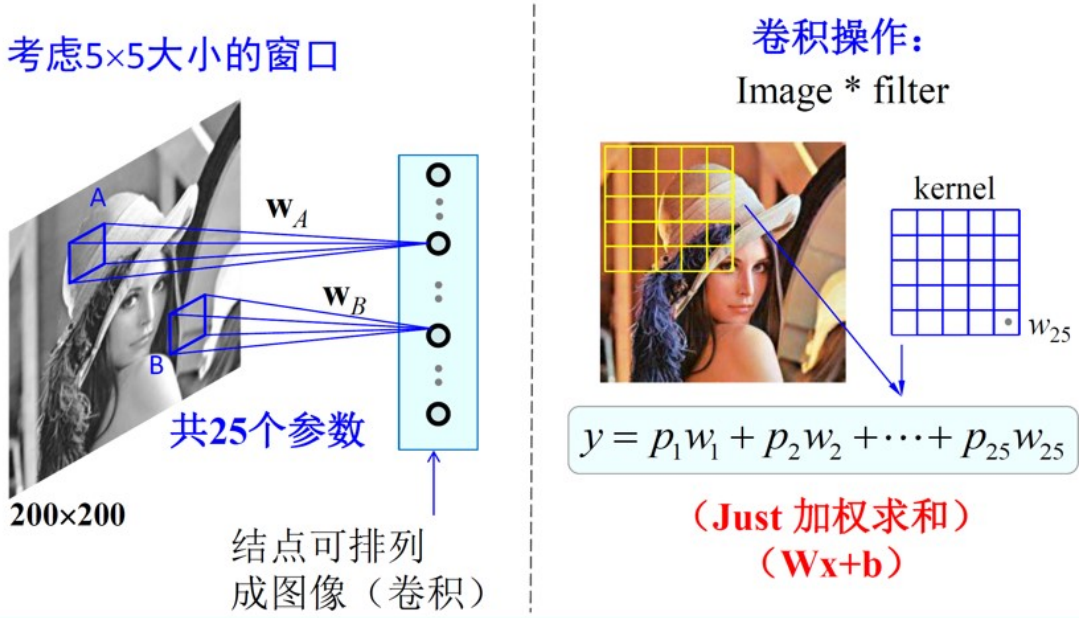
误差反向传播(Error Back Propagation)网络的网络拓扑结构是多层前向网络，在 BP 网络中，同层节点之间不存在相互连接，层与层之间多采用全互连方式，且各层的连接权值可调。

BP 网络的学习过程是由工作信号的正向传播和误差信号的反向传播组成的。

2. CNN 中的卷积和池化计算

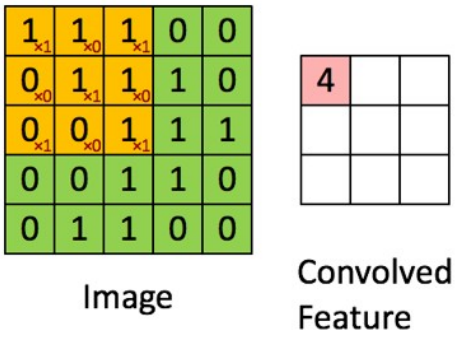
卷积神经网络：基本原理

- 利用图像局部结构进行卷积操作



CNN的Convolution过程

如图，原图像是5*5大小，有25个神经元，用一个3*3的卷积核对它进行卷积，得到了如右图所示的卷积后的Feature map。
该特征图大小为3*3。



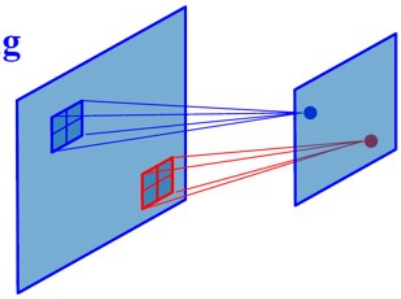
假设一种卷积核只提取出图像的一种特征，所以一般要多个卷积核来提取不同的特征，所以每一层一般都会有多张Feature map。

同一张Feature map上的神经元共用一个卷积核，这大大减少了网络参数的个数。

卷积神经网络：降采样 Pooling

- 降采样：增强网络的鲁棒性，抗形变、位移干扰
  - Average pooling 取平均
  - Max pooling 取最大

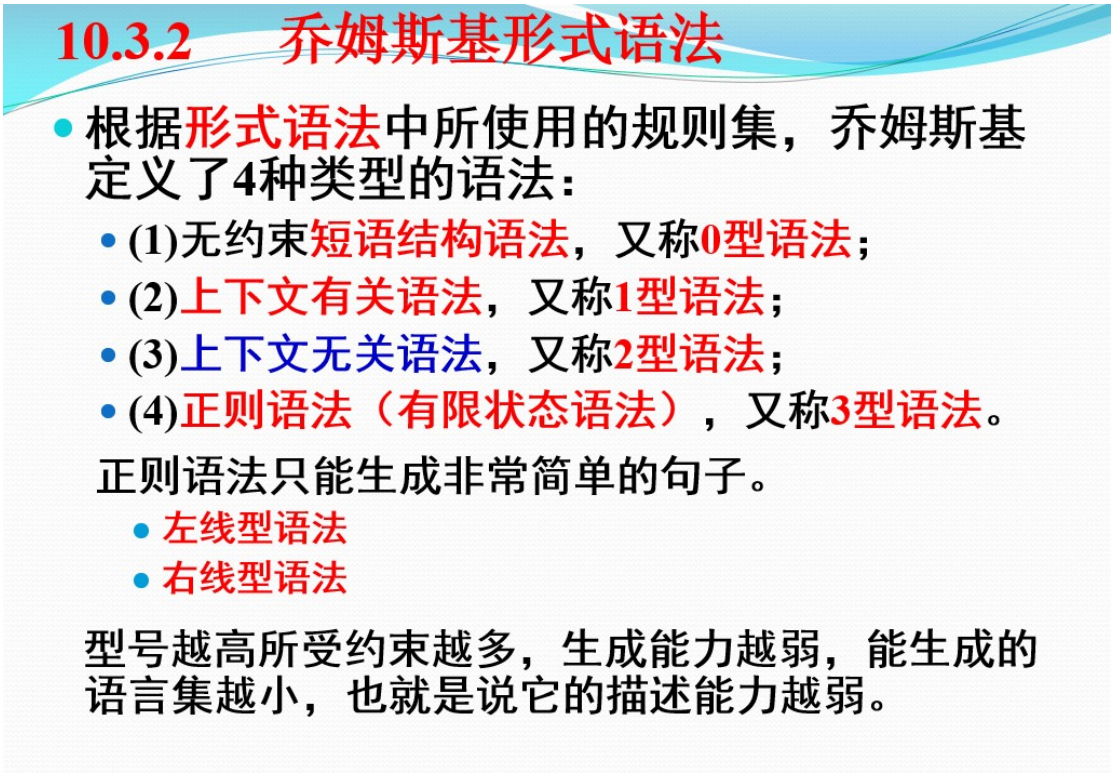
2×2区域内pooling



图像大小196×196　　降采样之后 98×98

第十章

1. 乔姆斯基语法体系将语法分成了哪 4 类？



2. 请用语法树结构解析句子

上下文无关文法(Context-free Grammars)是乔姆斯基提出的一种对自然语言语法知识进行形式化描述的方法。在这种文法中，语法知识是用重写规则表示的。作为例子，下面给出了一个英语的很小的子集。

语句 → 句子 终标符
句子 → 名词短语 动词短语
动词短语 → 动词 名词短语
名词短语 → 冠词 名词
名词短语 → 专用名词
冠词 → **the**
名词 → **professor**
动词 → **wrote**
名词 → **book**
动词 → **trains**
专用名词 → **Jack**
终标符 →.

这就是一个英语子集的上下文无关文法
在该文法中，“语句”是一个特殊的非终极符，称为起始符。 27

例 利用上述上下文无关文法，给出如下语句的分析树。

The professor trains Jack.

语句 → 句子 终标符
句子 → 名词短语 动词短语
动词短语 → 动词 名词短语
名词短语 → 冠词 名词
名词短语 → 专用名词
冠词 → **the**
名词 → **professor**
动词 → **wrote**
名词 → **book**
动词 → **trains**
专用名词 → **Jack**
终标符 →.