《人工智能导论：模型与算法》习题参考答案

第一章 绪论

1. **B**

2. **D**

3. **B**

4. **D**

5. **C**

6. **B**

7. 略

8. 略

9. 略

10. 参考答案：

强化学习有环境、智能体、状态、奖励、决策等诸多要素，涉及序列决策过程，智能体之前作出的决策会影响智能体当前的状态，从而影响“未来”的决策过程。而监督学习中，对每一个样本输入做出的决策不会影响到“未来”的决策。

监督学习的每次决策后得到的反馈是“最终反馈”，它包含了最佳决策的信息。而强化学习每次决策后得到的反馈只是当前步的反馈。

11. 略

第二章 逻辑与推理

1. **C**

2. **D**

3. **D**

4. **D**

5. 参考答案：
a) 不是命题，因为无法判断真假；
b) 不是命题，因为不是陈述句；
c) 是命题，其真值为真。

6. 参考答案：
应用归结法：
(1) $\alpha \vee \beta$ (已知);
(2) $\neg\beta \vee \gamma$ (b 进行蕴涵消除);
(3) $\alpha \vee \gamma$ (由 1 和 2);
(4) $\neg(\alpha \vee \gamma)$ (c 使用 De Morgan 定律)
(5) 3 和 4 矛盾，因此原命题集是不可满足的。

7. 参考答案：
a) $\neg(\forall x)(Basketball\_player(x) \rightarrow Higher\_than(x, 1.8m))$。其中，$Basketball_player(x)$表示$x$是篮球运动员，$Higher_than(x, 1.8m)$表示$x$的身高超过 1 米 8。

b) $(\forall x)\big(Real(x) \rightarrow More\_than(Square(x),\ 0)\big)$。其中，$Real(x)$表示$x$是实数，$Square(x)$表示$x$的平方，$More_than(Square(x),\ 0)$表示$x$的平方大于等于 0。

8. 参考答案：
(1) $\neg(\forall x)(F(x) \rightarrow G(x))$
(2) $(\exists x)\neg(F(x) \rightarrow G(x))$
(3) $(\exists x)\neg(\neg F(x) \vee G(x))$
(4) $(\exists x)(F(x) \wedge \neg G(x))$
(5) $F(a) \wedge \neg G(a)$ (存在量词消去)
(6) $F(a)$ (由 5 知)
(7) $\neg G(a)$ (由 5 知)
(8) $(\forall x)(F(x) \rightarrow G(x) \vee H(x))$
(9) $F(a) \rightarrow G(a) \vee H(a)$ (全称量词消去)
(10) $G(a) \vee H(a)$ (6 和 9 的假言推理)
(11) $H(a)$ (由 7 和 10 知)
(12) $F(a) \wedge H(a)$ (由 6 和 11 知)
(13) $(\exists x)(F(x) \wedge H(x))$ (存在量词引入)

9. 参考答案：

在给定目标谓词$Mother(x,y)$之后，可如下表来构造背景知识样例和训练样例。

背景知识样例集合	$Sibling$(Ann, Mike) $Couple$(James, David) $Father$(David, Ann) $Father$(David, Mike)	目标谓词训练样例集合	$Mother$(James, Mike) ¬$Mother$(James, David) ¬$Mother$(David, Ann) ¬$Mother$(David, Mike) ¬$Mother$(Ann, Mike)

按照表 2.3.1 中 FOIL 算法所列步骤依次将每个候选前提约束谓词加入到推理规则中，并计算所得新的推理规则对应的 FOIL 增益值。基于计算所得 FOIL 增益值来确定最佳前提约束谓词。下表给出了添加前提约束谓词后信息增益的计算结果。

推理规则		推理规则涵盖的正例和反例数		FOIL 信息增益值
目标谓词	前提约束谓词	正例	反例	信息增益值
$Mother(x,y)$ ←	空集	$m_+ = 1$	$m_- = 4$	/
$Mother(x,y)$ ←	$Father(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 2$	NA
	$Father(x,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 2$	NA
	$Father(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$Father(y,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$Father(z,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$Father(z,y)$	$\widehat{m_+} = 1$	$\widehat{m_-} = 3$	0.32
	$Sibling(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$Sibling(x,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$Sibling(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$Sibling(y,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$Sibling(z,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$Sibling(z,y)$	$\widehat{m_+} = 1$	$\widehat{m_-} = 2$	0.74
	$Couple(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$\boldsymbol{Couple(x,z)}$	$\boldsymbol{\widehat{m_+} = 1}$	$\boldsymbol{\widehat{m_-} = 1}$	**1.32**
	$Couple(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$Couple(y,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$Couple(z,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 2$	NA
	$Couple(z,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA

由于$Couple(x,z)$的 FOIL 增益值最大，因此选择$Couple(x,z)$加入推理规则，得到$Couple(x,z) \rightarrow Mother(x,y)$这一新的推理规则，并将训练样例集合中与该推理规则不符样例去掉。此时，背景知识样例和训练样本如下表：

背景知识样例集合	$Sibling$(Ann, Mike) $Couple$(James, David) $Father$(David, Ann) $Father$(David, Mike)	目标谓词训练样例集合	$Mother$(James, Mike) $\neg Mother$(James, David)

接着，再用相同方法继续将其他谓词逐一作为前提约束谓词加入推理规则进行考察，用 FOIL 增益值来判断选取最优推理规则。

推理规则		推理规则涵盖的正例和反例数		FOIL 信息增益值
现有规则	拟加入前提约束谓词	正例	反例	信息增益值
$Mother(x,y) \leftarrow Couple(x,z)$		$m_+ = 1$	$m_- = 1$	/
$Mother(x,y) \leftarrow Couple(x,z)$	$\wedge Father(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Father(x,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Father(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Father(y,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Father(z,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge \boldsymbol{Father(z,y)}$	$\boldsymbol{\widehat{m_+} = 1}$	$\boldsymbol{\widehat{m_-} = 0}$	**1**
	$\wedge Sibling(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Sibling(x,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Sibling(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Sibling(y,z)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Sibling(z,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Sibling(z,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA
	$\wedge Couple(x,y)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 1$	NA
	$\wedge Couple(x,z)$	$\widehat{m_+} = 1$	$\widehat{m_-} = 1$	0
	$\wedge Couple(y,x)$	$\widehat{m_+} = 0$	$\widehat{m_-} = 0$	NA

$\wedge Couple(y,z)$	$\widehat{m_+}=0$	$\widehat{m_-}=0$	NA
$\wedge Couple(z,x)$	$\widehat{m_+}=0$	$\widehat{m_-}=0$	NA
$\wedge Couple(z,y)$	$\widehat{m_+}=0$	$\widehat{m_-}=0$	NA

当$Father(z,y)$作为前提约束谓词加入到推理规则后$FOIL_Gain$值最大，因此将$Father(z,y)$加入，得到新的推理规则$Father(z,y) \wedge Couple(x,z) \rightarrow Mother(x,y)$。当$x$= James、$y$=Mike 和$z$=David 时，该推理规则覆盖训练样本集合的正例$Mother$(James, Mike)、且不覆盖任意反例，因此算法学习结束。

当学习得到$(\forall x)(\forall y)(\forall z)\big(Father(z,y) \wedge Couple(x,z) \rightarrow Mother(x,y)\big)$这一推理规则后，由题意可知$Father$(David, Ann)和$Couple$(James, David)，因此可推理得到新的知识$Mother$(James, Ann)，即 James 和 Ann 的关系为$Mother$关系。

10. 参考答案：

(1) 目标关系：$Mother$

(2) 对于目标关系$Mother$，生成四组训练样例，一个为正例、三个为负例：

正例：(James, Mike)

负例：(James, David)，(David, Ann)，(David, Mike)

(3) 从知识图谱采样得到路径，每一路径链接上述每个训练样例中两个实体：

(James, Mike)对应路径：$Couple \rightarrow Father$

(James, David)对应路径：$Mother \rightarrow Father^{-1}$（$Father^{-1}$与$Father$为相反关系）

(David, Ann)对应路径：$Father \rightarrow Sibling$

(David, Mike)对应路径：$Couple \rightarrow Mother$

(4) 对于每一个正例/负例，判断上述四条路径可否链接其包含的两个实体，将可链接（记为 1）和不可链接（记为 0）作为特征，于是每一个正例/负例得到一个四维特征向量：

(James, Mike)：$\{[1,0,0,0],1\}$

(James, David)：$\{[0,1,0,0],-1\}$

(David, Ann)：$\{[0,0,1,0],-1\}$

(David, Mike)：$\{[0,0,1,1],-1\}$

(5) 依据(4)中的训练样本，训练分类器M。

(6) 预测。对于样例 (James, Ann)，得到其特征值为$[1,0,0,0]$，将特征向量输入到分类器M中，分类器M给出分类结果为1，$Mother$(David, Ann)成立。

11. 参考答案：

下面我们用本章节式 2.4.4（调整公式）来求解是否存在“女性歧视”问题。用$X=1$表示男性，$X=0$ 表示女性；$Y=1$表示录取；$Z=$ 0,1,2,3 分别表示英语、俄语、西班牙语和意大利语，则有：

$$P\big(Y=1\big|do(X=1)\big)$$

$$=P(Y=1|X=1,Z=0)P(Z=0)+P(Y=1|X=1,Z=1)P(Z=1)$$
$$+P(Y=1|X=1,Z=2)P(Z=2)+P(Y=1|X=1,Z=3)P(Z=3)$$

带入表 1 中数据，可得男性被录取的因果效应为：

$$P\big(Y=1\big|do(X=1)\big)$$

$$=0.62\times\frac{(825+108)}{(2175+849)}+0.63\times\frac{(560+25)}{(2175+849)}$$
$$+0.33\times\frac{(417+375)}{(2175+849)}+0.06\times\frac{(373+341)}{(2175+849)}=0.41376$$

类似地，可以求得女性被录取的因果效应为：

$$P\big(Y=1\big|do(X=0)\big)$$

$$=0.82\times\frac{(825+108)}{(2175+849)}+0.68\times\frac{(560+25)}{(2175+849)}$$
$$+0.35\times\frac{(417+375)}{(2175+849)}+0.07\times\frac{(373+341)}{(2175+849)}=0.49274$$

最后，我们发现女性录取率（0.49274）比男性录取率（0.41376）要高，即不存在所谓的“女性歧视”。

12. 参考答案：

使用乘积分解规则：

$$P\big(X_1,X_2,X_3,X_4,X_5,X_6,X_i,X_j\big)$$
$$=P(X_1)\times P(X_2|X_4,X_5)\times P(X_3|X_1)\times P(X_4|X_1,X_3,X_5)\times P(X_5)$$
$$\times P(X_6|X_3)\times P(X_7|X_4,\,X_6,\,X_8)\times P(X_8|X_5)$$

外生变量：X_1,X_5；内生变量：X_2,X_3,X_4,X_6,X_7,X_8

13*. 参考答案：

为了阻塞节点X_6和节点X_8，我们需要让从节点X_6到节点X_8的所有路径满足D-分离性质。从节点X_6到节点X_8共有以下 7 条路径：

P1 $X_6\rightarrow X_7\leftarrow X_8$ (基于定义 2.18(2)，节点X_7不能在限定集Z中)

P2 $X_6\leftarrow X_3\rightarrow X_4\rightarrow X_7\rightarrow X_8$ (基于定义 2.18，节点X_3或X_4或X_7需要出现在限定集Z中)

P3 $X_6\leftarrow X_3\rightarrow X_4\leftarrow X_5\rightarrow X_8$ (基于定义 2.18，节点X_3或X_5出现在Z中，或者节点X_4不出现在Z中)

P4 $X_6\leftarrow X_3\rightarrow X_4\rightarrow X_2\leftarrow X_5\rightarrow X_8$ (基于定义 2.18，节点X_3或X_4或X_5出现在Z中，或者节点X_2不出现在Z中)

P5　$X_6 \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \rightarrow X_7 \rightarrow X_8$ (基于定义 2.18，节点X_3或X_1或X_4或X_7出现在Z中即可)
P6　$X_6 \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \leftarrow X_5 \rightarrow X_8$ (基于定义 2.18，节点X_3或X_1或X_5出现在Z中，或者节点X_4不出现在Z中)
P7　$X_6 \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \rightarrow X_2 \leftarrow X_5 \rightarrow X_8$ (基于定义 2.18，节点X_3或X_1或X_4或X_5出现在Z中，或者节点X_2不出现在Z中)
P8: $X_6 \rightarrow X_7 \leftarrow X_4 \leftarrow X_5 \rightarrow X_8$
P9: $X_6 \rightarrow X_7 \leftarrow X_4 \rightarrow X_2 \leftarrow X_5 \rightarrow X_8$

综上，能让从节点X_6到节点X_8的所有路径满足D-分离性质的限定集为：
所有包含节点X_3或$\{X_4, X_5\}$且不包含节点X_7的限定集，例如$\{X_3\}$，$\{X_3, X_4\}$等。

第三章　搜索与求解

1. **B**

A 如果图不连通，则可能不存在路径。如果图中存在负值回路（当然还有其他情况），则可能不存在最短路径。

B 显然不是最优的。

C 在这种情况下，节点所在层数和其路径长度是成正比的，因此优先扩展浅层节点等价于优先扩展路径代价小的节点，这在图搜索中是最优的（可参见 Dijkstra 算法）。

D 因为图搜索是在树搜索的基础上进一步剪枝，因此扩展的节点数量通常更少。

2. **D**

A 只有可容的启发函数才不会过高估计从当前节点到目标结点之间的实际代价。

B 如果存在负值边，则很容易构造反例。

C 启发函数通常是对当前节点到目标节点距离的估计，评价函数不一定有实际意义。

D 根据对 A*算法的分析，不难证明。

3. **A**

A 只需要重新定义黑方的动作为每次落两子即可。

B 导致问题中信息不完全，因此 Minimax 算法无法求解。

C 导致问题不再是两人对抗问题，每个人的目标不能再简单地用最大化/最小化某一个人的分数来衡量。

D 使该问题不是零和博弈。白方最大化自己的分数不一定必须最小化黑方的分数。

4. **D**

A、B、C 显然正确。

D 中置信上界的含义是样本取值以极大的概率不会超过置信上界，并不是说不可能超过。

5. **B**

A 选择过程中 UCB1 算法即体现了探索与利用的平衡。

B 只要有一个子节点未被扩展，算法就会进入扩展步骤。

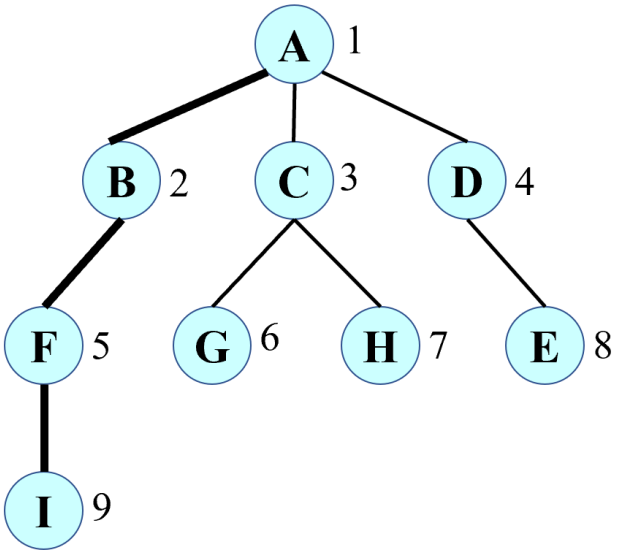
C 模拟步骤的策略不一定要和选择步骤相同，模拟步骤通常会采取更简单的策略。

D 对。更新当前路径上的节点，且不在搜索树中的当然不用更新。

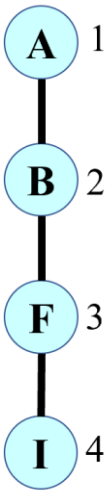
6. 参考答案：

下图中粗线表示路径，节点旁的数字表示扩展顺序。

(1)



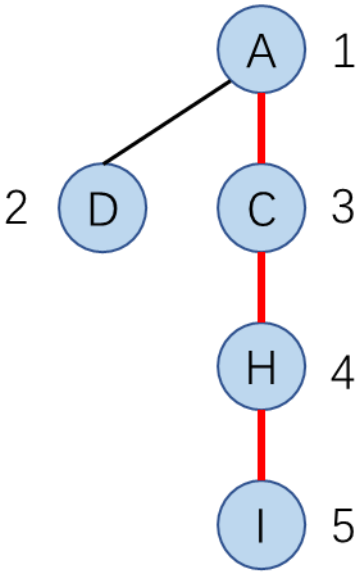
(2)



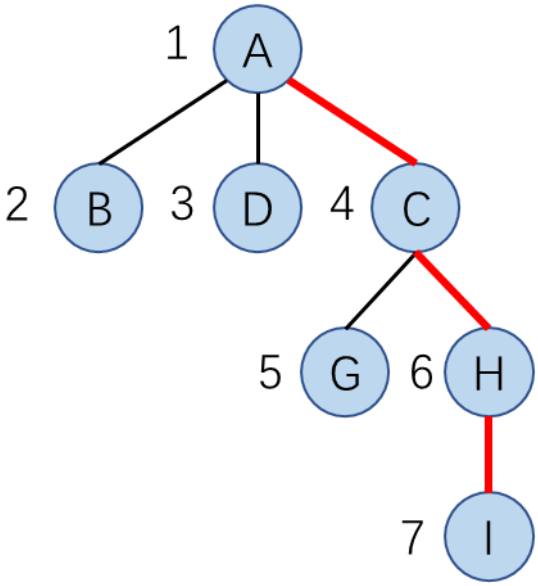
7. 参考答案：

下图中红线表示路径，节点旁的数字表示扩展顺序。

(1)



(2)



8. 参考答案：

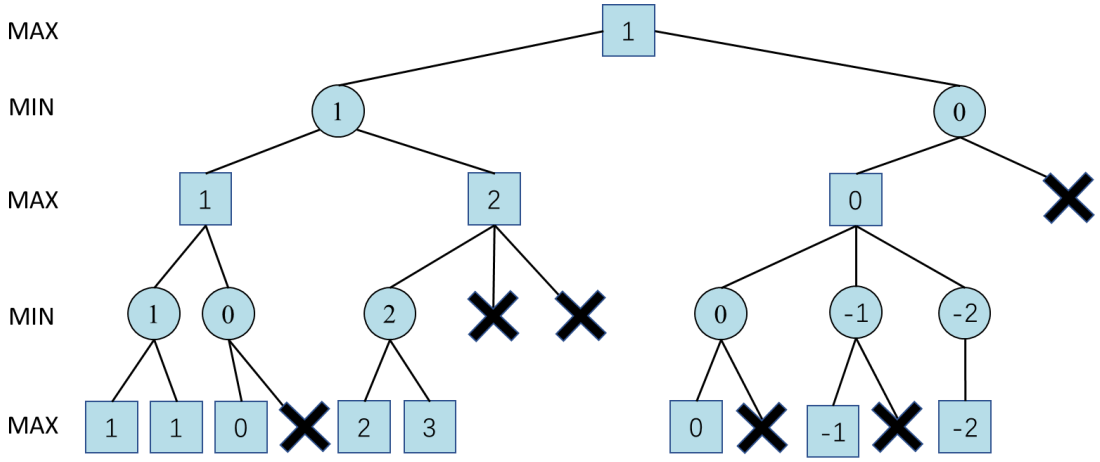
(1) 不矛盾。贪婪最佳优先搜索并不具有最优性，本题中的例子只是它能找到最优解的一个特例。“A*搜索是在已知信息下同类搜索策略中最优的”，其含义是：额外信息仅包括当前的启发函数时，所有能够保证最优性（即在任意有最短路径的问题中都能找到最短路径）的算法中，A*算法扩展的节点数量是最少的。

(2) 启发函数在满足可容性或一致性的基础上，其值越接近当前节点到终止节点的最小代价，搜索的效率越高。当启发函数值等于当前节点到终止节点的最小代价时，算法每一步都会朝着最优的方向探索，以$O(m)$的复杂度得到最优解。

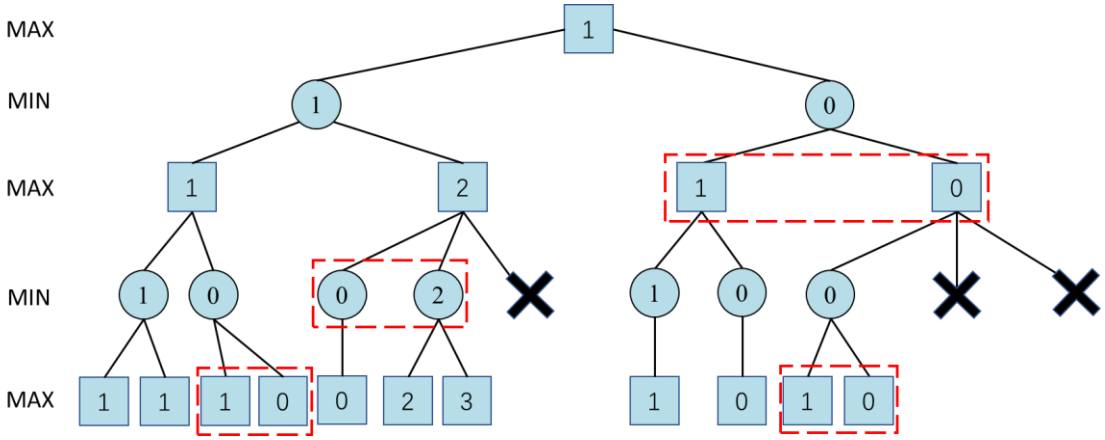
9. 参考答案：

扩展情况如下。显然扩展顺序会影响最终扩展的节点数量（算法时间效率）。

(1) 扩展节点数量为 20。

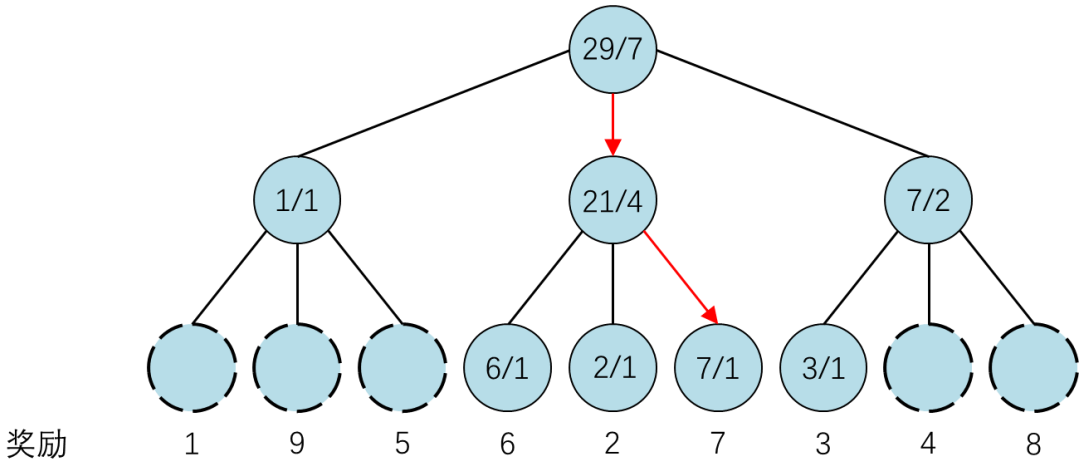


(2) 扩展节点数量为 25。

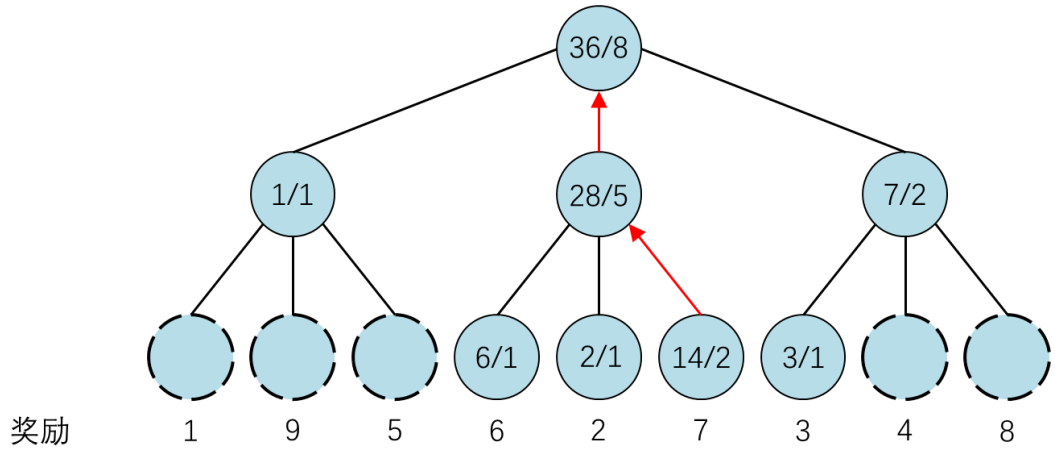


10. 参考答案：

(1) 第一步三个节点的UCB值从左到右分别为$\frac{1}{1}+\sqrt{\frac{2\ln 7}{1}}=2.97,\frac{21}{4}+\sqrt{\frac{2\ln 7}{4}}=6.24,\frac{7}{2}+\sqrt{\frac{2\ln 7}{2}}=4.89$，因此第一步选择第二层中间的节点。第二步三个节点的 UCB 值从左到右分别为$\frac{6}{1}+\sqrt{\frac{2\ln 4}{1}}=7.67$，$\frac{2}{1}+\sqrt{\frac{2\ln 4}{1}}=3.67$，$\frac{7}{1}+\sqrt{\frac{2\ln 4}{1}}=8.67$，因此第二步选择奖励为 7 的节点。如下图所示。



(2) 由于此时已经到达叶子节点，因此不需要进行扩展和模拟过程，反向传播后结果如下图所示



(3) 算法在很长一段时间内都会选择奖励为 7 的节点，而不会探索奖励为 9 的节点。当实验次数足够多时，第二层左侧的节点的 UCB 值最终会超过第二层中间节点的 UCB 值，因此只要实验次数足够多，算法是有可能探索到奖励为 9 的节点的。如果希望提高算法的效率，可考虑加大探索的力度，即取一个更大的超参数C。不难验证，在原题中的状态下，取$C = 10$即可令算法选择第二层左侧的节点。

11. 参考答案：

假设$s_1 \to \cdots \to s_i$的代价为C_1，$s_i \to \cdots \to s_k$的代价为C_2，由于$s_1 \to s_2 \to \cdots \to s_k$是一条最短路径，因此从$s_1$到$s_k$的最小代价$C^* = C_1 + C_2$。已知路径$P$是从$s_1$到$s_i$的最短路径，因此$P$的代价$C_1^* \le C_1$，所以$C^* = C_1 + C_2 \ge C_1^* + C_2$，即路径$P \to s_{i+1} \to \cdots \to s_k$的代价小于等于从$s_1$到$s_k$最短路径的代价，因此$P \to s_{i+1} \to \cdots \to s_k$必然也为一条最短路径。

12. 参考答案：

(1) 因为记录该分数是为了让当前节点的父节点选择一个置信上界最大的孩子节点，由于要最大化父节点玩家的收益相当于最小化当前玩家的收益，因此此处需要减去当前玩家的而收益。如果要修改为加上终局得分，该得

分应该从当前节点的父节点或子节点玩家的角度来计算。

(2) 关键在于修改BackPropagate函数的第三行，由于不存在两个玩家对抗，此处只要加上终局奖励得分（或代价的相反数）即可。

13*. 参考答案：

当搜索树的高度（最长路径长度）为m且是一棵满b叉树时，假设 alpha-beta 剪枝扩展的最少节点数量为$f(m)$。如本章正文中图 3.3.11 所示，观察其第二层的节点，不难发现：因为第二层最左侧节点继承了根节点$(-\infty,+\infty)$的范围，因此对该节点所在的子树进行搜索等价于树的高度为$m-1$的原问题，因此扩展节点数量最少为$f(m-1)$；考虑第二层的其余$b-1$个节点所在的子树，最优情况下这些高度为$m-1$的子树的每个第一层节点扩展1个子节点，每个第二层节点扩展b个子节点，每个第三层节点扩展1个子节点，每个第四层节点扩展b个子节点，如此循环。因此每棵子树扩展节点数为

$$
\begin{aligned}
&1+1+b+b+\cdots+b^{\left\lfloor\frac{m-2}{2}\right\rfloor}+b^{\left\lfloor\frac{m-1}{2}\right\rfloor}\\
&\le 2\left(1+b+\cdots+b^{\left\lfloor\frac{m-1}{2}\right\rfloor}\right)\\
&=2\frac{\left(b^{\left\lfloor\frac{m-1}{2}\right\rfloor+1}-1\right)}{(b-1)}\\
&<2\frac{b^{\left\lfloor\frac{m-1}{2}\right\rfloor+1}}{\frac{b}{2}}\\
&=4b^{\left\lfloor\frac{m-1}{2}\right\rfloor}
\end{aligned}
$$

由此可得递推关系

$$
\begin{aligned}
f(m)&<f(m-1)+4(b-1)b^{\left\lfloor\frac{m-1}{2}\right\rfloor}+1\\
f(m-1)&<f(m-2)+4(b-1)b^{\left\lfloor\frac{m-2}{2}\right\rfloor}+1\\
\cdots&<\cdots
\end{aligned}
$$

不等式左右两边同时求和，可得

$$
\begin{aligned}
f(m)&<f(0)+m+4(b-1)\left(b^{\left\lfloor\frac{0}{2}\right\rfloor}+b^{\left\lfloor\frac{1}{2}\right\rfloor}+\cdots+b^{\left\lfloor\frac{m-1}{2}\right\rfloor}\right)\\
&\le 1+m+8(b-1)\left(1+b+\cdots+b^{\left\lfloor\frac{m-1}{2}\right\rfloor}\right)\\
&=1+m+8(b-1)\frac{b^{\left\lfloor\frac{m-1}{2}\right\rfloor+1}-1}{b-1}\\
&=8b^{\left\lfloor\frac{m+1}{2}\right\rfloor}+m-7\\
&=O(b^{\left\lfloor\frac{m+1}{2}\right\rfloor}+m)
\end{aligned}
$$

在b和m足够大时，显然$b^{\left\lfloor\frac{m+1}{2}\right\rfloor} > m$，因此算法在最优情况下的时间复杂度为$O(b^{\left\lfloor\frac{m+1}{2}\right\rfloor})$。

14*. 参考答案：

(1) 令启发函数$h(n) \equiv 0$，则图搜索的 A*算法退化为 Dijkstra 算法。对于任意一个节点n，假设通过动作a能得到后继节点n'，当图中没有负值边，即单步代价$c(n,a,n') \geq 0$时，有$h(n') + c(n,a,n') = c(n,a,n') \geq 0 = h(n)$，可知恒为 0 的启发函数满足一致性。因此此时的图搜索 A*算法——Dijkstra 算法——是最优的。

(2) 跟据原状态转移图G，构造一个新的状态转移图G'，其中状态和状态转移关系不变，只对状态转移代价进行调整。新的单步代价定义为$c'(n,a,n') \coloneqq c(n,a,n') + h(n') - h(n)$，若启发函数满足一致性，则显然$c'(n,a,n') \geq 0$。根据(1)中结论，可知在 Dijkstra 算法在G'能够找到最优解。

首先证明，给定起始和终止状态s_1和s_k，则在图G'中找到的任意一条最短路径，必然也是G中的最短路径。考虑G'中的最短路径$n_1, n_2, \cdots, n_k$（按照本章中的定义，严格来说这不是一条路径，但通过这些节点的状态能找到一条路径。以下为了方便说明，将节点序列也称为路径。），其代价为

$$\begin{aligned}
&c'(n_1,a_1,n_2) + c'(n_2,a_2,n_3) + \cdots + c'(n_{k-1},a_{k-1},n_k) \\
&= [c(n_1,a_1,n_2) + h(n_2) - h(n_1)] + \cdots \\
&\qquad + [c(n_{k-1},a_{k-1},n_k) + h(n_k) - h(n_{k-1})] \\
&= c(n_1,a_1,n_2) + c(n_2,a_2,n_3) + \cdots + c(n_{k-1},a_{k-1},n_k) + h(n_k) - h(n_1)
\end{aligned}$$

当初始状态和终止状态给定时，$h(n_k) - h(n_1)$与路径本身无关，因此可以认为路径在G'中的代价是它在G中代价加上一个与路径无关的常数，因此G'中的最短路径必然也是G中的最短路径。

接着证明，给定起始和终止状态s_1和s_k，对于任意一个在G上搜索的 A*算法流程，必然存在一个在G'上的 Dijkstra 算法流程与其拥有相同的节点扩展顺序（称在不同算法中对应路径相同的节点为同一个节点），因此这两个算法会找到同一条从s_1到s_k的路径。不妨称G上的 A*算法为算法$A1$，G'上的 Dijkstra 算法为算法$A2$。对于$A1$搜索树中的任意节点n，可找到$A2$搜索树中对应相同路径的节点n'，根据上一段中的论证过程，可知$g(n') = g(n) + h(n) - h(n_1)$，其中$n_1$为根节点。那么算法$A1$（A*算法）中评价函数为$f(n) = g(n) + h(n)$，$A2$算法（Dijkstra 算法）中评价函数$f(n') = g(n')$，因此可得等式$f(n) = f(n') + h(n_1)$。即，当根节点确定时，算法$A1$和算法$A2$中对应同一路径的节点其评价函数值只相差一个常数。根据这个性质，不难用数学归纳法证明，对于任意一个算法$A1$可能导致的节点扩展顺序（此处强调任意是因为评价函数相同的节点扩展顺序可能不确定），都存在一个$A2$算法的扩展顺序与之相同。由于篇幅限制，此处省略具体的数学归纳法证明。

综上所述，给定起始和终止状态s_1和s_k，若 A*算法找到了一条路径P，则该路径必然也能在图G'上被 Dijkstra 算法找到。根据问题(1)中的结论，路径P是图G'上从s_1到s_k的最短路径。又根据上述证明，路径P也是图G上从状态s_1和s_k的最短路径，至此 A*算法的最优性得证。

第四章　监督学习

1. **A**

2. **B**

3. **F**

4. **C**

5. **A**

6. **A**

7. 参考答案（A 图是正确的）：
   1. 对于每个数据集，随着模型复杂度增大，模型在训练集上的错误率会不断下降，而在测试集上的错误率会先下降后上升。
   2. 随着模型复杂度增大，在更大的数据集 A 上模型更难拟合，因此也就不容易过拟合，但具有更好的泛化性。所以合理的猜测是，曲线（A, Train）会在（B, Train）的上方，曲线（A, Test）会在（B, Test）的下方，而曲线（A, Test）达到过拟合的转折点会比而曲线（B, Test）更靠后一些。

8. 略

10. 参考答案：
（1）该标准化项与参数w无关，该项对w的导数永远为 0，对w的优化求解没有作用。
（2）L2 标准化项通过惩罚过大的参数w来避免过拟合，λ小于 0 意味着该损失函数倾向于更大的w，从而激励过拟合，失去了标准化的作用。

11. 参考答案：
（1）心情指数大于 1 出去玩，等于 1 不出去玩。
（2）
迭代前：

1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

迭代后：

1/16	1/16	4/16	1/16	1/16	1/16	1/16	1/16	1/16	4/16

（3）有同伴就出去玩，没同伴不出去玩

$\frac{1}{2}\ln 3$

（4）三次迭代的分类器分别为：
C_1：心情指数大于 1 出去玩，等于 1 不出去玩
C_2：有同伴出去玩，没同伴不出去玩
C_3：天气好出去玩，天气不好不出去玩

强分类器可表示为：

$$sign(\frac{1}{2}\ln(4)\,C_1+\frac{1}{2}\ln(3)\,C_2+\frac{1}{2}\ln\left(\frac{17}{7}\right)C_3)$$

12*. 参考答案：

$$\log\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)}$$

$$\propto \log\frac{P(y=1)}{p(y=0)}+\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1)-\frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0)$$

$$\log\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)}$$

$$\propto \log\frac{P(y=1)}{p(y=0)}-\frac{1}{2}(\mu_1)^{\mathrm{T}}\Sigma^{-1}(\mu_1)+\frac{1}{2}(\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu_0)$$

$$+x^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_0)$$

$$=b+w^{\mathrm{T}}x$$

其中$b=\log\frac{P(y=1)}{P(y=0)}-\frac{1}{2}(\mu_1)^{\mathrm{T}}\Sigma^{-1}(\mu_1)+\frac{1}{2}(\mu_0)^{\mathrm{T}}\Sigma^{-1}(\mu_0)$是一个常量，$w=\Sigma^{-1}(\mu_1-\mu_0)$是一个线性参数向量。证毕。

第五章　无监督学习

1. **D**

2. **C**

3. **D**

4. **D**

5. **C**

6. 略

7. 参考答案：

在每一次聚类过程中，每个数据点都被分配到与其距离最近聚类中心所在集合，即 $\underset{c_{j\in C}}{\operatorname{argmin}}||x_i - c_j||^2$，因此在聚类步骤中$\sum_{i=1}^{K}\sum_{x\in G_i}||x - c_i||^2$中的每一项都保持不变或减小，仅当算法收敛时才会每一项都保持不变。因此，只要算法没有收敛，在每一次聚类后目标函数都会减小。

设一组数据点的聚类中心为z，则对于该组数据$\{x_i\}$，$\sum_{i=1}^{K}(x-z)^2 = \sum_{i=1}^{K}((x-\bar{x})^2 + 2(x-\bar{x})(\bar{x}-z) + (\bar{x}-z)^2) = \sum_{i=1}^{K}(x-\bar{x})^2 + 2(\bar{x}-z)\sum_{i=1}^{K}(x-\bar{x}) + \sum_{i=1}^{K}(\bar{x}-z)^2 = \sum_{i=1}^{K}(x-\bar{x})^2 + \sum_{i=1}^{K}(\bar{x}-z)^2$，当$z=\bar{x}$时取最小值。因此在更新聚类中心的步骤中，对每组数据取均值为新的中心将最小化每一个$\sum_{x\in G_i}||x - c_i||^2$。同样，只要算法没有收敛，在每一次更新聚类中心后目标函数都会减小。

综上，在算法收敛前目标函数都将严格递减。因此，k-means 算法不可能两次访问完全相同的聚类情况。而将有限数量的数据点分为 k 类的总可能是有限的，能保证不循环的 k-means 算法一定能够收敛。

8. 参考答案：

初始化两枚硬币的概率为 0.30 和 0.70

迭代次数	硬币 A 为正面次数	硬币 A 为反面次数	硬币 B 为正面次数	硬币 B 为反面次数	硬币 A 投掷正面概率 θ_A	硬币 B 投掷正面概率 θ_B
1	6.82	18.19	17.18	7.81	0.27	0.69
2	1.89	21.12	22.11	4.87	0.08	0.82
3	1.52	15.16	22.48	10.83	0.09	0.67

4	1.09	10.36	22.91	15.64	0.09	0.59
5	1.01	9.79	22.99	16.21	0.09	0.59

9. 参考答案：

k-means 算法可以被看做 EM 算法的一种特殊实现，其隐变量即为各聚类中心。在 E 步骤中，通过欧氏距离来估计各数据点最有可能归属于哪个聚类中心；在 M 步骤中，通过计算均值更新聚类中心位置来最大化这些数据点属于该聚类中心的可能性。

10. 参考答案:

a) 样本均值是对总体均值无偏估计的证明过程

证明：

$$E[\bar{x}] = E\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right]$$
$$= \frac{1}{n}E[x_1 + x_2 + \cdots + x_n]$$
$$= \frac{1}{n}(E[x_1] + E[x_2] + \cdots + E[x_n])$$
$$= \frac{1}{n}(\mu + \mu + \cdots + \mu)$$
$$= \frac{n\mu}{n}$$
$$= \mu$$

可见，样本均值是对总体均值的无偏估计。

b) $E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2\right] = \frac{n-1}{n}\sigma^2$

证明：

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left((\bar{x} - \mu) - (x_i - \mu)\right)^2\right]$$
$$= E\left[\frac{1}{n}\sum_{i=1}^{n}((\bar{x} - \mu)^2 - 2(\bar{x} - \mu)(x_i - \mu) + (x_i - \mu)^2)\right]$$
$$= E\left[(\bar{x} - \mu)^2 - \frac{2(\bar{x} - \mu)}{n}\sum_{i=1}^{n}(x_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

$$= E\left[(\bar{x}-\mu)^2 - 2(\bar{x}-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right] - E[(\bar{x}-\mu)^2]$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$= \frac{n-1}{n}\sigma^2$$

在上面的证明过程中，利用了如下的结果：

由于$E(X^2) = [E(X)]^2 + Var(X)$、$Var(a+X) = Var(X)$、$Var(aX) = a^2Var(X)$

因此

$$E[(\bar{x}-\mu)^2] = [E(\bar{x}-\mu)]^2 + Var(\bar{x}-\mu) = Var(\bar{x})$$

$$= Var\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

从$E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-x_i)^2\right] = \frac{n-1}{n}\sigma^2$可知，$\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-x_i)^2$是对总体方差的一个低估，因此在用样本方差来代替总体方差时，需要将样本方差定义为$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$，这才能从样本方差出发，得到总体方差$\sigma^2$，这是一个无偏估计。

答案：

$$\frac{n}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-x_i)^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}-x_i)^2\right] = \frac{n}{n-1}\times\frac{n-1}{n}\sigma^2 = \sigma^2$$

第六章 深度学习

1. **D**

2. **i 和 iii**

3. **D**

4. **D**

5. **C**

6. **C**

7. **A**

8. 参考答案：

数据：深度学习适合处理大数据，机器学习算法更适用于小数据；
硬件：深度学习由于巨大的计算量，需要大量计算资源，比如 GPU，机器学习算法对计算资源的需求相对较低；
特征构建：深度学习试图从数据中学习特征，机器学习中许多特征都需要由行业专家确定，并手工构造；
解决问题方式：深度学习通常利用“端到端”的方式构建模型，机器学习通常将问题分为几个步骤，每个步骤逐一解决，然后将结果组合。

9. 参考答案：

（1）根据图 6.2.可知，在异或操作中，点（0,0）与（1,1）为 0，点（0,1）和（1,0）为 1，由于感知机是线性分类器，无法构建超平面可以将这两类分开。
（2）输入层维度为 2，用于接收两个操作数；至少有一层隐藏层，且隐藏层中的神经元需要包含非线性激活函数；输出层维度为 1，用于输出结果。

10. 参考答案：

不能初始化为 0，也不能被同时初始化为其他相同的值。
如果参数被初始化为相同的值后，在误差反向传播过程中，同一层的神经元所接收到的误差都相同，更新后这些参数的值仍然相同。不管经过多少轮迭代，同一层神经元的参数保持相同，因此不同的神经元无法学习到不同特征的重要程度，失去了深度神经网络学习特征的能力。

11. 参考答案：

（1）$\mathrm{L} = \sum_{i=1}^{n+1} -y_i * logy'_i$

（2）$\mathrm{l}_{n+1} = -y_{n+1} * logy'_{n+1}$

$\mathrm{l}_n = -y_n * logy'_n$

$$\frac{\partial L}{\partial w_{n-1}} = \frac{\partial l_{n+1}}{\partial y'_{n+1}} * \frac{\partial y'_{n+1}}{\partial h_{n+1}} * \frac{\partial h_{n+1}}{\partial h_n} * \frac{\partial h_n}{\partial w_{n-1}} + \frac{\partial l_n}{\partial y'_n} * \frac{\partial y'_n}{\partial h_n} * \frac{\partial h_n}{\partial w_{n-1}}$$

12*. 参考答案：

RNN 的计算公式可以写为：

$$c_t = W \cdot [x_t, c_{t-1}] + b$$

假设时序长度为 T，则在梯度反向传播过程中，T 时刻的误差l_T对 W 的梯度为：

$$\frac{\partial l_T}{\partial W} = \sum_{i=1}^{T} \frac{\partial l_T}{\partial c_T} * (\prod_{j=i+1}^{T} \frac{\partial c_j}{\partial c_{j-1}}) * \frac{\partial c_T}{\partial W}$$

$\prod_{j=i+1}^{T} \frac{\partial c_j}{\partial c_{j-1}}$中每一项都时介于 0-1 之间的小数，当 T 很大时，$\prod_{j=i+1}^{T} \frac{\partial c_j}{\partial c_{j-1}}$值趋近于 0，从而产生梯度消失问题。

在 LSTM 中，$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$，因此$\frac{\partial c_t}{\partial c_{t-1}} = f_t$，由于$f_t$通常为 1 或者为 0，当为 1 时，所有梯度能够在 LSTM 中传递，当为 0 时，说明上一时刻的信息对当前时刻没有影响，因此没有必要传回梯度更新参数。这样，LSTM 解决了 RNN 的梯度消失问题。

第七章　强化学习

1. **A**

2. **D**

3. **A**

4. **B**

5. **B；C**

6. 参考答案：

根据公式（7.1.5）所示的价值函数的贝尔曼方程联立方程组：

$$\begin{cases} V_\pi(s_1) = R(s_1, 上, s_3) + \gamma V_\pi(s_3) = 0 + 0.99 \times V_\pi(s_3) \\ V_\pi(s_2) = R(s_2, 上, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times V_\pi(s_4) \\ V_\pi(s_3) = R(s_3, 上, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times V_\pi(s_d) \\ V_\pi(s_4) = 0 \\ V_\pi(s_d) = 0 \end{cases}$$

解得：

$$\begin{cases} V_\pi(s_1) = -0.99 \\ V_\pi(s_2) = 1 \\ V_\pi(s_3) = -1 \\ V_\pi(s_4) = 0 \\ V_\pi(s_d) = 0 \end{cases}$$

7. 参考答案

由于机器人每一步只能向上或者向右移动一个方格，首先计算状态s_3选择这两个不同动作后分别所得动作-价值函数取值：

$$q_\pi(s_3, 上) = \sum_{s' \in S} \mathrm{P}(s'|s_3, 上)\left[R(s_3, 上, s') + \gamma V_\pi(s')\right]$$

$$= 1 \times (-1 + 0.99 \times 0) + 0 \times \cdots = -1$$

$$q_\pi(s_3, 右) = \sum_{s' \in S} \mathrm{P}(s'|s_3, 右)\left[R(s_3, 右, s') + \gamma V_\pi(s')\right]$$

$$= 1 \times (1 + 0.99 \times 0) + 0 \times \cdots = 1$$

可见，智能体在状态s_3选择“向上移动一个方格”行动所得回报$q_\pi(s_3, 上)$值为-1、选择“向右移动一个方格”行动所得回报$q_\pi(s_3, 右)$值为1。显然，智能体在状态s_3应该选择“向右移动一个方格”行动，这样能够获得更大的回报。

于是，经过策略优化后，状态s_3处的新策略为$\pi'(s_3) = \text{argmax}_a q_\pi(s, a) =$右，则将$s_3$处的策略从“上”更新为 “右”。

8. 参考答案

1）第一次迭代：

$$q_\pi(s_1, 上) = R(s_1, 上, s_3) + \gamma V_\pi(s_3) = 0 + 0.99 \times 0 = 0$$

$$q_\pi(s_1, 右) = R(s_1, 右, s_2) + \gamma V_\pi(s_2) = 0 + 0.99 \times 0 = 0$$

$$\pi(s_1) = \text{argmax}_a q_\pi(s_1, a) = 上, \qquad V_\pi(s_1) = q_\pi(s_1, \pi(s_1)) = 0$$

$$q_\pi(s_2, 上) = R(s_2, 上, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$q_\pi(s_2, 右) = R(s_2, 右, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$\pi(s_2) = \text{argmax}_a q_\pi(s_2, a) = 上, \qquad V_\pi(s_2) = q_\pi(s_2, \pi(s_2)) = 1$$

$$q_\pi(s_3, 上) = R(s_3, 上, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$q_\pi(s_3, 右) = R(s_3, 右, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$\pi(s_3) = \text{argmax}_a q_\pi(s_3, a) = 右, \qquad V_\pi(s_3) = q_\pi(s_3, \pi(s_3)) = 1$$

此时的价值函数为：

1	0
0	1

2）第二次迭代：

$$q_\pi(s_1, 上) = R(s_1, 上, s_3) + \gamma V_\pi(s_3) = 0 + 0.99 \times 1 = 0.99$$

$$q_\pi(s_1, 右) = R(s_1, 右, s_2) + \gamma V_\pi(s_2) = 0 + 0.99 \times 1 = 0.99$$

$$\pi(s_1) = \text{argmax}_a q_\pi(s_1, a) = 上, \qquad V_\pi(s_1) = q_\pi(s_1, \pi(s_1)) = 0.99$$

$$q_\pi(s_2, 上) = R(s_2, 上, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$q_\pi(s_2, 右) = R(s_2, 右, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$\pi(s_2) = \text{argmax}_a q_\pi(s_2, a) = 上, \qquad V_\pi(s_2) = q_\pi(s_2, \pi(s_2)) = 1$$

$$q_\pi(s_3, 上) = R(s_3, 上, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$q_\pi(s_3, 右) = R(s_3, 右, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$\pi(s_3) = \text{argmax}_a q_\pi(s_3, a) = 右, \qquad V_\pi(s_3) = q_\pi(s_3, \pi(s_3)) = 1$$

此时的价值函数为：

1	0
0.99	1

3）第三次迭代：

$$q_\pi(s_1, 上) = R(s_1, 上, s_3) + \gamma V_\pi(s_3) = 0 + 0.99 \times 1 = 0.99$$

$$q_\pi(s_1, 右) = R(s_1, 右, s_2) + \gamma V_\pi(s_2) = 0 + 0.99 \times 1 = 0.99$$

$$\pi(s_1) = \text{argmax}_a q_\pi(s_1, a) = 上, \qquad V_\pi(s_1) = q_\pi(s_1, \pi(s_1)) = 0.99$$

$$q_\pi(s_2, 上) = R(s_2, 上, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$q_\pi(s_2, 右) = R(s_2, 右, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$\pi(s_2) = \text{argmax}_a q_\pi(s_2, a) = 上, \qquad V_\pi(s_2) = q_\pi(s_2, \pi(s_2)) = 1$$

$$q_\pi(s_3, 上) = R(s_3, 上, s_d) + \gamma V_\pi(s_d) = -1 + 0.99 \times 0 = -1$$

$$q_\pi(s_3, 右) = R(s_3, 右, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times 0 = 1$$

$$\pi(s_3) = \text{argmax}_a q_\pi(s_3, a) = 右, \qquad V_\pi(s_3) = q_\pi(s_3, \pi(s_3)) = 1$$

当迭代 3 次以后，价值函数收敛，算法终止，此时的价值函数为：

1	0
0.99	1

9. 参考答案：

根据算法 7.2.6 中的 Q 学习算法，s_1为初始状态，根据当前策略求出智能体应该采取的动作$a = \text{argmax}_a q_\pi(s_1, a) =$ 上，执行这个动作，得到奖励$R = 0$和进入下一状态$s' = s_3$，因此可如下更新对应的动作-价值函数：

$$q_\pi(s_1, \text{上}) \leftarrow q_\pi(s_1, \text{上}) + \alpha[R + \gamma \max_{a'} q_\pi(s', a') - q_\pi(s, a)]$$

$$= 0.1 + 0.5 \times [0 + 0.99 \times \max\{0, 0.1\} - 0.1] = 0.0995$$

此时的 q 函数为：

0.1/0	0/0
0.0995/0	0.1/0

同时令当前状态为s_3，此时智能体应该采取的动作$a = \text{argmax}_a q_\pi(s_3, a) =$ 上，执行这个动作，得到奖励$R = -1$和进入下一状态$s' = s_d$，因此可如下更新对应的动作-价值函数：

$$q_\pi(s_3, \text{上}) \leftarrow q_\pi(s_3, \text{上}) + \alpha[R + \gamma \max_{a'} q_\pi(s', a') - q_\pi(s, a)]$$

$$= 0.1 + 0.5 \times [-1 + 0.99 \times \max\{0, 0.1\} - 0.1] = -0.4005$$

此时算法达到终止状态s_d，该片段结束。此时的 q 函数为：

−0.4005/0	0/0
0.0995/0	0.1/0

此时每个状态的策略为：

→	
↑	↑

10. 参考答案：

函数：Sarsa
输入： 马尔可夫决策过程$MDP = (S, A, P, R, \gamma)$
输出： 策略π

```
随机初始化 q_π
repeat
    s ← 初始状态
    repeat
        a ← EpsGreedy(s, q_π, ε)
        执行动作 a，观察奖励 R 和下一个状态 s'
        a' ← EpsGreedy(s', q_π, ε)
        q_π(s,a) ← q_π(s,a) + α[R + γq_π(s',a') − q_π(s,a)]
        s ← s'
    until s 是终止状态
until q_π 收敛
π(s) := arg max_a q(s,a)
```

11*. 参考答案：

根据公式(7.3.3)，

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \mathbb{E}_{s,a\sim\pi}\left[q_{\pi_{\boldsymbol{\theta}}}(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - b(s) + b(s)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - b(s)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right] + \mathbb{E}_{s,a\sim\pi}[b(s)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)]$$

其中

$$\mathbb{E}_{s,a\sim\pi}[b(s)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)] = \mathbb{E}_{s\sim\pi}\left[\sum_a \pi_{\boldsymbol{\theta}}(s,a)b(s)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[b(s)\sum_a \pi_{\boldsymbol{\theta}}(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[b(s)\sum_a \pi_{\boldsymbol{\theta}}(s,a)\frac{\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(s,a)}{\pi_{\boldsymbol{\theta}}(s,a)}\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[b(s)\nabla_{\boldsymbol{\theta}}\sum_a \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s\sim\pi}[b(s)\nabla_{\boldsymbol{\theta}} 1]$$

$$= 0$$

因此

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - b(s)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

假设$h(s,a)$为 Actor-Critic 算法的基准函数，仿照上面的推导，

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \mathbb{E}_{s,a\sim\pi}\left[q_{\pi_{\boldsymbol{\theta}}}(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - h(s,a) + h(s,a)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - h(s,a)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right] + \mathbb{E}_{s,a\sim\pi}[h(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)]$$

其中，对$\mathbb{E}_{s,a\sim\pi}[h(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)]$中的$a$的期望进行展开，

$$\mathbb{E}_{s,a\sim\pi}[h(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)] = \mathbb{E}_{s\sim\pi}\left[\sum_a \pi_{\boldsymbol{\theta}}(s,a)h(s,a)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[\sum_a \pi_{\boldsymbol{\theta}}(s,a)h(s,a)\frac{\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(s,a)}{\pi_{\boldsymbol{\theta}}(s,a)}\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[\sum_a h(s,a)\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(s,a)\right]$$

$$= \mathbb{E}_{s\sim\pi}\left[\nabla_{\boldsymbol{\theta}}\sum_a h(s,a)\pi_{\boldsymbol{\theta}}(s,a)\right]$$

由于$h(s,a)$含有变量a，无法再同原推导一样将其放到求和符号外面。而$\mathbb{E}_{s\sim\pi}[\nabla_{\boldsymbol{\theta}}\sum_a h(s,a)\pi_{\boldsymbol{\theta}}(s,a)]$在一般情况下都不为 0，因此无法再得到形如

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \propto \mathbb{E}_{s,a\sim\pi}\left[\left(q_{\pi_{\boldsymbol{\theta}}}(s,a) - h(s,a)\right)\nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s,a)\right]$$

的式子，即$h(s,a)$不再适合作为基准函数。

第八章 人工智能博弈

1. **A**

2. **D**

3. **A**

4. **A**

5. **A**

6. **C**

7. 参考答案：

两位猎手同时猎鹿或同时猎兔都是纳什均衡解，这两个纳什均衡解是全合作（一起猎鹿）和完全不合作（独自猎兔）。不同的纳什均衡所有参与者的收益总和不一定相等。从囚徒困境的例子中可知，纳什均衡解未必一定是最优解。

8. 参考答案：

玩家 B 在两轮博弈后，遗憾值如下表：

玩家 B 在两轮后所得到的遗憾值

遗憾值\策略	**石头**	**剪刀**	**布**
第一轮	0	0	0
第二轮	2	1	0
$Regret_A^2$	2	1	0

按照遗憾匹配算法，第二轮过后，玩家 B 选择出石头的概率为$\frac{2}{3}$，选择出剪刀的概率为$\frac{1}{3}$，选择出布的概率为0。

9. 参考答案：

在当前策略下，初始节点为 2，行动序列路径$h = \{P \to B\}$产生的概率：

$$\pi_{-A}^{\sigma}(h) = \pi_B^{\sigma}(h) = 1 \times 0.5 = 0.5$$

由于在$\{2PB\}$节点选择加注和过牌的概率均为 50%，所以当前策略下从当前节点到达两个终结状态的概率分别为：

$$\pi^{\sigma}(h, z_1) = 0.5, \pi^{\sigma}(h, z_2) = 0.5$$

已知$u_A(z_1) = -1$和$u_A(z_2) = \frac{-2+2}{2} = 0$，故$v_A(\sigma, h) = -0.5$，同理，

$v_A\left(\sigma_{\{2PB\}\to B}, h\right) = 0$，故$r_A(\{2PB\}, B) = r_A(h, B) = 0.5$。

10*. 参考答案：

在价高者得的拍卖中，每次叫价之后的竞价都可以看成是博弈树的一颗子树，

但是由于不知道别人可以接受的最高价格是多少，在之后的竞价中通过猜测别人可接受的最高价格后所叫的价格是一种安全子博弈。

11. 参考答案：

使用 G-S 算法可以在一轮中得到稳定的匹配结果，即$(1,\mathrm{B}),(2,\mathrm{C}),(3,\mathrm{A})$，考虑 G-S 算法是男性向最喜欢的女性表白，可以对换表白与选择的角色，即女性向最喜欢的男性表白，可以得到另一种稳定的匹配结果，即$(2,\mathrm{A}),(3,\mathrm{B}),(1,\mathrm{C})$。此外，不难发现，$(1,\mathrm{A}),(2,\mathrm{B}),(3,\mathrm{C})$也是一种稳定的匹配结果，对比三种匹配结果，第一种对男性有利，每位男性都可以与自己最喜欢的女性结婚，而第二种匹配对女性有利，每位女性都可以与自己最喜欢的男性结婚，第三种相对双方较为均衡，不论男女都与自己第二喜爱的对象结婚。可以看出男性贪心表白，女性进行选择的匹配是一种男性占优的匹配，这也是 G-S 算法的特征。

12. 参考答案

周一：1；周二：3；周三：5；周四：4；周五：2；周六：6 周日：7