题型：（比去年难，因为去年是选修）（平时分50%，考试50%）

单选\*15（2分） 概念（ppt，均值函数，策略搜索，多智能体）

判断\*5（2分） 概念

计算\*2（10分） 出一些数据，写出步骤比如：值函数的计算（可能会给公式，不用计算器）（比如，折扣累计汇回报）

证明题\*2（10分） 所有证明题都是ppt上的，都是推导过的。（一定要注意上下标）（maybe值函数方法一个，随机策略方法一个）

算法论述题\*2（10分） （要考察两次论文里面的东西，需要找同学看一下其他组的ppt。还要加上一些常识）

出题原则：尽量覆盖所有知识点

复习：（全部都是重点，这里只是我听到的，不一定全）

1. 概述：

选择、判断

强化学习能解决的问题。训练过程。强化学习与其他机器学习的区别和联系。其他的优化方法（比较）。强化学习如何解决问题（这里的概念和例子，过程等等）。算法历史。强化学习的分类。强化学习的发展趋势。强化学习的路线图。在机器人中的应用（应该不考）。

1. 多臂赌博机：

大量概念（选择、判断、计算）

基本概念。多种策略。从多臂赌博机到马尔科夫决策过程。马尔科夫决策过程（马尔科夫性，基本概念）。折扣累计回报（很有可能是计算）。策略的定义（选择和判断）。值函数的定义。贝尔曼方程（将方程形式写成状态值函数的形式）。

1. 动态规划

……

策略评估（可能计算：给一张表，计算下一个状态的表）

伪代码中的公式。

策略改进（证明，容易出现上下标的问题）（计算，下一个状态的策略）

策略迭代和值迭代（算法论述题，两者之间的区别等）

1. MC方法

基本概念。

计算（给两条轨迹，用不同的方法计算值函数）

蒙特卡洛和动态规划，时间差分的区别和联系

蒙特卡洛策略改进（证明：on-policy mc策略改进证明）

On-policy和off-policy的区别，并且两种方法的例子（选择判断）

蒙特卡洛和动态规划，时间差分的区别和联系（我们的ppt两部分分开讲的）（可以比较方差和误差，产生的原因等）（算法论述）

1. 深度神经网络

Dqn（特点，伪代码，技巧，变种：在书上找变种）

1. 策略梯度

常见的直接策略搜索。（多种方法）

策略参数化（如何将策略参数化）

策略梯度（计算或者证明：需要自己推导一遍策略梯度的公式）

Reforce（证明：策略梯度是无偏的，但是方差很大等）（公式很重要）

1. TRPO及PPO

替代回报函数（证明）

带约束的优化问题（证明：……和为什么是单调递减的）

优化参数化的策略

……

共轭梯度法搜索可行方向

……

1. 随机策略

Ac框架相容函数（证明）

随机策略与确定性策略比较

确定性AC算法

确定性策略梯度算法中如何评估行为值函数？

相容函数

Ddpg，d4pg

多智能体强化学习（证明、选择判断、算法论述多与单的区别等等）

博弈论（概念）

那什均衡（精准理解概念，计算那什均衡点）

随机博弈

多智能体