**МИНИСТЕРСТВО науки и высшего образования Российской Федерации**

федеральное Государственное бюджетное образовательное учреждение высшего образования

**«ТЮМЕНСКИЙ ИНДУСТРИАЛЬНЫЙ университет»**

|  |
| --- |
| **Утверждаю:**  \_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_» \_\_\_\_ |

**ФОНД ОЦЕНОЧНЫХ СРЕДСТВ**

**для проверки сформированности компетенций**

|  |  |
| --- | --- |
| дисциплины: | **Обучение с подкреплением** |
| направление подготовки: | **01.04.02 Прикладная математика и информатика** |
| направленность (профиль): | **Машинное обучение и анализ данных** |
| форма обучения: | **Очная, очно-заочная, заочная** |

Фонд оценочных средств рассмотрен на заседании кафедры бизнес-информатики и математики

Протокол № \_\_\_ от «\_\_\_» \_\_\_ \_\_\_\_\_\_\_

**ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ И ИНДИКАТОРОВ ИХ ДОСТИЖЕНИЯ ПО**

**УЧЕБНОЙ ДИСЦИПЛИНЕ (МОДУЛЮ)**

|  |  |
| --- | --- |
| Код и наименование компетенции | Код и наименование индикатора достижения компетенции (ИДК) |
| ПКС-2 Способен разрабатывать (совершенствовать) и внедрять новые методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными | ПКС-2.1. Способен разрабатывать методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными |
|
|
| ПКС-2.2. Способен оптимизировать и внедрять методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными в профессиональной деятельности |
|
|

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕРКИ СФОРМИРОВАННОСТИ КОМПЕТЕНЦИЙ**

***Для оценки*** ПКС-2 *Способен разрабатывать (совершенствовать) и внедрять новые методы, модели, алгоритмы, технологии и инструментальные средства работы с большими данными*

**10 тестовых вопросов открытого типа с ответами по дисциплине "Обучение с подкреплением"**

1. Что такое обучение с подкреплением (Reinforcement Learning)?

Ответ:

Обучение с подкреплением — это метод машинного обучения, в котором агент учится взаимодействовать с окружающей средой. Цель агента — находить оптимальную стратегию поведения, которая максимизирует ожидаемую кумулятивную награду на основе обратной связи от среды.

2. Что такое функция ценности (Value Function) в обучении с подкреплением?

Ответ:

Функция ценности (Value Function) оценивает, насколько хороша текущая стратегия или действие в состоянии с точки зрения ожидаемой будущей награды. Она отражает "ценность" нахождения агента в конкретном состоянии или выполнения конкретного действия.

3. Объясните разницу между обучением с учителем (Supervised Learning) и обучением с подкреплением (Reinforcement Learning).

Ответ:

- В обучении с учителем агент учится на основе заранее размеченных данных, где для каждого входа известен правильный ответ.

- В обучении с подкреплением агент изучает оптимальную стратегию через взаимодействие со средой, не обладая заранее подготовленными "правильными ответами". Вместо этого агент получает награду (reward) или штраф (penalty) за свои действия.

4. Что такое Markov Decision Process (MDP), и почему он важен в обучении с подкреплением?

Ответ:

Марковский процесс принятия решений (MDP) — это математическая модель задач обучения с подкреплением. Он описывает агента, состояние среды, множества действий, вероятность переходов между состояниями и распределение наград. Важность MDP заключается в том, что он формализует взаимодействие агент-среда, позволяя создавать алгоритмы для поиска оптимальных стратегий.

5. Что такое политика (Policy) в контексте обучения с подкреплением?

Ответ:

Политика (Policy) — это стратегия агента, которая определяет, какое действие следует выполнять в каждом состоянии. Политика может быть детерминированной (каждому состоянию соответствует одно действие) либо стохастической (действия выбираются с определенными вероятностями).

6. Что такое Q-функция или функция действия, и как она используется в алгоритмах Q-Learning?

Ответ:

Q-функция (или функция ценности действия) оценивает ожидаемую суммарную награду, которую агент получит, начав в определенном состоянии, выбрав конкретное действие и затем следуя оптимальной политике. В Q-Learning Q-функция обновляется на каждом шаге для приближения оптимальной стоимости действий.

7. Какие основные отличия между Model-Based и Model-Free методами обучения с подкреплением?

Ответ:

- Model-Based методы: строят модель среды (например, вероятности переходов и награды) и используют её для планирования оптимальной стратегии.

- Model-Free методы: не строят явной модели среды, а изучают оптимальную стратегию напрямую через взаимодействие со средой.

8. Объясните суть DQN (Deep Q-Network) и его отличие от классического Q-Learning.

Ответ:

DQN (глубокие Q-сети) — это улучшенный вариант Q-Learning, использующий нейронные сети для аппроксимации Q-функции, вместо хранения Q-таблиц. Основное отличие — способность DQN эффективно работать в задачах с огромным пространством состояний и действий благодаря используемой нейросети.

9. Что такое эксплорация и эксплуатация (exploration и exploitation) в RL? Почему важен баланс между ними?

Ответ:

- Эксплорация (Exploration): исследование новых действий, которые агент ранее не выполнял, чтобы узнать о среде.

- Эксплуатация (Exploitation): использование уже изученных действий, которые приносят максимальную награду.

Баланс важен, так как чрезмерная эксплуатация может привести к игнорированию потенциально лучших действий, а чрезмерная эксплорация замедлит обучение оптимальной стратегии.

10. Опишите принцип работы алгоритма Policy Gradient.

Ответ:

Алгоритм Policy Gradient оптимизирует стратегию напрямую, без создания Q-функции, путем изменения параметров стохастической политики. Цель — максимизировать ожидаемую суммарную награду, используя градиентный метод для обновления параметров политики в направлении увеличения награды.

**10 тестовых вопросов закрытого типа**

**(Выберите один правильный ответ)**

1. Что обозначает термин "агент" в обучении с подкреплением?

a) Среда, в которой происходят действия

b) Алгоритм, взаимодействующий со средой \*(Верный ответ)\*

c) Источник наград

d) Действия, выполняемые системой

2. Как называется метод, позволяющий агенту делать выбор между исследованием среды и использованием уже изученной стратегии?

a) Q-Learning

b) Policy Gradient

c) Баланс эксплорации и эксплуатации \*(Верный ответ)\*

d) Bellman Equation

3. Какова основная цель функции награды (reward function) в RL?

a) Генерация случайных данных

b) Показать агенту, насколько хорошо он выполняет действия \*(Верный ответ)\*

c) Описать возможные действия агента

d) Определить структуру политики

4. Какой метод RL использует нейронные сети для аппроксимации функции ценности?

a) TD-Learning

b) DQN \*(Верный ответ)\*

c) SARSA

d) Markov Decision Process

5. Как определяется "оптимальная политика" в задачах RL?

a) Политика, которая минимизирует время выполнения задачи

b) Политика, которая соответствует случайным действиям

c) Политика, максимизирующая ожидаемую награду \*(Верный ответ)\*

d) Политика, использующая максимальное количество ресурсов

6. Какая из этих задач лучше всего подходит для обучения с подкреплением?

a) Классификация изображений

b) Прогнозирование временных рядов

c) Игра в шахматы \*(Верный ответ)\*

d) Создание текстового отчета

7. Что означает термин "эпизод" в контексте RL?

a) Весь процесс обучения агента

b) Последовательность состояний и действий до достижения терминального состояния \*(Верный ответ)\*

c) Одно действие агента

d) Структура среды

8. В чем отличие алгоритма Q-Learning от SARSA?

a) SARSA оптимизирует случайные действия

b) Q-Learning максимизирует награды в реальном времени

c) SARSA учитывает выбранные агентом действия, Q-Learning — предполагает оптимальные действия \*(Верный ответ)\*

d) В Q-Learning используются только нейронные сети

9. Чем отличается Model-Based RL от Model-Free RL?

a) Model-Based RL строит модель среды и использует её для планирования \*(Верный ответ)\*

b) Model-Based RL не взаимодействует со средой

c) Model-Free RL требует больше данных о среде

d) Model-Based RL работает только с Q-функциями

10. Что такое Bellman Equation в обучении с подкреплением?

a) Уравнение для вычисления ожидаемой награды от действий \*(Верный ответ)\*

b) Описание механизма обучения нейронной сети

c) Метод построения модели среды

d) Алгоритм выбора действий в реальном времени

**10 тестовых вопросов комбинированного типа**

**(Выберите один или несколько правильных ответов и/или дайте краткий ответ)**

1. Какие компоненты входят в обучение с подкреплением?

a) Агент \*(Верный ответ)\*

b) Среда \*(Верный ответ)\*

c) Система мониторинга

d) Награда \*(Верный ответ)\*

2. Назовите цели политики (policy) в RL.

Ответ:

Политика определяет стратегию агента, выбирая действие для каждого состояния с целью максимизации суммарной награды.

3. Какие типы масштабируемых алгоритмов часто используются в RL?

a) Policy Gradient \*(Верный ответ)\*

b) Q-Learning \*(Верный ответ)\*

c) Backpropagation

d) Deep Q-Network \*(Верный ответ)\*

4. Какие отличительные признаки есть у задач обучения с подкреплением?

a) Наличие функции награды \*(Верный ответ)\*

b) Прямая разметка данных

c) Взаимодействие агента со средой \*(Верный ответ)\*

d) Использование больших обучающих выборок

5. Какие основные различия между стохастической и детерминированной политикой?

Ответ:

- Стохастическая политика: действия выбираются с определенной вероятностью, что позволяет исследовать среду.

- Детерминированная политика: каждое состояние однозначно определяет действие.

6. Какой из перечисленных алгоритмов RL относится к классу Model-Free методов?

a) Q-Learning \*(Верный ответ)\*

b) Policy Gradient \*(Верный ответ)\*

c) Monte Carlo Methods \*(Верный ответ)\*

d) SARSA \*(Верный ответ)\*

7. В каких задачах используется модель среды (Model-Based RL)?

a) Планирование маршрутов \*(Верный ответ)\*

b) Оптимизация систем управления \*(Верный ответ)\*

c) Прогнозирование погодных условий

d) Выбор текущих решений в неизвестной среде

8. Что такое переобучение в контексте RL, и как его избежать?

Ответ:

Переобучение в RL означает, что агент запоминает конкретные последовательности действий вместо обобщения стратегии. Чтобы избежать этого, используют случайное инициирование состояний, регуляризацию или смешивание эпизодов.

9. Какие стратегии используются для управления балансом между эксплорацией и эксплуатацией?

a) Epsilon-Greedy \*(Верный ответ)\*

b) Softmax \*(Верный ответ)\*

c) Backtracking

d) Upper Confidence Bound (UCB) \*(Верный ответ)\*

10. Какие улучшения предлагает DQN перед традиционным Q-Learning?

a) Использование нейросети для аппроксимации Q-функции \*(Верный ответ)\*

b) Оптимизация политики напрямую

c) Использование метода Replay Buffer для хранения опыта \*(Верный ответ)\*

d) Увеличение вероятности случайного выбора действий

**Критерии оценивания уровня сформированности компетенций**

Оценка «отлично» выставляется обучающемуся, если им даны правильные ответы более, чем на 91% вопросов.

Оценка «хорошо» выставляется обучающемуся, если им даны правильные ответы на 76-90% вопросов.

Оценка «удовлетворительно» выставляется обучающемуся, если им даны правильные ответы на 61-75% вопросов.

Оценка «неудовлетворительно» выставляется обучающемуся, если им даны правильные ответы менее чем на 60% вопросов.