Aprendizaje por Reforzamiento

10/27/2025

1. Información General

- **CIP:**
- **Carga Académica:**
- **Requisitos:**
- **Profesor:** Diego Ascarza (diego.ascarza@tec.mx)
- **Horario:** Jueves de 18:30 a 21:00

2. Intención del curso

La intención de este curso es introducir a los estudiantes las herramientas y conceptos básicos del aprendizaje por reforzamiento (RL). El aprendizaje por reforzamiento se enfoca en la toma de decisiones en ambientes dinámicos e inciertos. En el curso se definirán la clase de problemas para los cuales RL está diseñado. Estos problemas incluyen aquellos donde se realizan decisiones secuenciales y aprendizaje a partir de la interacción con el ambiente. El curso también proveerá insights teóricos y aplicados para entender cómo los algoritmos asociados a RL operan y dónde suelen ser más efectivos. Finalmente, el curso pondrá énfasis en aplicaciones a la economía, en particular, a modelos macroeconómicos dinámicos y estocásticos.

3. Fin del aprendizaje

Al finalizar este curso, los estudiantes podrán:

- Modelar procesos de decisión Markoviana finitos (MDPs) e identificar los métodos de RL apropiados para resolverlos.
- Entender e implementar algoritmos básicos del RL tales como aquellos asociados a programación dinámica, métodos de Monte Carlo, temporal difference y policy gradient.
- Aplicar RL a problemas que involucren comportamiento adaptativo, control óptimo con un énfasis en temas de economía.

4. Temas y subtemas

1. Introducción
    a) Elementos del RL.
    b) Ejemplos de aplicación de métodos de RL.
    c) Alcances y limitaciones.
    d) Un ejemplo extendido: Tic-Tac-Toe.
2. Multi-Armed Bandits
    a) El problema del *k-armed bandit*.
    b) Métodos de acción-valor.
    c) Implementación incremental.
    d) Tracking de un problema no-estacionario.
    e) Valores iniciales optimistas.
    f) Cotas superiores de confianza de selección de acciones.
    g) Algoritmos basados en gradiente para bandit problems.
    h) Búsqueda asociativa (bandits contextuales).
3. Procesos de decisión Markovianos finitos
    a) La interfase Agent-Environment.
    b) Objetivos y rewards.
    c) Retornos y episodios.
    d) Políticas y funciones de valor.
    e) Políticas óptimas y funciones de valor óptimas.
    f) Optimalidad y aproximación.
4. Programación Dinámica
    a) Evaluación de política (predicción).
    b) Mejora de política.
    c) Iteración de política.
    d) Iteración de valor.
    e) Programación dinámica asincrónica.
    f) Iteración de política generalizada.
    g) Eficiencia de la programación dinámica.
5. El Modelo de Crecimiento Neoclásico
    a) Ambiente.
    b) Preferencias.

   c) Tecnología.
   d) Mercados y equilibrio competitivo.
   e) El problema del planificador social.
   f) Representación recursiva del problema del planificador social.
   g) Resolviendo el problema del planificador social con iteración de la función de valor.

6. Los métodos de Monte Carlo

   a) Predicción.
   b) Estimación de valor de acción.
   c) Control.
   d) Control sin inicio exploratorio.
   e) Predicción off-policy con importance sampling.
   f) Implementación incremental.
   g) Control en off-policy Montecarlo.

7. Aprendizaje por diferencia temporal (TD).

   a) Predicción en TD.
   b) Ventajas de los métodos de TD.
   c) Optimalidad del TD(0).
   d) Sarsa: Control en TD on-policy.
   e) Q-learning: Control TD off-policy.
   f) Sarsa esperado.
   g) Maximización de sesgo y aprendizaje doble.
   h) Juegos, afterstates y otros casos especiales.

8. Bootstrapping en n etapas.

   a) Predicción en TD en n etapas.
   b) Sarsa en n etapas.
   c) Aprendizaje off-policy en n etapas.
   d) Aprendizaje off-policy sin importance sampling.

5. Objetivos específicos de aprendizaje por tema

1. Presentar las ideas fundamentales del RL, incluyendo el aprendizaje mediante prueba y error, la presencia de recompensas diferidas y la dinámica de interacción entre un agente y su entorno. Los estudiantes comprenderán estos conceptos básicos y podrán identificar una amplia gama de aplicaciones potenciales, especialmente en el ámbito económico.

2. Explicar el trade-off entre exploración y explotación en problemas de decisión secuencial, enfatizando su importancia en RL Ilustrar estos conceptos mediante modelos de multi-armed bandits, desarrollando la intuición sobre métodos basados en valores de acción y demostrando que es posible aprender incluso en entornos simples sin estados definidos.

3. Entender el framework formal utilizado en RL: el proceso de decisión Markoviano. Aprender a definir estados, acciones, transiciones, rewards y cómo las políticas y las funciones de valor se relacionan con el desempeño de largo plazo.

4. Comprender el marco formal que sustenta el aprendizaje por refuerzo: los procesos de decisión de Markov (MDP). Aprender a definir correctamente los elementos clave —estados, acciones, transiciones y recompensas— y entender cómo las políticas y funciones de valor determinan el desempeño a largo plazo del agente.

5. Analizar el modelo de crecimiento neoclásico tanto en su forma secuencial como en su representación recursiva mediante programación dinámica. Aplicar los métodos de iteración de la función de valor y de iteración de políticas para resolver el problema del planificador social, comprendiendo su estructura como un problema de decisión intertemporal.

6. El objetivo es capacitar a los estudiantes en el uso de técnicas que permiten aprender funciones de valor a partir de muestras de episodios, sin requerir conocimiento previo del entorno o modelo. Se enfatizará cómo los métodos Monte Carlo pueden aplicarse en contextos donde solo se observa el resultado de trayectorias completas.

7. Comprender la lógica y el funcionamiento del aprendizaje por TD, destacando cómo combina elementos de los métodos Monte Carlo y la programación dinámica. Aprender a implementar algoritmos clave como TD(0), Sarsa y Q-learning y analizar su utilidad en sistemas dinámicos que generan retroalimentación continua.

8. Profundizar en el análisis del compromiso entre sesgo y varianza en la estimación de funciones de valor mediante métodos de múltiples etapas. Los estudiantes explorarán cómo los rollouts de longitud intermedia pueden aumentar la eficiencia y estabilidad del aprendizaje, con aplicaciones relevantes en tareas de predicción y planificación dentro de contextos económicos.

6. Metodología de enseñanza y actividades de aprendizaje sugeridas

6.1. Actividades de aprendizaje bajo la conducción del profesor:

1. Lecturas dirigidas y discusión en clase fomentando la comprensión profunda y rigurosa de los tópicos de la materia.

2. Horas de oficina para resolver preguntas asociadas a lo discutido en la materia.

6.2. Actividades de aprendizaje independiente:

1. Actualización constante de aplicaciones de RL a la política pública.

2. Participación en seminarios y conferencias impartidas por expertos en la materia.

7. Tiempo estimado por tema

- Tema 1: 2 horas
- Tema 2: 3 horas
- Tema 3: 3 horas
- Tema 4: 4 horas
- Tema 5: 3 horas
- Tema 6: 3 horas
- Tema 7: 3 horas
- Tema 8: 3 horas
- Tema 9: 3 horas
- Tema 10: 3 horas

Total: 30 horas

8. Políticas de evaluación sugeridas

Los alumnos realizarán tres tareas y una presentación final. Las tareas son individuales. Las tareas se entregan cada tres semanas. El trabajo y la presentación final serán grupales (en grupos de 3 alumnos). Las ponderaciones son como siguen:

- Tareas (60 %)..
- Presentación final (40 %).

9. Bibliografía sugerida

El libro de texto que se utilizará como referencia es Reinforcement Learning: an Introduction (Sutton y Barton 2da edición). Asimismo, el profesor proveerá de diapositivas a los alumnos con contenido basado en dicho libro de texto.

10. Perfil del profesor

Doctor en Economía por la Universidad de Minnesota, Estados Unidos. El profesor Ascarza trabaja en la intersección entre la Macroeconomía, Economía de la Salud y Finanzas Públicas. Sus temas de investigación explotan métodos empíricos y estructurales para estudiar temas de Seguridad Social, Reformas de Sistemas de Pensiones, Salud Mental y Mercados Inmobiliarios.

11. Política de uso de inteligencia artificial

El uso de herramientas de inteligencia artificial (IA) puede ser un recurso valioso para apoyar el aprendizaje, siempre que se utilice de manera ética y responsable. Para esta clase:

11.1. Uso permitido

Los estudiantes pueden emplear IA para tareas de apoyo como:

- Generar ideas iniciales de código o pseudocódigo.
- Consultar explicaciones sobre conceptos de aprendizaje por reforzamiento.
- Depurar errores básicos en programación.
- Revisar bibliografía y obtener resúmenes de artículos relevantes.

11.2. Uso restringido

- No se permite presentar como propio código, ensayos o resultados generados íntegramente por IA sin modificaciones ni análisis crítico.
- Los estudiantes deben citar explícitamente cuando usen ideas, fragmentos de texto o código sugeridos por IA.
- Se espera que cualquier uso de IA vaya acompañado de una reflexión personal que muestre comprensión del material.

11.3. Objetivo pedagógico

- El propósito de la clase es que el estudiante desarrolle su propio razonamiento y habilidades técnicas en aprendizaje por reforzamiento.
- El uso de IA debe servir como andamiaje de aprendizaje, no como sustituto del esfuerzo individual.

11.4. Evaluaciones

- En tareas individuales donde se indique explícitamente, queda prohibido el uso de IA.
- El incumplimiento de esta política se considerará una falta de integridad académica.

11.5. Recomendación

- Se invita a los estudiantes a documentar en un anexo breve cómo usaron IA (si la usaron) en cada entrega, para fomentar la transparencia.