Universidad Nacional de Ingeniería (UNI)
Escuela Profesional de
Ciencia de la Computación
Sílabo 2027-I

1. CURSO

AI265. Aprendizaje por Refuerzo (Electivo)

2. INFORMACIÓN GENERAL

2.1 Curso	:	AI265. Aprendizaje por Refuerzo
2.2 Ciclo	:	CS300 Ciclo.
2.3 Créditos	:	4
2.4 Horas	:	2 HT; 4 HP;
2.5 Duración del periodo	:	16 semanas
2.6 Condición	:	Electivo
2.7 Modalidad de aprendizaje	:	Presencial
2.8 Prerrequisitos	:	AI263. Introducción al Aprendizaje de Máquina. (6^{to} Sem)

3. PROFESORES

Atención previa coordinación con el profesor

4. INTRODUCCIÓN AL CURSO

El aprendizaje por refuerzo (RL) es fundamental para sistemas autónomos. Este curso cubre algoritmos clave como Q-Learning, DQN y Policy Gradients, con aplicaciones en robótica y juegos.

5. OBJETIVOS

- Diseñar agentes RL para entornos estocásticos.
- Implementar algoritmos RL usando frameworks modernos (OpenAI Gym, TensorFlow).
- Analizar problemas éticos en sistemas autónomos basados en RL.

6. RESULTADOS DEL ESTUDIANTE

AG-C08) Análisis de Problemas: Identifica, formula y analiza problemas complejos de computación. (Usage)

AG-C09) Diseño y Desarrollo de Soluciones: Diseña, implementa y evalúa soluciones para problemas complejos de computación. (Usage)

7. TEMAS

Unidad 1: Fundamentos de RL (18 horas)	
Resultados esperados: AG-C08,AG-C09	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Introducción al paradigma del Aprendizaje por Refuerzo: agente, entorno, acciones, estados, recompensas. ▪ Formalización de problemas secuenciales: Procesos de Decisión Markovianos (MDPs) y Procesos de Decisión Markovianos Parcialmente Observables (POMDPs). ▪ Concepto de horizonte y factor de descuento (gamma). ▪ La Ecuación de Bellman: fundamento de optimalidad para funciones valor estado (V) y acción-estado (Q). ▪ Métodos de Programación Dinámica: Evaluación y Mejora Iterativa de Políticas. ▪ Métodos de Monte Carlo: aprendizaje a partir de episodios completos. ▪ Algoritmos de diferencia temporal (TD): TD(0), SARSA, y el balance entre bias y varianza. ▪ Q-Learning (off-policy): derivación, propiedades de convergencia y parámetros (learning rate, epsilon-greedy).	▪ Modelar problemas como MDPs [Evaluar] ▪ Implementar Q-Learning [Usar]
Lecturas : [SB18]	

Unidad 2: Métodos Avanzados de RL Clásico (14 horas)	
Resultados esperados: AG-C08,AG-C09	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Limitaciones de los métodos tabulares y la maldición de la dimensionalidad. ▪ Aproximación de funciones valor: sustitución de tablas por modelos paramétricos (lineales, no lineales). ▪ Algoritmos de Gradiente Temporal: minimización del error de Bellman. ▪ Eligibilidad Traces: unificación de Monte Carlo y TD (TD(δ)). ▪ Planificación y Aprendizaje Integrados: Modelos del entorno y algoritmos Dyna-Q. ▪ Métodos basados en Políticas (Policy Search): introducción al gradiente de política. ▪ Balance entre exploración y explotación: estrategias avanzadas (UCB, exploración basada en entropía).	▪ Modelar problemas como MDPs [Evaluar] ▪ Implementar Q-Learning [Usar]
Lecturas : [SB18]	

Unidad 3: RL Profundo (Deep RL) (20 horas)	
Resultados esperados: AG-C08,AG-C09	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Revolución del RL con Deep Learning: estabilización del entrenamiento. ▪ Deep Q-Networks (DQN): arquitectura, función de pérdida, y mecanismos de estabilización (Target Network, Experience Replay). ▪ Extensiones de DQN: Double DQN, Dueling DQN, Prioritized Experience Replay. ▪ Política Estocástica en espacios continuos: introducción a los métodos Actor-Critic. ▪ Actor-Critic Methods: descomposición del gradiente, función ventaja (A2C). ▪ Métodos de Actor-Critic Avanzados: A3C (Asynchronous Advantage Actor-Critic) y su arquitectura paralela. ▪ Policy Gradient directos: REINFORCE con y sin baseline. ▪ Introducción a métodos de máxima entropía (Soft Actor-Critic - SAC) y aprendizaje offline.	▪ Entrenar agentes DQN en entornos complejos [Evaluar] ▪ Optimizar políticas con A3C [Usar]
Lecturas : [Mni+15], [Sil+17]	

Unidad 4: Ética, Seguridad y Aplicaciones en RL (12 horas)	
Resultados esperados: AG-C08,AG-C09	
Temas	**Aprendizaje esperado (*Learning Outcomes*)**
▪ Sesgos en el diseño de funciones de recompensa: especificación incorrecta, recompensas no alineadas y optimización de recompensas por hackeo (reward hacking). ▪ Desafíos de la generalización y robustez en agentes entrenados con RL. ▪ Seguridad en sistemas autónomos: evitación de fallos catastróficos, aprendizaje seguro (Safe RL) y alineación (AI Alignment). ▪ Casos de estudio: análisis de fallos en simulaciones y robots reales. ▪ Ataques adversarios en RL: ataques a los sensores, a la función de política y a la función de recompensa del agente. ▪ Implicaciones sociales: automatización, toma de decisiones y transparencia (agentes interpretables). ▪ Marcos éticos para el desarrollo y despliegue de agentes autónomos.	▪ Identificar riesgos éticos en aplicaciones de RL [Familiarizarse] ▪ Proponer estrategias de mitigación [Usar]
Lecturas : [SB18]	

8. PLAN DE TRABAJO

8.1 Metodología

Se fomenta la participación individual y en equipo para exponer sus ideas, motivándolos con puntos adicionales en las diferentes etapas de la evaluación del curso.

8.2 Sesiones Teóricas

Las sesiones de teoría se llevan a cabo en clases magistrales donde se realizarán actividades que propicien un aprendizaje activo, con dinámicas que permitan a los estudiantes interiorizar los conceptos.

8.3 Sesiones Prácticas

Las sesiones prácticas se llevan en clase donde se desarrollan una serie de ejercicios y/o conceptos prácticos mediante planteamiento de problemas, la resolución de problemas, ejercicios puntuales y/o en contextos aplicativos.

9. SISTEMA DE EVALUACIÓN

Sesiones Teóricas:
Las sesiones de teoría se llevan a cabo en clases magistrales donde se realizarán actividades que propicien un aprendizaje activo, con dinámicas que permitan a los estudiantes interiorizar los conceptos.

Sesiones de Laboratorio:
Para verificar que los alumnos hayan alcanzado el logro planteado para cada una de las unidades de aprendizaje, realizarán actividades que les permita aplicar los conocimientos adquiridos durante las sesiones de teoría y se les propondrá retos que permitan evaluar el desempeño de los alumnos.

Exposiciones individuales o grupales:
Se fomenta la participación individual y en equipo para exponer sus ideas, motivándolos con puntos adicionales en las diferentes etapas de la evaluación del curso.

Lecturas:
A lo largo del curso se proporcionan diferentes lecturas, las cuales podrían ser evaluadas. El uso de herramientas Online permite a cada estudiante acceder a la información del curso, e interactuar fuera de aula con el profesor y con los otros estudiantes.

La nota final NF depende de de las siguientes notas:

- EP: Examen Parcial.
- EF: Examen Final.
- P: Evaluacion permanente (Promedio de prácticas y Trabajos calificados).

La nota final NF se calcula así:

$$NF = \frac{EP + EF + P}{3}$$

10. BIBLIOGRAFÍA BÁSICA

[Mni+15] Volodymyr Mnih et al. ■Human-level control through deep reinforcement learning■. En: *Nature* 518 (2015), págs. 529-533.

[Sil+17] David Silver et al. ■Mastering the game of Go without human knowledge■. En: *Nature* 550 (2017), págs. 354-359.

[SB18] Richard Sutton y Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.