Reinforcement Learning - Laboratorio 4 -

Estefanía Elvira, Cayetano Molina, Priscilla González

Task 1

Responda a cada de las siguientes preguntas de forma clara y lo más completamente posible.

1. **¿Cómo afecta la elección de la estrategia de exploración (exploring starts vs soft policy) a la precisión de la evaluación de políticas en los métodos de Monte Carlo?**

   Exploring Starts: lo que se realiza acá inicia de manera aleatoria cada par de estado y acción para que así cada acción logre ser explorada por cada diferente estado que se tenga.

   Soft Policy: acá no se elige una acción de forma específica por lo que se permite cierto grado de exploración.

   Ahora bien, sabiendo un poco de la definición, el exploring starts lo que hace es brindar una estimación más precisa de cada estado y acción para que las acciones sean mejor exploradas por los diferentes estados. La estrategia de soft policy hace que las precisiones o evaluaciones sean mucho más realistas. Pero también existe la posibiludad de que esta estrategia no logre explorar los estados y acciones de forma adecuada por lo que la evaluación podría ser menos precisa.

   a. **Considere la posibilidad de comparar el desempeño de las políticas evaluadas con y sin explorar los inicios o con diferentes niveles de exploración en políticas blandas.**

      Esto va a depender mucho del contexto en el que se esté trabajando. Una exploring starts puede ser ventajosa al tener una mejor cobertura y una mejor precisión, pero puede ser que en otro contexto el proceso de exploración no sea lo más óptimo por lo que la soft policy puede convenir más ya que se adapta más a este tipos de comportamiento.

2. **En el contexto del aprendizaje de Monte Carlo fuera de la póliza, ¿cómo afecta la razón de muestreo de importancia a la convergencia de la evaluación de políticas? Explore cómo la razón de muestreo de importancia afecta la estabilidad y la convergencia.**

   La razón de muestreo de importancia se usa en el aprendizaje de Monte Carlo fuera de la póliza para corregir la distribución de las muestras cuando se evalúa una política objetivo diferente a la política comportamental que generó los datos. En términos simples, se calcula como la probabilidad de una secuencia de acciones bajo la política objetivo dividida por la probabilidad de la misma secuencia bajo la política comportamental.

   Impacto en la Convergencia y la Estabilidad

   - **Varianza de las Estimaciones:** El uso de la razón de muestreo de importancia puede introducir alta varianza en las estimaciones de la política objetivo. Esto ocurre especialmente cuando las políticas objetivo y comportamental son muy diferentes, lo que resulta en razones de muestreo extremas (muy grandes o muy pequeñas). La alta varianza puede ralentizar la convergencia y hacer las estimaciones menos estables.

   - **Convergencia Lenta:** La alta varianza debida a la razón de muestreo de importancia puede requerir más muestras para lograr una estimación precisa de la política objetivo, lo que ralentiza la convergencia.

   - **Regularización y Técnicas de Reducción de Varianza**: Se pueden emplear técnicas de regularización, como la truncación de las razones de muestreo de importancia (capping), para mitigar el impacto negativo en la estabilidad y mejorar la convergencia.

3. **¿Cómo puede el uso de una soft policy influir en la eficacia del aprendizaje de políticas óptimas en comparación con las políticas deterministas en los métodos de Monte Carlo? Compare el desempeño y los resultados de aprendizaje de las políticas derivadas de estrategias épsilon-greedy con las derivadas de políticas deterministas.**

Política Épsilon-Greedy

La política épsilon-greedy es una política soft que, con probabilidad $1-\epsilon$ selecciona la acción con el mayor valor esperado (exploitation), y con probabilidad ϵ\epsilon, selecciona una acción al azar (exploration). Esto ayuda a equilibrar la exploración del espacio de acción y la explotación de las mejores acciones conocidas.

Política Determinista

Una política determinista selecciona siempre la acción con el mayor valor esperado, sin incorporar exploración directa. Esto puede llevar a una exploración insuficiente y a quedarse atrapado en políticas subóptimas.

Comparación de Desempeño

1. **Exploración vs. Explotación**:
   - La política épsilon-greedy generalmente conduce a una mejor exploración del espacio de acción, lo que ayuda a evitar quedarse atrapado en óptimos locales y facilita el descubrimiento de la política óptima.
   - Las políticas deterministas pueden ser rápidas en explotar las mejores acciones conocidas, pero corren el riesgo de converger a una política subóptima debido a la falta de exploración.
2. **Velocidad de Convergencia**:
   - Las políticas deterministas pueden converger más rápido en entornos simples donde la exploración no es tan crítica.
   - Las políticas épsilon-greedy, aunque pueden ser más lentas en converger debido a la exploración, suelen proporcionar mejores resultados en problemas complejos donde la estructura de recompensas es más difícil de descubrir sin exploración.
3. **Estabilidad del Aprendizaje**:
   - Las políticas épsilon-greedy tienden a proporcionar estimaciones más estables y robustas al ruido en los datos debido a su naturaleza exploratoria.
   - Las políticas deterministas pueden ser más sensibles a errores y a variaciones en el entorno debido a su falta de exploración

4. **¿Cuáles son los posibles beneficios y desventajas de utilizar métodos de Monte Carlo off-policy en comparación con los on-policy en términos de eficiencia de la muestra, costo computacional? y velocidad de aprendizaje?**

	Ventajas	Desventajas
On-policy	● La velocidad que maneja cuando las políticas son usadas y diseñadas correctamente ya	● Puede que no se exploren todos los estados y las acciones ya que van a de la mano de una solo

	que se evalúan ellas mismas • El agente puede llegar a ser más directo ya que por medio de las políticas el agente va aprendiendo	política • Se pueden llegar a ejecutar muchas veces ya que se necesitan políticas específicas en muchas ocasiones lo que hace que es costo y el tiempo computacional sea mucho más alto
Off-policy	• El agente puede llegar a ser más óptimo ya que suele no utilizar la misma política sino que recurre a más haciendo así que se exploren más los estados y acciones	• Al usar diferentes políticas para ajustar sus estimaciones, el aprendizaje puede llegar a ser un poco más lento que utilizar una on-policy

Referencias

GeeksforGeeks. (2024, 14 de enero). *Monte Carlo Policy Evaluation*. GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/monte-carlo-policy-evaluation/

Mohan, S. (2022, 21 de diciembre). *Monte Carlo Methods for Reinforcement Learning - Nerd For Tech - Medium*. Medium; Nerd For Tech. https://medium.com/nerd-for-tech/monte-carlo-methods-for-reinforcement-learning-d30d874dd817

Sutton, A. Barto, G. (1998). Monte Carlos Methods. https://www.tu-chemnitz.de/informatik/KI/scripts/ws0910/ml09_5.pdf.

Thomas, P., Theocharous, G., & Ghavamzadeh, M. (2015). High-Confidence Off-Policy Evaluation. *AAAI Conference on Artificial Intelligence*. Recuperado de https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10016/10002

Precup, D., Sutton, R. S., & Singh, S. (2000). Eligibility Traces for Off-Policy Policy Evaluation. *Proceedings of the Seventeenth International Conference on Machine Learning*. Recuperado de https://www.cs.mcgill.ca/~dprecup/publications/ai1999.pdf

Reinforcement Learning
- Laboratorio 4 -

Reinforcement Learning
- Laboratorio 4 -