Inteligencia Artificial - 1er Semestre 2014

udp FACULTAD DE INGENIERÍA

Examen (Pauta)

11 de Julio de 2014

Profesor: Alejandro Figueroa

- Está prohibido el uso de teléfonos celulares durante el desarrollo de la prueba.
- La prueba debe responderse con un lápiz de tinta indeleble, de lo contrario no hay opción a correcciones.
- Cualquier alumno que sea sorprendido intentando copiar será calificado con una nota 1.
- Está prohibido conversar durante la prueba. Recuerde que su compañero puede estar concentrado y el ruido puede perturbarlo en el desarrollo de su prueba.
- Utilice sólo las hojas entregadas para escribir sus respuestas.

1. Búsqueda en Espacios Combinatorios (35 puntos).

a. Compare 2-opt con 3-opt. Tenga presente que la pauta contempla tres diferencias y cuatro similitudes (15 puntos).

2-opt	3-opt
Corta los arcos e intercambia los destinos	Corta los arcos e intercambia los destinos
Corta 2 arcos	Corta 3 arcos
1 forma de reconectarlos	2 formas de reconectarlos
Sólo cambia la solución si mejora	Sólo cambia la solución si mejora
Cuando no hay más mejora termina	Cuando no hay más mejora termina
App. 5% sobre la cota inferior de Held-Karp.	App. 3% sobre la cota inferior de Held-Karp.
Técnica de mejoramiento de soluciones	Técnica de mejoramiento de soluciones

b. ¿Cuáles son las dos diferencias principales entre las metaheurísticas y las técnicas completas? Ejemplifique ambos grupos (5 puntos).

Metaheurísticas	Técnicas Completas
No garatizan encontrar el óptimo	Garatizan que encuentran el óptimo
Utiliza el azar	No utiliza el azar
ACS, BCO, PSO, Hill-Climbing	Fuerza Bruta, Branch and Bound, Backtracking

c. Durante el curso, en especial la primera parte, estudiamos el problema del vendedor viajero (TSP). Diseñe las componentes de un algoritmo genético que busque el camino más corto: Inicialización, función objetivo, cromosoma, mutación, cruzamiento, condición de término, y el mecanismo de selección (15 puntos).

Componente	Implementación
Función Objetivo	$C(P) = \sum_{i=1}^{n-1} d_{P(i),P(i+1)} + d_{P(n),P(1)}$ P una solución factible, $d_{p(i),P(i+1)}$ es la distancia de una ciudad a la otra. Además, n es el número de ciudades. Hay que minimizar
Cromosoma	Tiene "n" celdas, que representan el orden de visita de las "n" ciudades. Cada una de las "n" ciudades debe estar en alguna casilla, ergo ninguna se puede repetir o estar ausente.
Mutación	Swap aleatorio. 2-opt/3-opt también podrían usarse acá.
Cruzamiento	*Enhanced Edge Recombination*
Condición de Término	Un número pre-determinado de iteraciones.
Mecanismo de Selección	Ruleta, es decir inversamente proporcional a la C(P) obtenida por la solución. También por torneos.
Inicialización	Vecino más cercano, random sin repetición.

2. Recuperación de Información (20 puntos).

a. Determine el tamaño del vocabulario de una colección de 100 millones de tokens, utilice k=100 y b=0.5 (10 puntos)

Respuesta: Utilizando la Ley de Heap, se tiene $M = k*T^b = 10^2*(10^8)^{0.5} = 10^2*10^4 = 10^6$

b. Si la frecuencia de la palabra "the" es 100, y además, es la palabra que más aparece en la colección, cuál sería la estimación de la frecuencia de la segunda, tercera y cuarta palabras más frecuentes. Suponga que estas palabras son "a", "of" y "in", respectivamente. (10 puntos).

Respuesta: Utilizando la Ley de Zipf, la frecuencia de "a" sería 100/2=50, la de "of" sería 100/3=33, y la de "in" sería 25.

3. Aprendizaje No-Supervisado (25 puntos)

a. Calcule el Purity de los siguientes clústeres (12 puntos).

Cluster	Ciencia	Deportes	Política	Purity
1	250	20	10	0.893
2	20	180	80	0.643
3	30	100	210	0.617
TOTAL	300	300	300	0.711

b. Compare Single-Link con Complete-Link Ejemplifique ambos grupos (13 puntos). La pauta contempla cuatro diferencias y/o similitudes.

Single-Link	Complete -Link
La distancia entre dos clústeres es la distancia de los dos puntos más cercanos.	Une dos clústers tal que su distancia máxima es la mínima entre todos los clústers.
Sensible a datos ruidosos	No es sensible a datos ruidosos
Complejidad $O(n^2)$	Complejidad $O(n^2 \log n)$
No tiene problemas con outliers	Problemas con outliers

4. Aprendizaje Supervisado (20 puntos)

a. ¿Qué es overfitting? ¿Cuáles son sus potenciales causas? (10 puntos)
   i. Es cuando los modelos no generalizan los datos de buena forma, ajustándose a errores de anotación o en la obtención de los datos.
   ii. Es cuando existe un clasificador c_1 que obtiene una mejor performance que otro c_2 en los datos de entrenamiento, pero en los datos de prueba pasa lo contrario.
   iii. Es producto, muchas veces, de ruido en los datos: a) mal etiquetamiento de los datos; b) valores equivocados en los atributos; y 3) el dominio es complejo y tiene un alto grado de "aleatoriedad".

b. ¿Cuál es la diferencia entre held-out y k-fold cross-validation? (10 puntos) Ejemplifique.

Held-out evaluation divide el conjunto de datos en dos partes: entrenamiento y prueba. Por ejemplo, si hay 10,000 ejemplos, toma un porcentaje, e.g., 10%, 20% para probar, y el resto para entrenar. Normalmente la fracción para probar es mucho menor que la porción para entrenar.

Al contrario, k-fold cross-validation divide el conjunto de entrenamiento en "k" grupos disjuntos de datos de igual tamaño. Después, se hacen k experimentos dejando uno de los k grupos para probar, y usa los k-1 grupos restantes para generar el modelo. K-fold cross-validation hace k experimentos dejando un fold diferente cada vez para probar (el k-ésimo). Held-out no prueba sobre todos los datos, en cambio cross-validation si. Por ende, el primero tiene a usarse cuando el conjunto de datos es grande, y el otro cuando los conjuntos son más reducidos. Por ejemplo, 10,000 datos y k=10 genera 10 grupos de 1,000 datos, ergo 10 experimentos.