Solemne 2 (Pauta)

26 de Junio 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 Tabú Search y Particle Swarm Optimization. Tenga presente que la pauta contempla seis diferencias y dos similitudes (15 puntos).

Tabú Search	Particle Swarm Optimization
Tiene memoria representada en una lista Tabú.	Cada partícula recuerda la mejor solución que ha estado.
Se mueve a un punto cercano en la vecindad de acuerdo a una función N(x).	El N(x) depende no sólo de la vecindad del punto, sino que también de la historia de la "sociedad" de partículas, y la historia de la propia partícula.
Maneja un punto de búsqueda en cada iteración	Maneja un conjunto de puntos(partículas) que forman una "sociedad" (búsqueda paralela)
Parte con una solución al azar	Parte con una solución al azar
N(x) sólo depende de la solución actual, es decir de la geografía definida por el movimiento.	Partículas comparten información social/geográfica de sus soluciones para mejorar la elección del próximo punto.
Termina después de un número determinado de iteraciones o cuando se cumple alguna condición	Termina después de un número determinado de iteraciones o cuando se cumple alguna condición
Estrategias diferentes para manejar la memoria: corto/largo plazo, soluciones completas o patrones, malas y buenas, etc.	La memoria es simple: cada partícula guarda su mejor solución, de las cuales se puede inferir la mejor global.
No es una técnica evolutiva, básicamente porque maneja un solo individuo.	Es una técnica evolutiva, ya que la próxima generación va a contener la amalgama de patrones en las soluciones de la generación actual.

b. ¿Cuál es la diferencia principal entre técnicas constructivas y de mejoramiento de soluciones? Ejemplifique ambos grupos (5 puntos).

Constructivas	Mejoramiento
Construyen una solución factible mientras recorren el espacio de búsqueda. Muchas veces privilegiando instanciaciones que puedan conducir a la mejor solución.	Necesitan una solución ya construida, normalmente factible. Esta solución es mejorada mediante cambios en su estructura.
Backtracking, Branch and Bound, ACS, BCO, etc.	Algoritmos Genéticos, Hill Climbing, Tabú Search, etc.

c. En la última parte del curso, en especial en la penúltima tarea, aprendimos que el desempeño de un clasificador depende de lo features escogidos. Más precisamente, discutimos que la elección de las palabras utilizadas para generar los modelos era importante para lograr una mayor efectividad en los pronósticos. Sin embargo, las bolsas de palabras eran normalmente grande. De ahí que escoger el subconjunto que logre el mejor modelo se transforma en un problema NP-completo. Diseñe las componentes de un algoritmo genético que busque este subconjunto óptimo de palabras: Inicialización, función objetivo, cromosoma, mutación, cruzamiento, condición de término, y el mecanismo de selección (15 puntos).

Componente	Implementación
Función Objetivo	El resultado del accuracy después de desarrollar cross-validation con las palabras que son incluidas por el cromosoma.
Cromosoma	Binario: 0 si la palabra no es incluida, 1 si lo es. El número de genes es el tamaño del diccionario.
Mutación	Swap aleatorio.
Cruzamiento	De 1 ó 2 puntos.
Condición de Término	Accuracy=100% o un número pre-determinado de iteraciones.
Mecanismo de Selección	Ruleta, es decir proporcional al accuracy obtenida. También por torneos.
Inicialización	Asignar unos/ceros aleatoriamente.

2. Recuperación de Información (20 puntos).

a. Nombre cuatro diferencias entre un lematizador y un stemmer. Ejemplifique (10 puntos)

Lematizador	Stemmer
Mapea cada palabra a su raíz, estudiando su forma.	Principalmente, corta cada palabra para obtener un stem. En su gran mayoría es heurístico, y se basa en un conjunto de reglas.
Comete pocos errores.	Comete muchos errores.
Detecta estructuras internas y patrones lingüísticos, e.g., mice → mouse.	Principalmente ataca sufijos y prefijos, e.g., palabras terminadas con “ed”, “ly”, etc.
Es dependiente de la categoría sintáctica de la palabra.	No toma en cuenta la categoría sintáctica de la palabra.
goes, going, went → go	taller, tallest → tall

a. ¿Qué diferencias hay entre las consultas que se hacen a un motor de recuperación de información (e.g., Google) y los documentos que deben recuperarse? Tenga presente que la pauta contempla seis diferencias y dos similitudes (10 puntos).

Consultas	Colección de Documentos
No tienen metadata.	Tienen metadata.
Son cortas.	Varian en largo, pero mucho más largos que las consultas.
Los sustantivos y nombres propios gobiernan las consultas.	Los verbos son la categoría más prominente. Hay uno por oración principalmente.
Las mayúsculas no son confiables.	El uso de las mayúsculas es altamente confiable.
En su gran mayoría, sólo 2-3 palabras.	Estructuras complejas: tablas, imágenes, etc.
Alta ambigüedad.	Baja ambigüedad.

3. Aprendizaje No-Supervisado (25 puntos)

a. Compare K-Means y Fuzzy C-Means. Tenga presente que la pauta contempla dos diferencias y seis similitudes (20 puntos).

K-Means	Fuzzy C-Means
Particional	Particional
K clusters (parámetro)	K clusters (parámetro)
Iterativo hasta: Número de iteraciones o un criterio de error	Iterativo hasta: Número de iteraciones o un criterio de error
Calcula centroides en cada iteración	Calcula centroides en cada iteración
Utiliza una métrica de distancia o similitud.	Utiliza una métrica de distancia o similitud.
Cada dato va a pertenecer a un clúster.	Cada dato va a tener un grado de membrecía a cada uno de los K clústeres.
No tiene otro parámetro a ajustar.	Coeficiente de difusión a ajustar.
Puede caer en un óptimo local	Puede caer en un óptimo local

b. ¿Cuál es la diferencia principal entre las métricas de distancia y similitud? Ejemplifique ambos grupos (5 puntos).

Distancia	Similitud
De 0 a infinito	-1 a 1
Entre más cercano a 0, más parecido son los vectores.	Entre más cercano a 1, más parecido son los vectores.
Euclidiana	Coseno

4. Aprendizaje Supervisado (20 puntos)

Dada los siguientes vectores etiquetados, calcule a) la entropía del conjunto de datos (5 puntos); y b) determine y fundamente si la palabra "cancer" o "battery" es mejor para comenzar un clasificador basado en árboles de decisión (15 puntos).

	a	cancer	tech	recovery	energy	battery	CLASE
1	1	0	1	1	1	0	Tech
2	1	1	0	1	0	0	Health
3	1	0	0	0	1	0	Tech
4	1	0	1	0	0	1	Tech
5	1	0	0	0	1	0	Health
6	1	0	0	0	0	1	Tech
7	1	1	0	0	1	0	Health
8	1	0	0	1	0	1	Health
9	1	0	0	1	1	0	Tech
10	1	0	0	1	0	1	Health
11	1	0	1	1	1	1	Tech
12	1	0	0	1	0	1	Health
13	1	0	0	0	1	0	Tech
14	1	0	1	0	0	0	Tech
15	1	0	1	0	1	0	Tech
16	1	0	1	0	0	1	Tech
17	1	1	0	0	1	1	Health
18	1	0	1	1	0	1	Health
19	1	1	0	1	1	0	Tech
20	1	0	1	0	0	0	Tech
Total	**20**	**4**	**8**	**9**	**10**	**9**	

Respuesta a): Hay doce ejemplos de la clase "Tech" y ocho de la clase "Health".

$$P(C_{tech}) = \frac{12}{20} = \frac{3}{5}$$

$$P(C_{health}) = \frac{8}{20} = \frac{2}{5}$$

$$Entropy(D) = -\left[\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right] = 0{,}970950594$$

Respuesta b): El feature “cancer” aparece en cuatro instancias y es binario, por ende su entropía está dada por:

$$Entropy_{cancer}(D) = -\left[\frac{4}{20}Entropy_{cancer}(D_4) + \frac{16}{20}Entropy_{cancer}(D_{16})\right] = 0{,}879086211$$

$$Entropy_{cancer}(D_1) = -\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right] = 0{,}811278124 \quad (3\,\text{Health y}\,1\,\text{Tech})$$

$$Entropy_{cancer}(D_0) = -\left[\frac{11}{16}\log_2\frac{11}{16} + \frac{5}{16}\log_2\frac{5}{16}\right] = 0{,}896038233 \quad (5\,\text{Health y}\,11\,\text{Tech})$$

$$Entropy_{battery}(D) = -\left[\frac{9}{20}Entropy_{battery}(D_9) + \frac{11}{20}Entropy_{battery}(D_{11})\right] = 0{,}91092724$$

$$Entropy_{battery}(D_1) = -\left[\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right] = 0{,}99107606 \quad (5\,\text{Health y}\,4\,\text{Tech})$$

$$Entropy_{battery}(D_0) = -\left[\frac{3}{11}\log_2\frac{3}{11} + \frac{8}{11}\log_2\frac{8}{1|}\right] = 0{,}84535094 \quad (3\,\text{Health y}\,8\,\text{Tech})$$

$$gain(cancer) = 0{,}970950594 - 0{,}879086211 = 0{,}09186438$$

$$gain(battery) = 0{,}970950594 - 0{,}91092724 = 0{,}06002335$$

El mayor gain es para “cancer”, por ende conviene partir con esta palabra el árbol de decisión.