udp Escuela de Informática y Telecomunicaciones

Examen: Inteligencia Artificial

martes 2 de julio de 2013

Profesor: Alejandro Figueroa

Ayudante: Nicolás Olivares

Tome nota de las siguiente consideraciones antes de comenzar:

- Puede utilizar una calculadora para desarrollar los ejercicios.
- El uso de un aparato electrónico móvil está prohibido.
- Si es sorprendido copiando, se le asignará la nota 1.
- Se prohibe el uso de cualquier tipo de apunte durante el examen.
- Está prohibido ir al baño con algún aparato electrónico (e.g., móvil y laptop)

Recuperación de Información (16 puntos)

1. **¿Cómo se clasifican las intenciones que tienen los usuarios al enviar sus consultas a una máquina de recuperación de información? Discuta las categorías principales (de primer nivel o más abstractas). Ejemplifique cada categoría (4 puntos).**
2. **¿Qué diferencias hay entre las consultas que se envían a un buscador y los documentos convencionales en términos de categorías sintácticas (POS)? (8 puntos).**
3. **Explique el algoritmo BSBI (4 puntos).**

Aprendizaje No-Supervisado (28 puntos)

1. **Compare clustering jerárquico y particional. (4 puntos)**
2. **¿Qué estrategias existen para inicializar los clústeres en K-means? (4 puntos)**
3. **Enuncie la distancia de Minkowski y Canberra (8 puntos).**
4. **Explique el problema de "Estandarización de Datos". Ejemplifique (12 puntos).**

Semántica Distribucional (36 puntos)

1. **¿Cuál es la idea principal de la semántica distribucional? (3 puntos)**
2. **¿Por qué se explica que los hiperónimos aparezcan juntos a sus hipónimos en un texto? Ejemplifique (5 puntos).**
3. **Calcule PMI (Pointwise Mutual Information) y NPMI para las siguientes palabras (15 puntos).**

Palabra	Frecuencia Global

Apple	71
Orange	31
Pineapple	45
Pear	54
Watermelon	25
Melon	35
Kiwi	23
Fruit	250
Color	289
Computer	412
Chocolate	321

Palabra 1	**Palabra 2**	**Frecuencia del Par**
Apple	Fruit	24
Orange	Color	12
Pineapple	Computer	3
Pear	Chocolate	12
Watermelon	Fruit	18
Kiwi	Color	3
Apple	Computer	31
Orange	Chocolate	14
Pineapple	Fruit	35
Pear	Color	32
Watermelon	Computer	5
Kiwi	Chocolate	10

Use N_{ij} = 100000 y N_i =10000. Use logaritmo en base 10.

4. **Explique la diferencia entre palabras que contrastan y opuestas. ¿En qué difieren dos palabras opuestas? (4 puntos)**
5. **Explique "Explicit Semantic Analysis" (9 puntos).**

Aprendizaje Supervisado (20 puntos)

1. **¿Qué es Overfitting?. Expliquelo en el contexto de la tarea 3d (6 puntos).**
2. **Dada el siguiente conjunto de datos (14 puntos):**

Atributo 1	**Atributo 2**	**Clase**
Adulto	Bien	T
Joven	Regular	F
Anciano	Mal	T
Adulto	Mal	F
Joven	Bien	T

Anciano	Regular	F
Adulto	Bien	T
Joven	Mal	F
Anciano	Bien	T
Joven	Mal	F

a) **Calcule la entropía de D (4/14 puntos).**
b) **Calcule el Information Gain para cada atributo de un árbol de decisión. Considere sólo el nodo inicial. ¿Cuál sería el nodo raíz? (10/14 puntos).**

udp Escuela de Informática y Telecomunicaciones

Examen: Inteligencia Artificial (Pauta)

martes 2 de julio de 2013

Profesor: Alejandro Figueroa

Ayudante: Nicolás Olivares

Tome nota de las siguiente consideraciones antes de comenzar:

- Puede utilizar una calculadora para desarrollar los ejercicios.
- El uso de un aparato electrónico móvil está prohibido.
- Si es sorprendido copiando, se le asignará la nota 1.
- Se prohibe el uso de cualquier tipo de apunte durante el examen.
- Está prohibido ir al baño con algún aparato electrónico (e.g., móvil y laptop)

Recuperación de Información (16 puntos)

1. **¿Cómo se clasifican las intenciones que tienen los usuarios al enviar sus consultas a una máquina de recuperación de información? Discuta las categorías principales (de primer nivel o más abstractas). Ejemplifique cada categoría (4 puntos).**

Navegación	El objetivo es ir a un sitio conocido específico que tengo en mente. La razón por la cual utilizo el motor de búsqueda es porque es más conveniente que tipiar la URL, o quizás no la sé.	aloha airlines duke university hospital kelly blue book
Información	Mi intención es aprender algo mediante la lectura o viendo alguna página web.	what is a supercharger baseball death and injury why are metals shiny color blindness help quitting smoking walking with weights pella windows phone card amsterdam universities
Recurso	La idea es obtener algún recurso disponible en páginas web.	kazaa lite measure converter ellis island lesson plans

2. **¿Qué diferencias hay entre las consultas que se envían a un buscador y los documentos convencionales en términos de categorías sintácticas (POS)? (8 puntos).**

Consultas Web	Documentos
Cortas (2.8 palabras promedio)	Largos .. párrafos.
La mayoría de las etiquetas que se utilizan para documentos son muy raras en las consultas. Por ejemplo, posesivos, adverbio, pronombres, verbos (2.35%).	Los documentos tienen oraciones completas por ende la gran mayoría tiene verbos, no así las consultas que están formadas mayoritariamente por frases sustantivas.
Los sustantivos propios abarcan cerca del 40.2%.	Los sustantivos propios abarcan el 13%.
Vemos principalmente un tipo de verbo.	35 tipos de verbos.
La más común son los sustantivos propios	La más común son los sustantivos.
El uso de las mayúsculas es inconsistente.	Las mayúsculas son útiles para indicar los sustantivos propios.
La categoría URI.	Normalmente no existe esa categoria.
Las consultas son formas resumidas de peticiones complejas.	No se observa este fenómeno.

3. **Explique el algoritmo BSBI (4 puntos).**
   Este algoritmos consta de los siguientes pasos:
   I. Segmenta la colección en partes iguales.
   II. Recolecta de cada documento pares <ID término, ID documento> y los acumula en memoria hasta que se llena un bloque de un tamaño pre-determinado.
   III. El bloque es invertido y escrito en el disco: ordenar los pares y genera los posting lists.
   IV. Almacena los resultados intermedios en el disco.
   V. Une los resultados intermedios en un índice final.

Aprendizaje No-Supervisado (28 puntos)

1. **Compare clustering jerárquico y particional. (4 puntos)**

Jerárquico	Particional
Genera un árbol donde la unión de los hijos forma los datos que tienen el padre.	Forma particiones disjuntas de un mismo nivel.
2 métricas de distancia: entre elementos y entre clústers.	1 métrica de distancia

Trabaja con clústers, solo el nivel de las hojas con datos.	Asignas datos a clústers.

2. **¿Qué estrategias existen para inicializar los clústeres en K-means? (4 puntos)**

   Varias formas: 1) aleatoria; 2) escoger K datos de manera aleatoria; y 3) Calcular el centroides de todos los datos, y se escoge el punto más lejando al centroide, y después al más lejano al primer punto, así sucesivamente.

3. **Enuncie la distancia de Minkowski y Canberra (8 puntos).**

$$dist(x_i, x_j) = \sqrt[h]{\left|x_{i1} - x_{j1}\right|^h + \ldots + \left|x_{ir} - x_{jr}\right|^h}$$

$$dist(x_i, x_j) = \frac{\left|x_{i1} - x_{j1}\right|}{x_{i1} + x_{j1}} + \ldots + \frac{\left|x_{ir} - x_{jr}\right|}{x_{ir} + x_{jr}}$$

   X_i y X_j son vectores de features, *r* es el número de atributos, y *h* es el parámetro de la distancia de Minkowski.

4. **Explique el problema de "Estandarización de Datos". Ejemplifique (12 puntos).**

   La idea de estandarización es que todos los atributos puedan tener igual impacto en el cómputo de la distancia. El objetivo es prevenir que hayan clústeres dominados por atributos con una alta variación.

   Suponga dos vectores <0.1, 20> y <0.9, 720>, la distancia euclidiana: 700.000457. Esta distancia está dominada casi completamente por 720 - 20 = 700.

   Solución: Transformar 20 y 720 a 0.02 y 0.72, lo que arroja la distancia 1.063.

Semántica Distribucional (36 puntos)

1. **¿Cuál es la idea principal de la semántica distribucional? (3 puntos)**

   La idea es deducir relaciones de palabras semánticas de una colección de documentos: "*palabras que aparecen frecuentemente en el mismo contexto*".

2. ¿Por qué se explica que los hiperónimos aparezcan juntos a sus hipónimos en un texto? Ejemplifique (5 puntos).

Al redactar un texto conviene utilizar hiperónimos para evitar la repetición de palabras ya empleadas anteriormente, como se hace en el siguiente ejemplo:
"*De repente, un descapotable rojo paró frente al banco. Del automóvil salieron dos individuos encapuchados, mientras otro esperaba en el vehículo*."

3. Calcule PMI (Pointwise Mutual Information) y NPMI para las siguientes palabras (15 puntos).

Palabra	Frecuencia Global
Apple	71
Orange	31
Pineapple	45
Pear	54
Watermelon	25
Melon	35
Kiwi	23
Fruit	250
Color	289
Computer	412
Chocolate	321

Palabra 1	Palabra 2	Frecuencia del Par	$p_{ij}/(p_i*p_j)$	pmi_{ij}	$log(p_{ij})$	$npmi_{ij}$
Apple	Fruit	24	1,35211268	0,13101288	-3,61978	0,03619352
Orange	Color	12	1,3394352	0,12692171	-3,92081	0,03237123
Pineapple	Computer	3	0,1618123	-0,7909884	-4,52287	-0,1748860
Pear	Chocolate	12	0,69228107	-0,1597175	-3,92081	-0,0407357
Watermelon	Fruit	18	2,88	0,45939249	-3,74472	0,12267715
Kiwi	Color	3	0,4513314 3	-0,3455044	-4,52287	-0,0763903
Apple	Comput	31	1,059756	0,025206	-	0,007184

	er		6	13	3,50863	02
Orange	Chocolate	14	1,40689378	0,14826131	-3,85387	0,03847074
Pineapple	Fruit	35	3,11111111	0,49291552	-3,45593	0,14262883
Pear	Color	32	2,0504934	0,31185838	-3,49485	0,08923369
Watermelon	Computer	5	0,48543689	-0,3138672	-4,30103	-0,0729749
Kiwi	Chocolate	10	1,35446296	0,13176713	-4	0,03294178

Use N_{ij} = 100000 y N_i =10000. Use logaritmo en base 10.

4. Explique la diferencia entre palabras que contrastan y opuestas. ¿En qué difieren dos palabras opuestas? (4 puntos)

Dos palabras opuestas tienen muchas propiedades en común pero difieren grandemente en una dimensión de significado (dimensión de oposición). Por ejemplo, las palabras gigante y enano son cosas vivientes, comen, caminan, piensan, pero son muy diferentes en la dimensión de altura.

También se pueden ver como palabras que tienen una relación binaria incompatible. Por ejemplo, "día" y "noche", día es "no noche".

5. **Explique "Explicit Semantic Analysis" (9 puntos).**

   La idea es representar un documento como un conjunto de artículos Wikipedia relacionados. Ese grupo de documentos forman un vector. Para ésto, se construye un índice invertido sobre los artículos de Wikipedia filtrados para poder encontrar los artículos que contienen una palabra. El filtrado consiste en eliminar "stubs", por ejemplo.

   Se usa un interpretador semántico que recorre las palabras de un texto y buscar los artículos relacionados mediante el índice. De esos artículos se recuperan sólo los k (e.g., 10) mejores. Ese ranking está dado por tf-idf. Como para un texto se recuperan varios artículos, estos se unen en un vector de conceptos, donde refleja la importancia del cada concepto encontrado en Wikipedia para el texto en cuestión. Este procedimiento permite forman vectores de pesos por fragmentos de textos con los cuales se puede calcular la relación semántica entre dos pares de textos.

Aprendizaje Supervisado (20 puntos)

3. **¿Qué es Overfitting?. Expliquelo en el contexto de la tarea 3d (6 puntos).**

   Es cuando se tienen muy buenos resultados con los datos de entrenamiento, pero muy malos con los datos de prueba. Esto indica que los datos no fueron generalizados de buena forma y se ajustaron a las particularidades, por ejemplo, los errores de anotación.

   En general, se dice que hay overfitting cuando existe un clasificador A que obtiene una performance que otro B en los datos de prueba, pero en los datos de prueba pasa lo contrario.

   En los árboles de decisión se ven representado por hojas muy profundas que cubren pocos datos.

   En el caso de las tablas, existe la posibilidad de que hayan sido anotadas descuidadamente, por ende, el modelo de aprendizaje no haya aprendido patrones de "descuido" y por ende, al etiquetar datos de prueba, por ejemplo lo anotados por otra persona, su performance sea muy mala. En ese caso, las generalizaciones hechas sobre los datos de entrenamiento son malas, producto de la mala etiquetación. Es decir, se hizo overfitting sobre el conjunto de datos.

   Otras razones de overfitting son: valores equivocados en los atributos, el dominio es complejo y tiene un gran grado de aleatoriedad.

4. Dada el siguiente conjunto de datos (14 puntos):

Atributo 1	Atributo 2	Clase
Adulto	Bien	T
Joven	Regular	F
Anciano	Mal	T
Adulto	Mal	F
Joven	Bien	T
Anciano	Regular	F
Adulto	Bien	T
Joven	Mal	F
Anciano	Bien	T
Joven	Mal	F

c) Calcule la entropía de D (4/14 puntos).

$$Entropia(D) = -\frac{5}{10} * \log_2 \frac{5}{10} - \frac{5}{10} * \log_2 \frac{5}{10} = 1\ bit$$

d) Calcule el Information Gain para cada atributo de un árbol de decisión. Considere sólo el nodo inicial. ¿Cuál sería el nodo raíz? (10/14 puntos).

$$Entropia_{Atributo1}(D_{Adulto}) = -\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3} = 0.918$$

$$Entropia_{Atributo1}(D_{Joven}) = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4} = 0.811$$

$$Entropia_{Atributo1}(D_{Anciano}) = -\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3} = 0.918$$

$$Entropia_{Atributo1}(D) = \frac{3}{10} * 0.918 + \frac{4}{10} * 0.811 + \frac{3}{10} * 0.91 = 0.87524\ bits$$

$$Entropia_{Atributo2}(D_{Bien}) = -\frac{4}{4}\log_2 \frac{4}{4} - \frac{0}{4}\log_2 \frac{0}{4} = 0$$

$$Entropia_{Atributo2}(D_{Mal}) = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4} = 0.811$$

$$Entropia_{Atributo2}(D_{Regular}) = -\frac{2}{2}\log_2 \frac{2}{2} - \frac{0}{2}\log_2 \frac{0}{2} = 0$$

$$Entropia_{Atributo2}(D) = \frac{4}{10} * 0 + \frac{4}{10} * 0.811 + \frac{2}{10} * 0 = 0.3244\ bits$$

$$gain(D, Atributo1) = 1 - 0.87524 = 0.1248$$

$$gain(D, Atributo2) = 1 - 0.3244 = 0.6756$$