INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia

Control 1

El siguiente control está pensado para ser respondido en un plazo de 2.5 horas (2 horas y 30 minutos). Recuerde poner su nombre en todas las hojas. Si lo desea puede utilizar las hojas de enunciado para responder.

Pregunta 1

Responda las siguientes preguntas de manera clara y precisa (1 punto cada una).

a) Explique cómo opera el mecanismo de entrenamiento de una red neuronal multi-capas que utiliza el algoritmo *Back-Propagation* con regla delta de aprendizaje. Puede apoyarse en ecuaciones, dibujos, etc. para aclarar su explicación.
b) Explique cómo opera el mecanismo de entrenamiento de una red neuronal de Kohonen. Puede apoyarse en ecuaciones, dibujos, etc. para aclarar su explicación.
c) Explique el mecanismo de aprendizaje de una red Bayesiana y de qué forma puede ser utilizada para la detección del molesto spam que llega cada cierto tiempo en los emails.
d) ¿De qué forma se puede disminuir el over fitting en la fase de entrenamiento de un árbol de decisión? Muestre con ejemplos gráficos el problema y la solución propuesta.
e) Explique el mecanismo de entrenamiento del algoritmo K-means utilizando Medois en vez de centroides. ¿Cuál es la principal diferencia en la creación de clusters?
f) Explique los conceptos de Aprendizaje supervisado, NO-supervisado y por Refuerzo.

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 1

Nombre:

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 1

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia

Pregunta 2

Dados los siguientes documentos:

- D1 = *Se produjo un bombazo en la estación escuela militar. La bomba ha dejado 14 heridos.*
- D2 = *Medios internacionales publican sobre el bombazo producido en Chile.*
- D3 = *Nadie se ha atribuido la explosión causada por una bomba detonada en escuela militar.*
- D4 = *En la tarde del lunes una bomba explotó en un basurero causando heridos.*

Debe determinar qué documento es más similar D1. Para esto es necesario lo siguiente:

I. Emplee los criterios de preprocesamiento de acuerdo a lo abordado en clases. Debe **justificar cada decisión** y ser consecuente con los criterios que dijo utilizaría.
II. Utilice como *features* unigrams de palabras y calcule la matriz TF-IDF.
III. Compare mediante la medida de similitud del coseno cual es el documento más parecido a D1.

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \times |d_2|} = \frac{\sum_{i=1}^{|V|}(w(t_i, d_1) \times w(t_i, d_1))}{\sqrt{\sum_{i=1}^{|V|} w(t_i, d_1)^2} \times \sqrt{\sum_{i=1}^{|V|} w(t_i, d_2)^2}}$$

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 2

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 2

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia

Pregunta 3

Responda las siguientes preguntas en forma clara y ordenada.

a) (1 pto) En base a la siguiente matriz de confusión calcule: *accuracy*, *precision* y *recall*.

	Valor real: Sí	Valor real: No
Predicción: Sí	410	150
Predicción: No	165	390

b) (1 pto) Un método para mejorar el valor de una función, sea maximizando o minimizando, es el método del gradiente. ¿Se trata de un método de optimización o de una heurística? Explique, además indique una ventaja y una desventaja de este método.

c) (1 pto) Explique el funcionamiento del modelo *k-nearest neighbour* (KNN). Indique, además, una modificación para que el análisis considere toda la data en vez de una muestra.

d) (1 pto) Considere el modelo de clusterización *k-means*. Dado que por construcción no entregará una solución óptima, explique un método que permita mejorar los resultados.

e) (2 ptos) Considere el modelo vectorial con una matriz término-documento *M* para sitios web. Explique cuál es el proceso por el cual se llega a poblar la matriz (desde el sitio web hasta el cálculo de frecuencias). Indique qué método puede usar un indexador web para entregar el sitio más acorde a una búsqueda.

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 3

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia

Hoja Respuesta Pregunta 3

Nombre: