APRENDIZAJE AUTOMÁTICO
TEMA 1

Nombre: José Arcos Aneas
D.N.I: 74740565-H

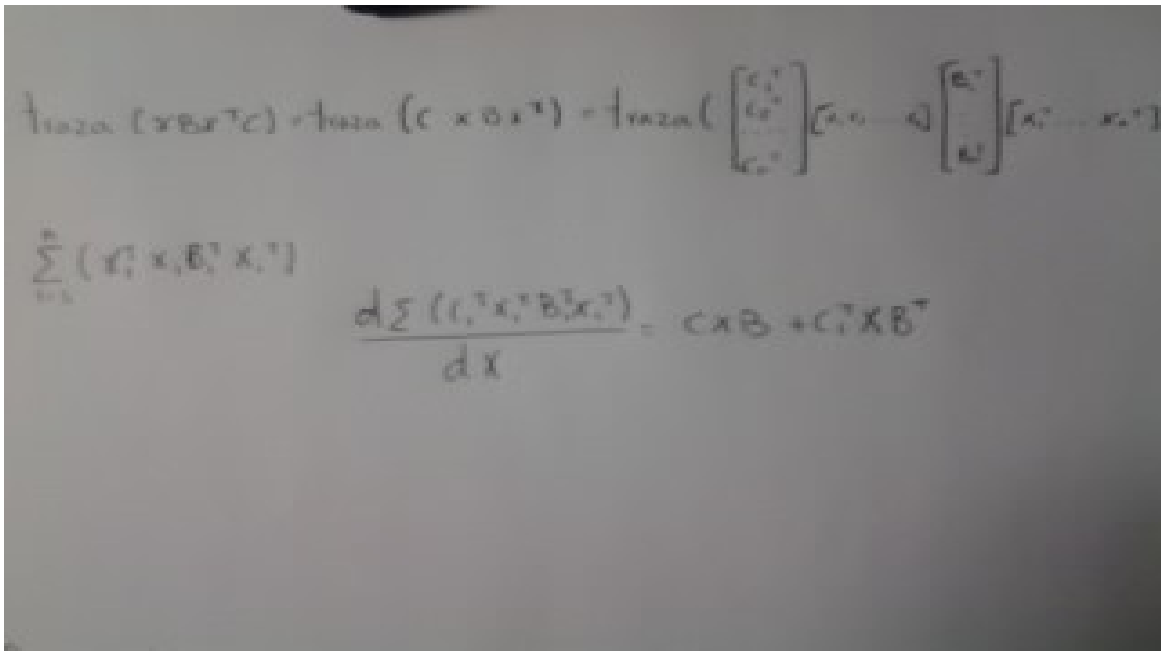
Bibliografía y documentación utilizadas:

- Wikipedia.
- Documentos de diferentes universidades (Valencia, Madrid, Granada)
- Ejecicios 5 y 6 de EJERCICIOS Alberto Quesada Aranda

Cuestionario de Teoría (12 puntos)

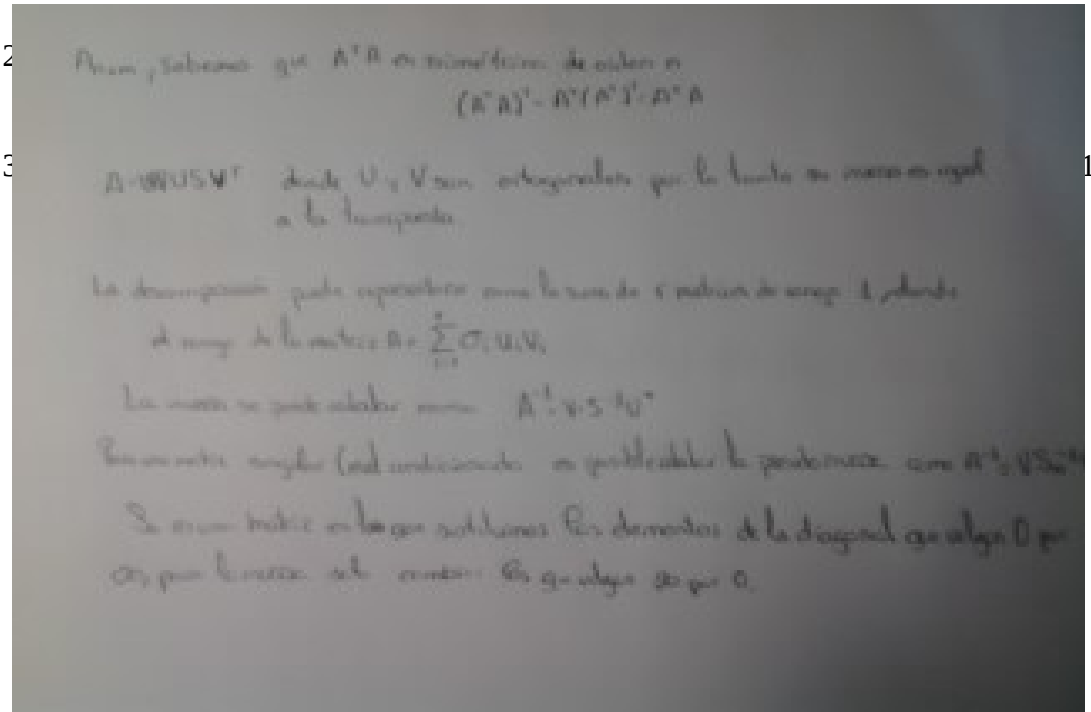
EJERCICIOS: (6 puntos)

1. Verificar que ∂/∂X traza(XBX T C)== CXB + CtXBt

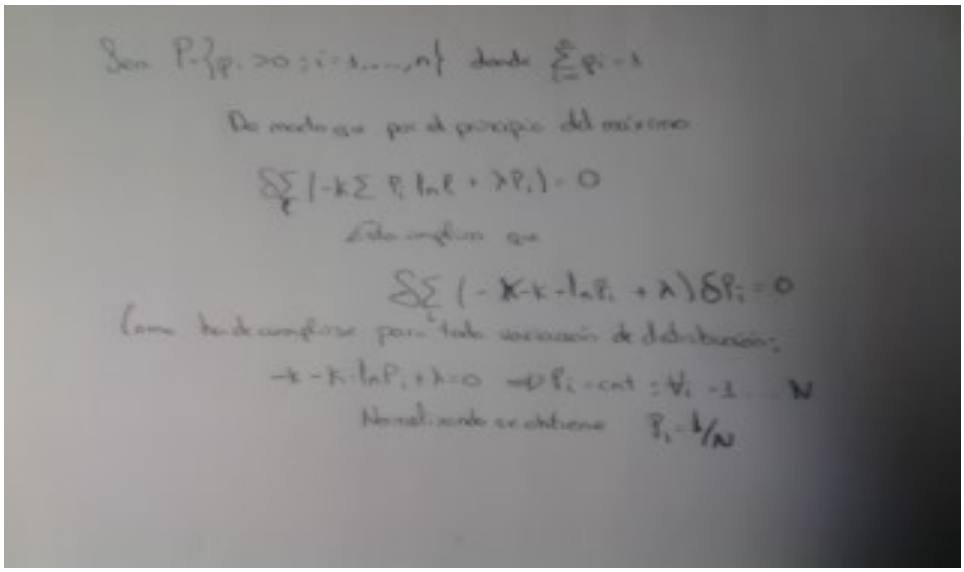


2 x

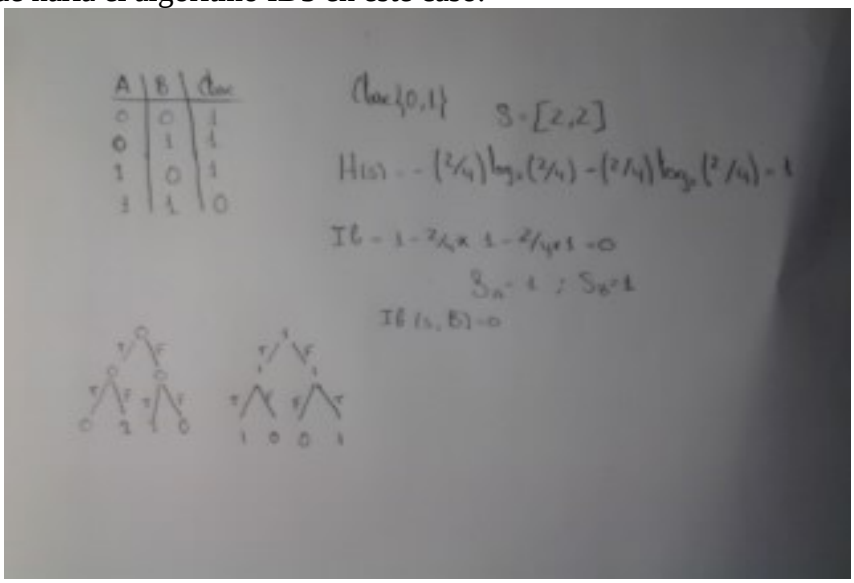
3 1)



4. Sean P = {pi > 0, i = 1, · · · , n} una distribución de probabilidad o discreta, es decir sumatorio(pi) = 1. Probar usando la técnica de los multiplicadores de Lagrange que la distribución de máxima entropía H(P) se obtiene cuando pi =1/n.



5. Suponga que le dan los siguientes 4 vectores de datos para entrenar un árbol de decisión. Calcule la ganancia de información de las dos características y valore lo que haría el algoritmo ID3 en este caso.



6. Sea H(label) = 0,95 la entropía de las etiquetas de un nodo N con 8 muestras. Supongamos tres posibles situaciones de ramificación dadas por las variables x1 , x2 , x3 con valores (T,F) y visualizadas en la figura. Analizar los tres supuestos y decir cual de los tres sería preferido por el algoritmo ID3 en términos de la ganancia de información. Extraer alguna conclusión sobre las situaciones que prefiere el algoritmo ID3 a la hora de ramificar.

H(label) = 0,95
Tenemos 8 muestras

H(x1 -> true) = -4/6 * log_base_dos (4/6) - 2/6 * log_base_dos(2/6) = 0,9182
H(x1 -> false) = 1
IG(label, x1) = 0,95 - 6/8 * 0,9182 - 2/8 * 1 = 0,01135

H(x2 -> true) = 1
H(x2 -> false) = -2/2 * log_base_dos(2/2) - 0/2 * log_base_dos(0/2) = 0
IG(label, x2) = 0,95 - 6/8 * 1 - 2/8 * 0 = 0,2

H(x3-> true) = 1
H(x3-> false)= -4/4 * log_base_dos(4/4) - 0/4 * log_base_dos(0/4) = 0
IG(label, x3) = 0,95 - 4/8 * 1 - 4/8 * 0 = 0,45

La ganancia de información es x3 y sera el primero que se tomara para hacer crecer al árbol.
El elegido es el que tiene un nodo hoja directamente en su ramificación y además su otra rama tiene menor diferencia con los valores mas pequeños.

CUESTIONES: (6 puntos)

1. **Identifique las diferencias esenciales entre las técnicas de aprendizaje supervisado las técnicas de aprendizaje no-supervisado. Justificar la respuesta.**

Respuesta:

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.

2. **¿Que diferencias esenciales hay entre un problema de regresión y uno de clasificación? Justificar la respuesta.**

Respuesta:

Con la regresión se trata de encontrar una función que permita aproximar los datos de una variables independiente en función de datos que dependientes. Los valores que devuelven son numéricos.

Con los clasificadores lo que hacemos en clasificar los datos en función de unas variables dependientes. con los clasificadores se obtiene un clasificador a partir de unos datos de entrenamiento. una ve construido el clasificador se debe medir su precisión el clasificador suele devolver etiquetar o algún parámetro que identifique los elementos de una clase.

3. **¿Cual es el objetivo más importante en el entrenamiento de cualquier técnica supervisada? Justificar la respuesta.**

Respuesta:

Construir un construir un conjunto (o grupo) para que los objetos que se han clasificado con ciertos, sean identificados. Los parámetros que se describen a los objetos deben ser discriminatorios para la clasificación.

4. **¿Que considera que es mejor, que un modelo una vez entrenado ajuste perfectamente los datos de entrenamiento o que no lo haga? Justificar la respuesta.**

Respuesta:

Es preferible un tamaño pequeño para entrenamiento entre un 10% y un 30% del total de muestra y el resto para test. Si optamos por un numero alto es posible que nos pasemos y nos ciñamos a los datos que hemos tenido para entrenar y con un nuevo valor no sepamos clasificarlo.

5. **¿Que es mejor usar un K grande o uno pequeño en la regla k-NN? Identificar los pros y contras de cada caso. Justificar la respuesta.**

Respuesta:

Un K muy grande nos llevaría a sobreajuste y un K muy pequeño no representaría las características de nuestra población.
La mejor elección de k depende fundamentalmente de los datos: los valores grandes de k reducen el efecto de ruido en la clasificación. Pero crean limites entre clases parecidas. Un buen k puede ser seleccionado por una optimización de uso.

6. **¿Es posible tener sobreajuste con un clasificador k-NN? En caso afirmativo ¿como se puede corregir? Justificar la respuesta**

Respuesta:

Si considerando un conjunto de k muy grande que conlleva a sobrevalorar la capacidad predictiva de los modelos obtenidos.
Lo arreglaríamos cogiendo un conjunto mas pequeño de k en el que aparezcan bien representadas las características de cada clase.

7. **Demostrar que en el plano el conjunto de puntos que están a igual distancia de dos prototipos dados p0 y p1 es una recta del plano definida por dichos prototipos.**

Respuesta:

8. **¿De que manera influye la dimensión del vector muestral en el resultado de la regla k-NN? Justificar la respuesta.**

Respuesta:

Entendiendo como dimensión como el número mínimo de coordenadas necesarias para especificar cualquier punto de un espacio. El tamaño influye en el numero de características que se han de tener en cuenta a la hora de crear el clasificador. Por lo que el calculo de la distancia euclídea sera algo mas complicado. Si con dimensión nos referimos al tamaño del vector de muestras tendremos mas posibilidades de crear un clasificador generalice mejor los datos de test.

9. **Identifique los cuatro aspectos que en su opinión son los más importantes a la hora de construir un árbol de decisión ID3. Justificar la respuesta.**

Respuesta:

Los aspectos mas importantes a la hora de construir un árbol ID3 son las siguientes:
1.Describir los objetos en términos de parejas, el atributo y su valor
2.Calcular la entropía para todas las clases.
3.Seleccionar el mejor atributo basado en la reducción de la entropía.
4.Iterar hasta que todos los objetos sean clasificados.

10. **¿Es posible tener sobreajuste con un modelo de árbol de clasificación? En caso afirmativo ¿como se puede corregir?, en caso negativo decir por que no. Justificar la respuesta.**

Respuesta:

Si, de hecho si hacemos crecer mucho un árbol de decisión en la etapa de entrenamiento, podemos aumentar el error por varianza. Esto es a lo que se denomina sobreajuste.
Por otro lado sino dejamos que crezca lo suficiente, podemos aumentar el error por sesgo.

Los problemas con la varianza son originados porque los datos de entrenamiento no son representativos de datos objetivos.
Si una hoja del árbol tiene pocos datos la regla asociado tendrá baja significancia estadística.

Para mejorar estos problemas de precisión se utilizan un mecanismo llamado poda. Tiene varios tipos:

Pre-poda: Parar la construcción del árbol en algunas ramas en tiempo de construcción
Post-poda: Construir un árbol posiblemente sobreajustado y podarlo después.

11. **Suponga que sobre un mismo conjunto de datos de entrenamiento es posible aplicar un clasificador k-NN o un árbol de decisión. Identifique los criterios que considere son los relevantes para decidir el uso de una u otra técnica. Haga las hipótesis que considere oportunas sobre el conjunto de datos.**

Respuesta:

Un clasificador con un numero medio de valores para test y con valores no muy dispersos en el que los grupos puedan estar bien formados sin muchos valores en las fronteras de las regiones que identifican cada clase puede ser una opción muy recomendable.
Los arboles de decisiones resumen los ejemplos de partida, permitiendo la clasificación de nuevos casos siempre y cuando no existan modificadores sustanciales en las condiciones bajo las cuales se

generaron los ejemplos que sirvieron para su construcción. Facilitan la interpretación de la decisión adoptada. Explica el comportamiento respecto a una determinada tarea de decisión. Y algo muy importante es que reduce el número de variables independientes. Por lo que seria bueno en un vector con muchas características.

12. **Suponga que tenemos una variable aleatoria X que puede tomar dos valores con probabilidad p y 1 − p. Calcular la expresión de la entropía H(X) como función de p y analizar que rango de valores de p serán los más informativos sobre la variable X.**

Respuesta:

La entropía de esta variable, aplicando la fórmula de la definición es:

$H[X]= -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p)$

En el caso general, tenemos infinitas distribuciones diferentes con este esquema dependiendo del valor de p, que recorre los reales en el intervalo [0,1].
El valor máximo de la entropía es para p=0,5

La máxima entropía sera $H[X]= \log_2(n)$