Artificial Intelligence
HW 4: Machine learning

Published: 2021/05/10
Due date: -
In class solutions: -

---

1. Knn. Una empresa de telecomunicaciones desea hacer su marketing telefónico de forma más eficiente debido a que en las campañas de nuevos productos, los tele-operadores tienen que llamar varias veces a los clientes y en la mayoría de los casos esos seguimientos no se traducen en nuevas contrataciones.

Se nos plantea la idea de construir un modelo que pueda predecir si un cliente nos va a contratar un nuevo producto a través de la campaña de marketing telefónico
en función de los datos del cliente, del estado actual de su seguimiento y de cuál fue el resultado con ese cliente en campañas anteriores.

Si desarrollamos un modelo k-NN con k=3. Cuál será la respuesta del modelo para un nuevo cliente con género="V", tiene ya un producto contratado= "NO" y producto a comprar = "telefonía fija".

Adjunto se muestra el historial de comportamiento de clientes previos:

Id cliente	**Género**	**Producto ofertado en la campaña**	**Tiene ya un producto contratado**	**Resultado de la campaña. Contrata producto (S/N)**
1	V	Móvil	NO	NO
2	M	Móvil	SI	NO
3	V	Internet	NO	SI
4	V	Internet	SI	NO
5	M	Internet	NO	SI
6	V	TV	SI	SI
7	M	Telefonía fija	NO	NO

SOLUCIÓN:

Los 3 más parecidos son id 7 (similitud 2/3). Id 1 (similitud 2/3). Id 3 similitud (2/3). Los demás tienen < similitud (=1/3) asique no hay empates. Si usamos distancia de hamming. Si usamos euclidea depende de la codificación que elijamos. Lo normal es elegir one hot encoding.

¿Porque? Pues porque no nos especifican que un producto de campaña valga o signifique más que otro. De este modo si codificamos móvil, internet, tv y fijo como 0,1,2,3, implicaría que es más parecido fijo que tele que fijo que móvil por ejemplo. En general debemos tomar que las variables cualitativas nominales (las no ordinales) se transforman en una variable booleana por cada valor distinto del atributo. Luego la codificación más apropiada sería. Genero : V= 0, M= 1. Movil (si/no) = 0/1. Internet (si/no) = 0/1. Tv (si/no) = 0/1. Fijo (si/no) = 0/1. Tiene producto ya: no= 0, si =1. La mayoría de lo que dicen los 3 knn es (NO-NO-SI). Luego la respuesta que da el modelo es NO.

2. Intro ML. Supongamos que disponemos del siguiente conjunto de datos destinados a ser utilizados para entrenar un predictor de *k* vecinos próximos.

N	x_{n1}	x_{n2}	x_{n3}	y_n
1	2.34	Sí	Rojo	Bien
2	3.21	Sí	Verde	Regular
3	2.01	No	Rojo	Mal
4	2.77	Sí	Verde	Bien
5	3.08	Sí	Azul	Bien
6	2.83	No	Rojo	Regular

1. ¿De qué tipo es el problema de predicción: clasificación, clasificación ordinal o regresión?
2. ¿Qué tipo de preprocesamiento realizarías para los datos? En concreto
2.1. ¿Modificarías de algún modo los valores numéricos de x_{n1}?
2.2. ¿Cómo codificarías los valores de x_{n2}?
2.3. ¿Cómo codificarías los valores de x_{n3}?
2.4. ¿Cómo codificarías los valores de y?
3. ¿Qué función de distancia utilizarías?
4. ¿Cuáles son los parámetros y los hiperparámetros de este modelo predictivo (k vecinos próximos?
5. Escribe el pseudocódigo de un algoritmo para determinar el número de vecinos próximos (*k*).

SOLUCIÓN:

1. .¿De qué tipo es el problema de predicción: clasificación, clasificación ordinal o regresión?
Clasificación ordinal: Mal < Regular < Bien
2. ¿Qué tipo de preprocesamiento realizarías para los datos? En concreto
2.1. ¿Modificarías de algún modo los valores numéricos de x_{n1}?
Centrar: Sustrayendo media, mediana, o punto medio del intervalo
Escalar: Dividiendo los valores del atributo una vez centrado por un factor de escala (desviación estándar, espaciado intercuartílico, radio del intervalo).
2.2. ¿Cómo codificarías los valores de x_{n2}?
Codificación binaria: No = 0; Sí = 1.
2.3. ¿Cómo codificarías los valores de x_{n3}?
Codificación 1 de 3: Rojo = 100; Verde = 010; Azul = 001.
2.4. ¿Cómo codificarías los valores de y?
Codificación de distancia: Mal = 0; Regular = 1; Bien = 2;
3. ¿Qué función de distancia utilizarías?
Se puede utilizar una distancia L_p con p > 0. Por ejemplo;
Distancia de Manhattan (L_1): d({\mathbf x}_n ,{ \mathbf x}_m) = \sqrt{\left| x_{n1} − x_{m1} \right|^2}

Distancia euclídea (L_2):
4. ¿Cuáles son los parámetros y los hiperparámetros de este modelo predictivo (k vecinos próximos)?
Parámetros: Los propios datos de entrenamiento
Hiperparámetros:
función de distancia (hiperparámetro funcional)
número de vecinos (k, hiperparámetro entero).
5. Escribe el pseudocódigo de un algoritmo para determinar el número de vecinos próximos (*k*).
Se puede utilizar cualquier método de validación (simple, K-fold, leave-one-out). Por ejemplo, mediante ***leave-one-out cross-validation***
ENTRADA:
Valores de k considerados: k Î {k_1, k_2, …, k_M}
Datos de entrenamiento:
SALIDA: Valor de k* determinado por leave-one-out cross-validation
For m = 1:M # valor de k considerado):
error(k_m) := 0

For n = 1:N # ejemplo del conjunto de entrenamiento)
Predice la clase del ejemplo $\mathbf{x}_n$ utilizando k_m vecinos próximos
Si la clase predicha no coincide con y_n
error(k_m) += 1
k^∗ = \arg \min_k \left[\left\{ error(k_m)\right\}_ {m=1}^M\right]

3. Árboles de decisión. Considera la siguiente base de datos:

ID ejemplo	Edad	X2	X3	Clase
1	20	0	0	Sí
2	30	0	0	No
3	50	0	0	Sí
4	20	0	1	No
5	40	1	1	Sí
6	60	1	1	Sí
7	30	1	2	No
8	20	1	2	No
9	40	1	2	No

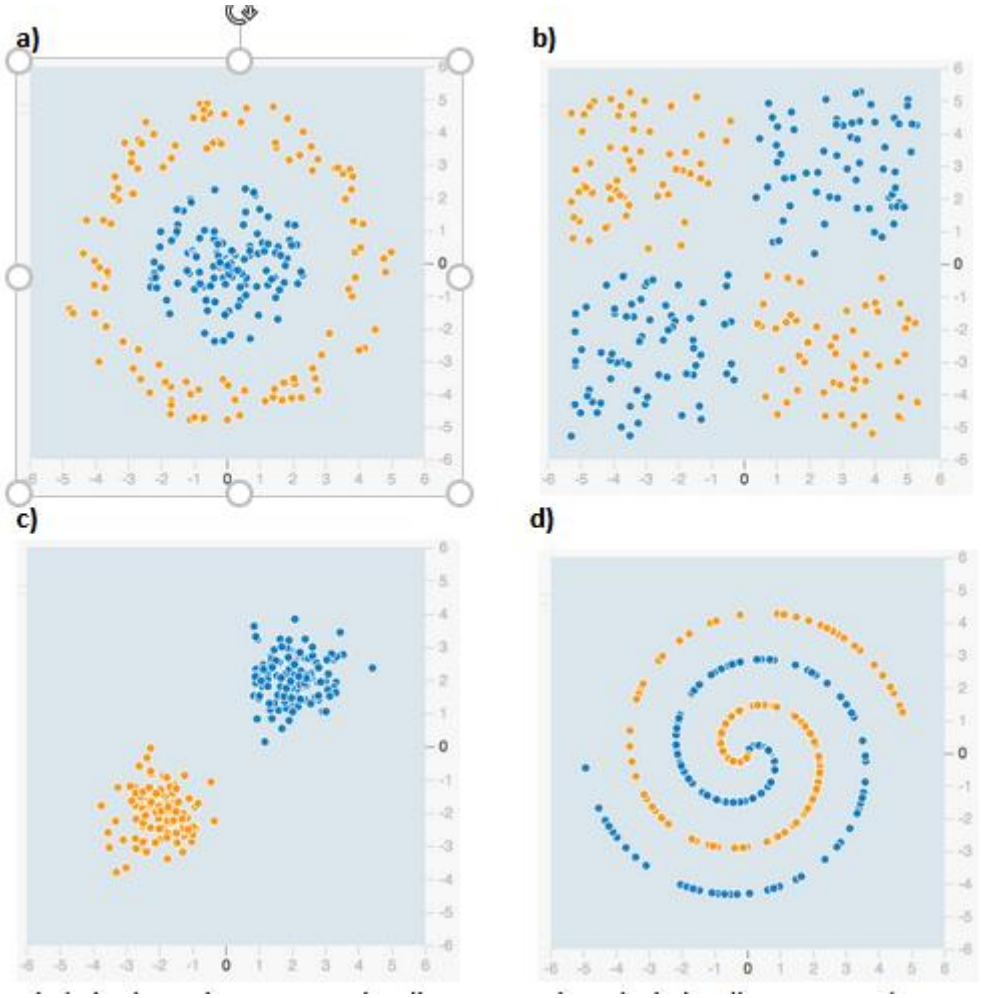
Responde razonadamente e incluye todos los cálculos intermedios que has realizado. No se darán por válidas respuestas sin los cálculos intermedios:

1. ¿Cuál es la entropía inicial de la clase?
2. ¿Cuáles son las preguntas candidatas a nodo raíz que va a evaluar C4.5?
3. ¿Cuál es la ganancia de información de cada una de esas preguntas candidatas?
4. ¿Qué pregunta elige finalmente C4.5 como nodo raíz?

SOLUCIÓN:

soon

4. Intro Deep learning. Supongamos que tenemos 4 datasets diferentes representados por las siguientes cuatro gráficas (tomado de https://playground.tensorflow.org):



El eje horizontal representa el atributo X1 y el vertical el atributo X2. Cada punto es un patrón (con valores particulares en estas dos variables), y el color (naranja / azul) representa la clase.

Responde justificadamente a las siguientes preguntas:

1) ¿Cuál o cuáles de esos datasets son linealmente separables?

2) ¿Qué problema o problemas de esos cuatro podrías resolver con total precisión con una regresión logística?

3) ¿Y con un árbol de decisión C4.5 con profundidad 1? (nodo raíz + 2 nodos hoja)

4) ¿Y si el árbol tiene profundidad 2?

5) ¿Y con una red neuronal de una capa oculta con 100 neuronas?

SOLUCIÓN:

1) c) es linealmente separable ya que las dos clases se pueden separar completamente por una recta

2) Solo c) se puede resolver con total precisión con una regresión logística ya que este método sólo puede resolver completamente problemas que son linealmente separables

3) Un árbol de decisión C4.5 va a realizar preguntas del tipo X1>umbral o X2>umbral, lo que introduce una separación de los puntos a través de una línea vertical u horizontal respectivamente. Si solo hay una pregunta de ese tipo, el único dataset que se puede resolver completamente es de nuevo el c), a través de la pregunta X1>0

4) Si el árbol tiene profundidad 2 el árbol puede realizar 1 (raíz) + 2 (hijos del raíz) separaciones del tipo X1>umbral o X2>umbral. El único dataset que se puede resolver completamente de esta forma es, en caso de que el árbol encuentre los cortes correctos, el b). Los cortes serían por ejemplo X1>0 (raíz) y X2>0 en ambos hijos. Otra posibilidad es X2>0 (raíz) y X1>0 en ambos hijos

5) Una red neuronal de una capa oculta con 100 neuronas va a combinar 100 rectas para construir la frontera de clasificación. Aparentemente esto es suficiente para resolver todos los datasets.

5. Árboles de decisión. Considera los siguientes datos que utilizaremos para entrenar un árbol de decisión con el algoritmo ID3 (la clase es la última columna):

Género	Fumador	Predisposición a enfermedad
V	Sí	Sí
V	Sí	Sí
V	No	No
M	Sí	No
V	No	No
M	No	Sí

a) ¿Qué pregunta pone en la raíz ID3? Detalla los cálculos que has necesitado realizar para contestar a esto.

b) Consideremos ahora la misma tabla pero con más datos:

Género	Fumador	Rango de edad	Predisposición a enfermedad
V	Sí	Anciano	Sí
V	Sí	Anciano	Sí
V	No	Anciano	No
M	Sí	Joven	No
V	No	Anciano	No
M	No	Anciano	Sí

Si desarrollamos un modelo k-NN con k=3, ¿cuál será la respuesta del modelo para un nuevo cliente con género=”V”, Fumador=“Sí” y Rango de edad = “joven”? Codifica las variables de la manera que creas más conveniente y explica dicha

SOLUCIÓN:

a) Priores de las clases

P(Clase = si) = 1/2;

P(Clase = no) = 1/2

H(Clase) = 1 bit

Partición utilizando Sexo

P(Sexo = V) = 4/6 = 2/3;

P(Clase = si | Sexo = V) = 1/2; P(Clase = no| Sexo = V) = 1/2;

H(Clase | Sexo = V) = 1 bit

P(Sexo = M) = 2/6 = 1/3;

P(Clase = si | Sexo = M) = 1/2; P(Clase = no| Sexo = M) = 1/2;

H(Clase | Sexo = M) = 1 bit

H(Clase | Sexo) = H(Clase | Sexo = V) P(Sexo = V)

+ H(Clase | Sexo = M) P(Sexo = M)

= (2/3)×1 + (1/3)×1 = 1 bit

GI(Clase | Sexo) = H(Clase) - H(Clase | Sexo) = 0

Partición utilizando Fumador

P(Fumador = Si) = 3/6 = 1/2;

P(Clase = si | Fumador = si) = 2/3; P(Clase = no| Fumador = si) = 1/3;

H(Clase | Fumador = si) = 0.9183 bits

P(Fumador = no) = 3/6 = 1/2;

P(Clase = si | Fumador = no) = 1/3; P(Clase = no| Fumador = no) = 2/3;

H(Clase | Fumador = no) = 0.9183 bits

H(Clase | Fumador) = H(Clase | Fumador = si) P(Fumador = si)

+ H(Clase | Fumador = no) P(Fumador = no)

= (1/2)× 0.9183 + (1/2)× 0.9183 = 0.9183 bits

GI(Clase | Fumador) = H(Clase) - H(Clase | Fumador) = 1 - 0.9183 = 0.0817 bits

Se elige “Fumador” en la raíz al tener la máxima ganancia de información (mínima entropía de la clase dada la pregunta)

b) Los 3 más parecidos son el 1,2, 4. Con distancia=2/3. Los demás tiene distancia 1/3. Si usamos distancia de Hamming. Si usamos Euclidea depende de la codificación que elijamos.

Fórmula:

Lo normal es elegir one hot encoding. En general debemos tomar que las variables cualitativas nominales (las no ordinales) se transforman en una variable booleana por cada valor distinto del atributo. En este caso cada atributo tiene sólo dos valores luego es muy fácil. Luego la codificación más apropiada sería. Genero : V= 0, M= 1. Fumador: no= 0, si =1. Rango edad: Joven = 0. Anciano =1. Codificado buscamos los tres vectores más parecidos a (0,1,0). La respuestas son sí, si no. Luego el modelo responde lo que la mayoría que es sí.

id	Género	Fumador	Rango de edad	Predisposición a enfermedad	distancia
1	0	1	1	Sí	1
2	0	1	1	Sí	1
3	0	0	1	No	2
4	1	1	0	No	1
5	0	0	1	No	2
6	1	0	1	Sí	2

6. **Intro ML.** Una empresa de telecomunicaciones desea hacer su marketing telefónico de forma más eficiente debido a que en las campañas de nuevos productos los teleoperadores tienen que llamar a los clientes y en la mayoría de los casos esas llamadas no se traducen en nuevas contrataciones.

Se nos plantea la idea de construir un modelo que pueda predecir si un cliente nos va a contratar un nuevo producto a través de la campaña de marketing telefónico en función de los datos del cliente y de cuál fue el resultado con ese cliente en campañas pasadas.

1. La información general que tenemos de los clientes es:

1. Número de cliente
2. Documento de identidad
3. Nombre y apellidos
4. Fecha de nacimiento
5. Educación: básica, secundaria, universitaria
6. Fecha inicio como cliente
7. Productos ya contratados: listado que puede incluir ninguno, uno o varios de los siguientes elementos: Telefonía fija, Internet, Móvil, TV.

2. La información disponible de las campañas anteriores es:

1. Número de cliente
2. Número de campaña
3. Producto ofrecido en la campaña. Es uno de los siguientes: Telefonía fija, Internet, Móvil, TV, otros servicios
4. Resultado de la campaña: Producto contratado / Producto no contratado

Planteamos como una tarea de clasificación la predicción de si un cliente nos contratará o no el nuevo producto. Para ello se deben tomar las siguientes decisiones (responde razonadamente a cada una de ellas):

1. ¿Cuál es la variable a predecir?
2. ¿Qué información de la disponible es relevante para la construcción del modelo? Es decir, por cada variable enumerada especifica: a) ¿La incluirías en el modelo o no? b) ¿Por qué?
3. Describe cómo obtendrías el conjunto de entrenamiento a partir de los datos generales de los clientes y la información de campañas anteriores. ¿Qué representa cada una de las filas del conjunto de entrenamiento? ¿cuáles serían los atributos?
4. ¿Qué transformaciones podrían ser útiles dentro de cada atributo seleccionado? Pon solo un par de ejemplos y justifícalos.
5. ¿Cuántas ramas tendría el nodo raíz de un árbol de decisión si el atributo más relevante para predecir fuera el atributo “Producto de la campaña"?
6. Si usamos k-nn como modelo predictivo necesitamos que todos los atributos sean numéricos. ¿Cómo codificarías numéricamente los atributos “Educación” y “Producto de la campaña”? Justifícalo.

Supongamos que después de preparar el dataset construimos un árbol de decisión y al evaluarlo en el conjunto de test obtenemos que de los 20 ejemplos de clientes que contrataron un nuevo producto el clasificador predijo 15 de ellos correctamente, y que de los 80 ejemplos de clientes que NO contrataron el nuevo producto el clasificador predijo 70 de ellos correctamente.

Se pide:

7. Calcular la matriz de confusión para esta evaluación
8. Calcular la precisión global (accuracy) del clasificador

SOLUCIÓN:

1. Predecir Resultado de la campaña
2. útiles: años que tiene, educación, meses o días como cliente, productos en posesión, resultados de las anteriores compras. Datos no útiles nombre y apellidos, puesto esto no es algo que se pueda generalizar ni entrar un modelo es algo concreto de cada caso.
3. hacer una tabla donde las x tienen un identificador por id cliente + id producto, el resto todos los atributos, y las ys a predecir es Resultado de la adquisición. Imaginamos los datos como 3 tablas excel y las juntamos las 3 en una sola con un atributo de clave de tabla.
4. Fecha por días desde la última entrevista, variable exógena: número de entrevistadores? Transformar la edad con la fecha a numérico? Realizar una normalización de la edad…, etc.
5. 5: Telefonía fija, Internet, Móvil, TV, otros servicios
6. Educacion es ordinal la paso a 0,1,2. Producto campaña no es ordinal por eso hago una variable dummie/booleana por cada de los posibles valores de producto a comprar pongo una nueva columna en la matriz y un 0 o 1 en caso de que exista o no.
7. TN= 70, FN=5, Fp= 10, TP= 15

	predicted	
real	15	5
	10	70

8. accuracy = 70+15/100=0.85

7. Árboles de decisión y k-NN. Considera los siguientes datos que utilizaremos para entrenar un árbol de decisión con el algoritmo C4.5:

X1	X2	X3	Clase

0	0	1	Sí
1	1	1	No
2	0	1	No
2	1	2	Sí
0	0	2	No
1	1	2	Sí

a) ¿Qué pregunta pone en la raíz C4.5? Detalla los cálculos que has necesitado realizar para contestar a esto.

b) Si desarrollamos con estos datos un modelo k-NN con k=3, ¿cuál será la respuesta del modelo para el caso X1=1, X2=1, X3=0?

SOLUCIÓN:

soon

8. Regresión logística. Tenemos el siguiente dataset:

X1	X2	Clase
0	0	1
1	2	0
1	0	1

Queremos construir una regresión logística usando este dataset como datos de entrenamiento.

Todos los pesos iniciales son 0 y la constante de aprendizaje es 0.5

¿Cuál sería el resultado tras la primera época de realizar un entrenamiento **batch** del modelo?

Escribe en tu solución todos los pasos necesarios para realizar estos cálculos.

SOLUCIÓN:

soon