Cuestiones teóricas de Introducción a la Minería de Datos

Jaime Lorenzo Sánchez

5 de febrero de 2023

Índice

1. DATOS

1.1. Muchas ciencias se basan en la observación en lugar de experimentos diseñados. Comparar los problemas de calidad de datos involucrados en la ciencia observacional con los de la ciencia experimental y la minería de datos

Las ciencias observacionales tienen el problema de no poder controlar completamente la calidad de los datos que obtienen, siendo necesario trabajar con los datos disponibles en lugar de datos de un experimento cuidadosamente diseñado. En ese sentido, el análisis de datos para la ciencia observacional se parece a la minería de datos.

1.2. Distinguir entre ruido y valores atípicos

1.2.1. ¿Es el ruido alguna vez interesante o deseable? ¿Existen excepciones?

Por definición, el ruido se refiere a la modificación de los valores originales. Existen excepciones como por ejemplo los valores atípicos (son objetos de datos con características considerablemente diferentes a la mayoría de los otros objetos de datos en el conjunto de datos).

1.2.2. ¿ Pueden los objetos de ruido ser atípicos?

Los objetos de ruido pueden ser atípicos, ya que la distorsión aleatoria de los datos suele ser responsable de los valores atípicos.

1.2.3. ¿Son los objetos de ruido siempre atípicos?

No. La distorsión aleatoria puede resultar en un objeto o valor muy parecido a uno normal.

1.2.4. ¿Los valores atípicos son siempre objetos de ruido?

No. A menudo, los valores atípicos simplemente representan una clase de objetos que son diferentes de los objetos normales.

1.2.5. ¿Puede el ruido convertir un valor típico en uno inusual, o viceversa?

Sí.

1.3. Considere el problema de encontrar los K vecinos más cercanos de un objeto de datos. Un programador diseña el Algoritmo 2.1 para esta tarea.

Algorithm 2.1 Algorithm for finding K nearest neighbors.

1: **for** $i = 1$ to *number of data objects* **do**
2: Find the distances of the i^{th} object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first K distances of the sorted list
5: **end for**

1.3.1. Describa los posibles problemas con este algoritmo si hay objetos duplicados en el conjunto de datos. Suponga que la función de distancia solo devolverá una distancia de 0 para objetos que son iguales.

Primero, el orden de los objetos duplicados en una lista de vecinos más cercanos dependerá de los detalles del algoritmo y del orden de los objetos en el conjunto de datos. En segundo lugar, si hay suficientes duplicados, la lista de vecinos más cercanos puede consistir sólo de duplicados. Tercero, un objeto puede no ser su propio vecino más cercano.

1.3.2. ¿Cómo solucionarías este problema?

Un enfoque es mantener solo un objeto para cada grupo de objetos duplicados. En este caso, cada vecino puede representar un solo objeto o un grupo de objetos duplicados.

1.4. Considere una matriz de término de documento, donde tf_{ij} es la frecuencia de i-ésimo palabra (término) en el j-ésimo documento y m es el número de documentos. Considere la transformación variable que está definida por $tf'_{ij} = tf_{ij} * log\frac{m}{df_i}$, donde df_i es el número de documentos en el que aparece el término i-ésimo y se conoce como la frecuencia del documento del término. Esta transformación se conoce como transformación inversa de frecuencia del documento.

1.4.1. ¿Cuál es el efecto de esta transformación si aparece un término en un documento? ¿En cada documento?

Los términos que aparecen en todo documento tienen peso 0, mientras que los que ocurren en un documento tienen un peso máximo, es decir, log m.

1.4.2. ¿Cuál podría ser el propósito de esta transformación?

Esta normalización refleja la observación de que los términos que ocurren en cada documento no tiene ningún poder para distinguir un documento de otro, mientras que los que son relativamente raros lo hacen.

1.5. Este ejercicio compara y contrasta algunas medidas de similitud y distancia

1.5.1. Para datos binarios, la distancia L1 corresponde a la distancia de Hamming; es decir, el número de bits que son diferentes entre dos vectores binarios. La similitud de Jaccard es una medida de la similitud entre dos vectores binarios. Calcule la distancia de Hamming y la similitud de Jaccard entre los siguientes dos vectores binarios: x = 0101010001 y = 0100011000

Distancia Hamming = 3. Similitud de Jaccard = $\frac{Numerodeunos-1}{numerobits-nmerodeceros-1} = \frac{2}{5} = 0'4$.

1.5.2. ¿Qué enfoque, la distancia de Jaccard o Hamming, es más similar al coeficiente de coincidencia simple y qué enfoque es más similar a la medida del coseno? Explique. (Nota: la medida de Hamming es una distancia, mientras que las otras tres medidas son similitudes, pero no dejes que esto te confunda).

La distancia de Hamming es similar a la SMC. De hecho, $SMC = \frac{distanciadeHamming}{nmerodebits}$. La medida de Jaccard es similar a la medida del coseno porque ambas ignoran los partidos 0-0.

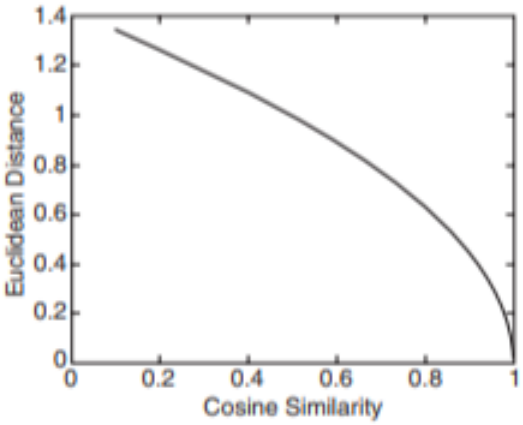
1.5.3. Suponga que está comparando qué tan similares son dos organismos de diferentes especies en términos de la cantidad de genes que comparten. Describe qué medida, Hamming o Jaccard, crees que sería más apropiada para comparar la composición genética de dos organismos. Explique. (Suponga que cada animal se representa como un vector binario, donde cada atributo es 1 si un gen particular está presente en el organismo y 0 en caso contrario).

Jaccard es más apropiado para comparar la composición genética de dos organismos; ya que queremos ver cuántos genes comparten estos dos organismos.

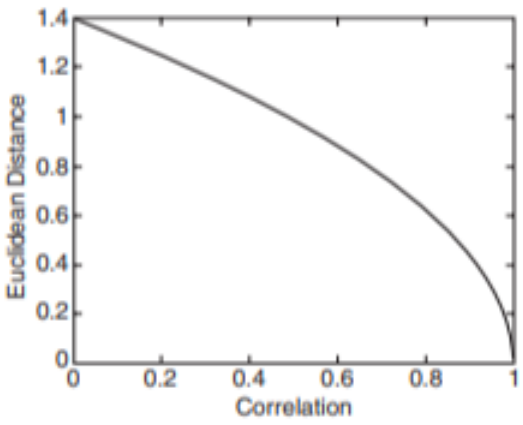
1.5.4. Si quisiera comparar la composición genética de dos organismos de la misma especie, por ejemplo, dos seres humanos, ¿usaría la distancia de Hamming, el coeficiente de Jaccard o una medida diferente de similitud o distancia? Explique. (Tenga en cuenta que dos seres humanos comparten $> 99,9\,\%$ de los mismos genes).

Dos seres humanos comparten más del $99,9\,\%$ de los mismos genes. Si queremos comparar la composición genética de dos seres humanos, debemos centrarnos en sus diferencias. Por lo tanto, la distancia de Hamming es más apropiada en esta situación.

1.6. Muestre que la métrica de diferencia de conjuntos dada por d(A, B) = tamaño(A menos B) + tamaño(B menos A) satisface los axiomas métricos dados en la página 70. A y B son conjuntos y (A menos B) es la diferencia de conjuntos.



(a) Relationship between Euclidean distance and the cosine measure.



(b) Relationship between Euclidean distance and correlation.

1(a). Como el tamaño de un conjunto es mayor o igual a 0, entonces $d(x, y) \geq 0$.

1(b). Si (A = B) $\rightarrow$ (A - B) = (B - A) = θ y por lo tanto d(x,y) = 0.

2. d(A,B) = tamaño(A - B) + tamaño(B - A) = tamaño(B - A) + tamaño(A - B) = d(B,A).

3. Primero, tenga en cuenta que d(A, B) = tamaño(A) + tamaño(B) - 2 * tamaño(A $\bigcap$ B).

- d(A,B) + d(B,C) = tamaño(A) + tamaño(B) - 2 * tamaño(A $\bigcap$ B) + tamaño(B) + tamaño(C) - 2 * tamaño(B $\bigcap$ C).

Dado que tamaño (A ∩ B) ≤ tamaño (B) y tamaño (B ∩ C) ≤ tamaño (B),
d(A, B) + d(B,C) ≥ tamaño(A) + tamaño(C)+2*tamaño(B) − 2*tamaño(B)
= tamaño(A) + tamaño(C) ≥ tamaño(A) + tamaño(C) − 2*tamaño(A ∩ C)
= d(A, C).

- d(A, C) ≤ d(A, B) + d(B,C)

1.7. Dada una medida de similitud con valores en el intervalo [0,1], describa dos formas para transformar este valor de similitud en un valor de disimilitud en el intervalo [0,∞].

d = $\frac{1-s}{s}$ y d = -log s.

1.8. Explicar por qué calcular la proximidad entre dos atributos suele ser más sencillo que calcular la similitud entre dos objetos.

En general, un objeto puede ser un registro cuyos campos (atributos) sean de diferente tipos. Para calcular la similitud general de dos objetos en este caso, necesitamos decidir cómo calcular la similitud de cada atributo y luego combinar estas similitudes. Por el contrario, los valores de un atributo son todos del mismo tipo, y así, si otro atributo es del mismo tipo, entonces el cálculo de la similitud es conceptual y computacionalmente sencilla.

1.9. Discuta las diferencias entre la reducción de dimensionalidad basada en agregación y la reducción de dimensionalidad basada en técnicas como PCA y SVD.

La dimensionalidad de PCA o SVD puede verse como una proyección de los datos en un conjunto reducido de dimensiones. En agregación, los grupos de dimensiones están combinados y, en algunos casos, la agregación puede ser vista como un cambio de escala. Por el contrario, la reducción de la dimensionalidad basada en técnicas como PCA y SVD no tienen tal interpretación.

2. CLASIFICACIÓN

2.1. La figura 4.13 muestra que la entropía y el índice de Gini aumentan monótonamente en el rango [0, 0'5] y ambos disminuyen monótonamente en el rango [0'5, 1]. ¿Es posible que la ganancia de información y la ganancia en el índice de Gini favorezcan atributos diferentes? Explique.

Sí, aunque estas medidas tienen un rango similar y un comportamiento monótono, sus respectivas ganancias, que son diferencias escaladas de las medidas, no necesariamente se comportan de la misma manera.

2.2. Si bien el método de arranque de .632 es útil para obtener una estimación confiable de la precisión del modelo, tiene una limitación conocida. Considere un problema de dos clases, donde hay igual número de ejemplos positivos y negativos en los datos.Suponga que las etiquetas de clase para los ejemplos se generan aleatoriamente. El clasificador utilizado es un árbol de decisiones no podado (es decir, un memorizador perfecto). Determine la precisión del clasificador usando cada uno de los siguientes métodos.

2.2.1. El método de retención, en el que dos tercios de los datos se usan para entrenamiento y el tercio restante se usa para pruebas.

Asumiendo que los ejemplos de entrenamiento y de test son igualmente representativos, el ratio de error de test será cercano al 50 %.

2.2.2. Validación cruzada de diez vecinos.

Asumiendo que los ejemplos de entrenamiento y de test son igualmente representativos, el ratio de error de test será cercano al 50 %.

2.2.3. El método de arranque .632.

El error de entrenamiento para un memorizador perfecto es del 100 %, mientras que la tasa de error para cada muestra de arranque es cercana al 50 %. Sustituyendo esta información en la fórmula para el método de arranque .632, la estimación del error es: $\frac{1}{b} * \sum_{i=1}^{b} [0{,}632 * 0'5 + 0{,}368 * 1] = 0{,}684$.

2.2.4. A partir de los resultados de las partes (a), (b) y (c), ¿qué método proporciona una evaluación más confiable de la precisión del clasificador?

El método de retención y validación cruzada de diez veces proporciona una mejor estimación del error que el método boostrap de .632.

2.3. Dados los conjuntos de datos que se muestran en las figuras 5.6, explique cómo se comportarían en estos conjuntos de datos el árbol de decisión, el bayesiano ingenuo y los clasificadores de vecinos más cercanos.

Analizamos el comportamiento de los datos en el árbol de decisión:

- Funcionará bien pues los atributos distintivos tienen mejor poder de discriminación que los atributos de ruido en términos de ganancia de entropía y probabilidad condicional.
- Tendrá un problema de sobreajuste debido a la cantidad relativamente grande de atributos distintivos.
- Funcionará pero dará como resultado un árbol de decisión bastante grande. Las primeras divisiones serán bastante al azar, porque puede que no encuentre una buena división inicial al principio. Si usa una división oblicua en lugar de sólo divisiones verticales y horizontales, entonces el árbol de decisión resultante será más compacto y altamente preciso.
- Tendrá un árbol grande para capturar los límites de decisión circulares.

Analizamos el comportamiento del NB:

- Funcionará bien pues los atributos distintivos tienen mejor poder de discriminación que los atributos de ruido en términos de ganancia de entropía y probabilidad condicional.
- No funcionará en absoluto pues se observa cierta dependencia de atributos. Los otros esquemas funcionarán mejor.
- Funcionará muy bien pues cada atributo discriminante tiene una mayor probabilidad condicional en una clase sobre otra, y la clasificación general se realiza multiplicando estas probabilidades condicionales individuales.
- No funcionará tan bien debido a la dependencia de atributos.

Analizamos el comportamiento del KNN:

- No funcionará bien pues se observa un número relativamente grande de atributos de ruido.
- Lo hará razonablemente bien debido a la cantidad relativamente grande de atributos distintivos.
- Funciona mejor que los demás modelos.

3. REGLAS DE ASOCIACIÓN

3.1. Para cada una de las siguientes preguntas, proporcione un ejemplo de una regla de asociación del dominio de la cesta de la compra que satisfaga las siguientes condiciones. Además, describa si tales reglas son subjetivamente interesantes.

3.1.1. Una regla que tiene un alto apoyo y una gran confianza.

Leche $\rightarrow$ Pan: Una regla tan obvia tiende a ser poco interesante.

3.1.2. Una regla que tiene un apoyo razonablemente alto pero poca confianza

Leche → Atún: Si bien la venta de leche y atún puede estar por encima del umbral de apoyo, no todas las transacciones que contienen leche también contienen atún. Una regla con poca confianza tiende a ser poco interesante.

3.1.3. Una regla que tiene poco apoyo y poca confianza.

Aceite de cocina → Detergente. Una regla con poca confianza tiende a ser poco interesante.

3.1.4. Una regla que tiene poco apoyo y alta confianza

Vodka → Caviar. Una regla con alta confianza tiende a ser interesante.

3.2. Supongamos que s1 y c1 son los valores de soporte y confianza de una regla de asociación r al tratar cada ID de transacción como una canasta de mercado. Además, sean s2 y c2 los valores de soporte y confianza de r al tratar cada ID de cliente como una canasta de mercado. Discuta si hay alguna relación entre s1 y s2 o c1 y c2.

No existe relación aparente entre s1,s2,c1 y c2.

3.3. Responda las siguientes preguntas utilizando los conjuntos de datos que se muestran en la Figura 6.6. Tenga en cuenta que cada conjunto de datos contiene 1000 elementos y 10 000 transacciones. Las celdas oscuras indican la presencia de elementos y las celdas blancas indican la ausencia de elementos. ¿Aplicaremos el algoritmo Apriori para extraer conjuntos de elementos frecuentes con minsup = 10 % (es decir, los conjuntos de elementos deben estar contenidos en al menos 1000 transacciones)?

3.3.1. ¿Qué conjuntos de datos producirán la mayor cantidad de conjuntos de elementos frecuentes?

El conjunto de datos (e) porque genera el conjunto de elementos más frecuentes junto con sus subconjuntos.

3.3.2. ¿Qué conjuntos de datos producirán la menor cantidad de conjuntos de elementos frecuentes?

Conjunto de datos (d) pues no produce conjuntos de elementos frecuentes en el umbral de soporte del 10 %.

3.3.3. ¿Qué conjuntos de datos producirán el conjunto de elementos más frecuente?

Conjunto de datos (e).

3.3.4. ¿Qué conjuntos de datos producirán conjuntos de elementos frecuentes con el soporte máximo más alto?

Conjunto de datos (b).

3.3.5. ¿Qué conjuntos de datos producirán conjuntos de elementos frecuentes que contengan elementos con niveles de soporte muy variados (es decir, elementos con soporte mixto, que van desde menos del 20 % hasta más del 70 %).

Conjunto de datos (e).

4. AGRUPAMIENTO

4.1. Identifique los conglomerados en la Figura 8.3 utilizando las definiciones basadas en el centro, la contigüidad y la densidad. Indique también el número de conglomerados para cada caso y dé una breve indicación de su razonamiento. Tenga en cuenta que la oscuridad o el número de puntos indica densidad. Si ayuda, suponga que basado en el centro significa K-means, basado en contigüidad significa enlace único y basado en densidad significa DBSCAN.

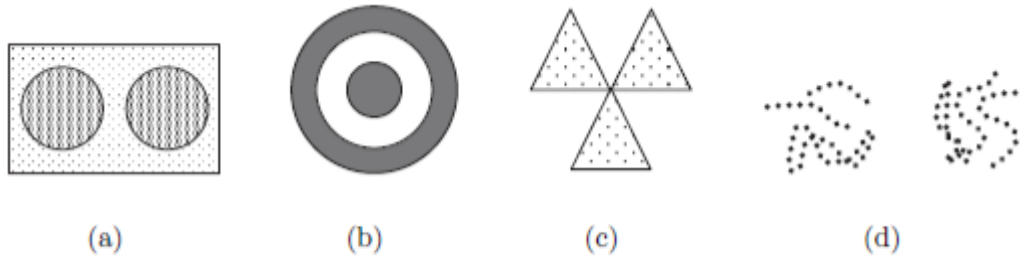


Figure 8.3. Clusters for Exercise 5.

Clúster (a): Tenemos que el rectángulo se divide en 2 por la mitad, luego tenemos un centro basado en 2 clusters (ruido incluido en ambos clústers). Las 2 regiones circulares se unen por el ruido, luego tenemos 1 clúster basado en contigüidad. Tenemos un clúster, para cada círculo, basado en densidad. Por tanto, tenemos 2 clúster basado en densidad, eliminando el ruido.

Clúster (b): Basados en el centro, tenemos que los 2 anillos forman un clúster. Basados

en la contigüidad, ambos anillos se unen por el ruido y forman un solo clúster. Basados en la densidad, tenemos 2 clúster (uno por cada anillo).

Clúster (c): Basados en el centro, tenemos un anillo por cada triángulo (3 clúster). Sin embargo, también es aceptable un clúster. Basados en contigüidad, los 3 triángulos se unirán en un solo clúster debido al ruido. Basados en densidad, tenemos un clúster por cada triángulo (3 clústers).

Clúster (d): Basados en el centro, tenemos que cada grupo de líneas forma un clúster (2 clúster en total). Basados en contigüidad, cada set de líneas forma un clúster (5 clúster en total). Basados en densidad, los 2 grupos de líneas definen 2 regiones separadas de alta densidad (2 clúster en total).

4.2. ¿Sería la medida del coseno la medida de similitud adecuada para usar con el agrupamiento de K-medias para datos de series de tiempo? ¿Por qué o por qué no? Si no, ¿qué medida de similitud sería más apropiada?

La medida del coseno es adecuada para datos dispersos. Como los datos de series de tiempo son datos densos de alta dimensión, la medida del coseno no es adecuada.

Si consideramos que la magnitud de tiempo de la serie es importante, sería más apropiada la distancia euclídea.

Si sólo son importantes las formas de la serie de tiempo, sería apropiada la correlación.

Si la comparación de la serie de tiempo debe tener en cuenta que una serie de tiempo puede llevar o retrasar a otra, o solo estar relacionada con otra durante períodos de tiempo específicos, se deben usar enfoques más sofisticados para modelar la similitud de la serie de tiempo.

4.3. El algoritmo líder (Hartigan [4]) representa cada grupo utilizando un punto, conocido como líder, y asigna cada punto al grupo correspondiente al líder más cercano, a menos que esta distancia esté por encima de un umbral especificado por el usuario. En ese caso, el punto se convierte en el líder de un nuevo grupo.

4.3.1. ¿Ventajas y desventajas del algoritmo líder en comparación a K-medias?

Ventajas: El algoritmo líder sólo requiere un escaneo de los datos y, por lo tanto, es más eficiente desde el punto de vista computacional. Aunque el algoritmo líder es dependiente del orden, para un ordenamiento fijo de los objetos, siempre produce el mismo conjunto de conglomerados.

Desventajas: No es posible establecer el número de grupos resultantes, excepto indirectamente, para el algoritmo líder. Además, el K-medias casi siempre produce mejores agrupamientos de calidad medidos por SSE.

4.3.2. Sugiera formas en las que se podría mejorar el algoritmo líder.

- Utilizar una muestra para determinar la distribución de distancias entre los puntos, de modo que el conocimiento obtenido de este proceso se pueda utilizar para establecer de manera más inteligente el valor del umbral.
- El algoritmo líder podría modificarse para agrupar varios umbrales durante un solo paso.

4.4. Supongamos que encontramos K conglomerados usando el método de Ward, bisectando K-medias y K-medias ordinarias. ¿Cuál de estas soluciones representa un mínimo local o global? Explique.

Aunque el método Ward elige un par de grupos para fusionarlos en función de minimizar SSE, no dispone de un paso de refinamiento como K-medias ordinarias. Del mismo modo,

la bisección K-medias no tiene un paso de refinamiento general.

Por lo tanto, a menos que se agregue un paso de refinamiento de este tipo, sólo el método k-medias ordinario genera un mínimo local. Sin embargo, ningún método de los indicados garantiza que se produzca un mínimo global.

4.5. Los algoritmos de agrupamiento jerárquico requieren un tiempo de O(m2 log(m)) y, en consecuencia, no son prácticos para usar directamente en conjuntos de datos más grandes. Una posible técnica para reducir el tiempo requerido es muestrear el conjunto de datos. Los grupos K se pueden extraer de este agrupamiento jerárquico tomando los grupos en el nivel K-ésimo del dendrograma. Luego, los puntos restantes se pueden asignar a un grupo en tiempo lineal, utilizando varias estrategias. Para cada uno de los siguientes tipos de datos o conglomerados, discuta brevemente si (1) el muestreo causará problemas para este enfoque y (2) cuáles son esos problemas. Enfóquese solo en los problemas causados por la característica particular mencionada. Finalmente, suponga que K es mucho menor que m.

4.5.1. Datos con clústeres de muy diferente tamaño.

Esto puede ser un problema, especialmente si la cantidad de puntos en un grupo es muy pequeña, debido a que un número menor de datos es más fácil de perder o agrupar incorrectamente que un número mucho mayor de datos, además de que el número de datos de menor tamaño puede estar, a veces, representado por un número menor de datos debido a la naturaleza del problema.

4.5.2. Alta dimensionalidad de los datos

Puede ser un problema porque los datos en alta dimensionalidad suelen ser escasos y es posible que sean necesarios más puntos para definir la estructura de un grupo en el espacio de alta dimensionalidad.

4.5.3. Datos con valores atípicos, es decir, puntos atípicos.

Puede ser un problema si encontrar el agrupamiento correcto depende de la presencia de valores atípicos, pues estos valores no son muy frecuentes y suelen omitirse. De lo contrario, es beneficioso.

4.5.4. Datos con regiones muy irregulares.

Esto puede ser un problema porque la estructura del borde puede perderse al muestrear ,a menos que se muestreen una gran cantidad de puntos.

4.5.5. Datos con cúmulos globulares.

Por lo general, esto no es un problema, ya que no es necesario muestrear tantos puntos para conservar la estructura de un cúmulo globular como uno irregular.

4.5.6. Datos con densidades muy diferentes.

En este caso, los datos tenderán a provenir de la región más densa. Para los conglomerados que, para empezar, no son muy densos, esto puede significar que ahora se tratan como ruido o valores atípicos.

4.5.7. Datos con un pequeño porcentaje de puntos de ruido.

Dado que nos interesa excluir el ruido, y dado que la cantidad de ruido es pequeña, esto podría ser beneficioso.

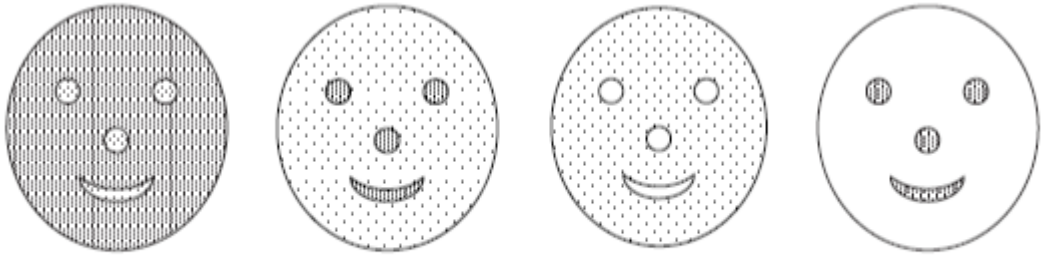
4.5.8. Datos no Euclidianos o Euclidianos

Esto no tiene un impacto particular.

4.5.9. Datos con muchos tipos de atributos y mixtos

Muchos atributos fueron discutidos bajo alta dimensionalidad, y los mixtos no tienen un impacto particular.

4.6. Considere las siguientes cuatro caras que se muestran en la figura 8.7. Nuevamente, la oscuridad o el número de puntos representa la densidad. Las líneas se usan solo para distinguir regiones y no representan puntos.



(a) (b) (c) (d)

4.6.1. Para cada figura, ¿podrías usar un solo enlace para encontrar los patrones representados por la nariz, los ojos y la boca? Explique.

Sólo para (b) y (d). Para (b), los puntos en la nariz, los ojos y la boca están mucho más juntos que los puntos entre estas áreas. Para (d) solo hay espacio entre estas regiones.

4.6.2. Para cada figura, ¿podrías usar K-medias para encontrar los patrones representados por la nariz, los ojos y la boca? Explique.

Sólo para (b) y (d). Para (b), K-means encontraría la nariz, los ojos y boca, pero también se incluirían los puntos de menor densidad. Para (d), K-means encontraría la nariz, los ojos y la boca directamente siempre que el número de grupos se establezca en 4.

4.6.3. ¿Qué limitación tiene el agrupamiento para detectar todos los patrones formados por los puntos de la figura 8.7(c)?

Las técnicas de agrupamiento sólo pueden encontrar patrones de puntos, no de espacios vacíos.

4.7. Una forma de dispersar una matriz de proximidad es la siguiente: para cada objeto (fila en la matriz), establezca todas las entradas en 0 excepto las correspondientes a los objetos k-vecinos más cercanos. Sin embargo, la matriz de proximidad dispersa normalmente no es simétrica.

4.7.1. Si el objeto a está entre los k vecinos más cercanos del objeto b, ¿por qué no se garantiza que b esté entre los k vecinos más cercanos de a?

Considere un conjunto denso de k + 1 objetos y un valor atípico, que está más lejos de cualquiera de los objetos que están entre ellos. Ninguno de los objetos en el conjunto denso tendrá el valor atípico en su lista de k vecinos más cercanos, pero el valor atípico tendrá k de los objetos del conjunto denso en su lista de k vecinos más cercanos.

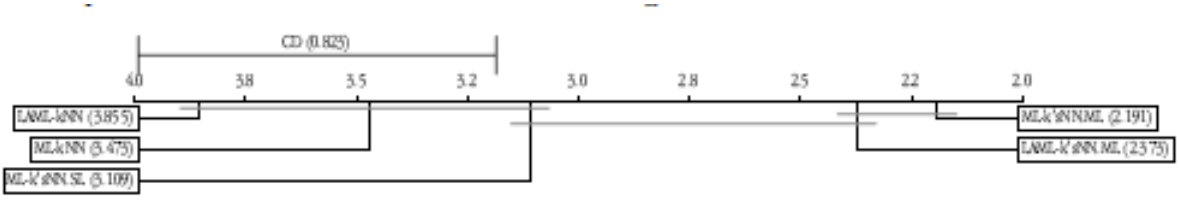
4.7.2. Sugiera al menos dos enfoques que podrían usarse para hacer simétrica la matriz de proximidad dispersa.

- Establecer la i-ésima entrada en 0 si la j-ésima entrada es 0, o viceversa.
- Establecer la i-ésima entrada en 1 si la j-ésima entrada es 1, o viceversa.

5. Cuestiones de examen

5.1. Enero de 2021

La gráfica siguiente muestra los rangos de Friedman y la diferencia crítica de un test de Nemenyi cuando se compara el rendimiento en clasificación de 5 algoritmos:



¿Qué información sobre el rendimiento de cada uno de los algoritmos y su relación entre ellos puede obtenerse de la gráfica? Enunciar para qué sirve o cómo funciona el test

de Nemenyi no se valora en esta pregunta. Se han de obtener conclusiones del ejemplo mostrado.

Solución:

- Los rangos de Friedman nos dan una ordenación de los algoritmos de mejor a peor rendimiento (derecha a izquierda).
- La diferencia crítica indica grupos de algoritmos cuyo rendimiento no tiene diferencias como serían los 2 más a la derecha o los 3 más a la izquierda.

5.1.1. Considere un problema de clasificación en 2 clases, Valor = Bajo,Alto, con los siguientes atributos:

- Aire Acondicionado = Funcionando, Bajo
- Motor = Bueno, Malo
- Kilometraje = Alto, Medio, Bajo
- Corrosión = sí, No

Considere un clasificador basado en reglas con el siguiente conjunto de reglas:

- Kilometraje = Alto $\rightarrow$ Valor = Bajo.
- Kilometraje = Bajo $\rightarrow$ Valor = Alto.
- Aire Acondicionado = Funcionando, Motor = Bueno $\rightarrow$ Valor = Alto.
- Aire Acondicionado = Funcionando, Motor = Malo $\rightarrow$ Valor = Bajo.
- Aire Acondicionado = Roto $\rightarrow$ Valor = Bajo.

Responda a las siguientes cuestiones:

a) ¿Son las reglas mutuamente excluyentes?

No, porque hay hay 2 reglas que se activan por el mismo registro (Aire Acondicionado = Funcionando).

b) ¿Es el conjunto de reglas exhaustivo?

Sí, porque la regla Aire acondicionado cubre cualquier combinación de conjuntos de atributos.

c) ¿Es necesario ordenar las reglas?

Sí, porque hay instancias que disparan más de una regla.

d) ¿Es necesario definir una clase por defecto para el conjunto de reglas?

No, porque el conjunto es exhaustivo.

5.1.2. En una regla de asociación X → Y se define el concepto Lift = $\frac{P(Y/X)}{P(Y)}$. Explique cuál es la utilidad de este concepto para evaluar una regla en relación a los conceptos de soporte y confianza de una regla.

Nota: Enunciar los conceptos de soporte, confianza o Lift no se valora en esta pregunta. Se ha de responder a la utilidad del concepto de Lift.

RESPUESTA: El concepto de Lift tiene como objetivo evitar el problema que surge con el concepto de confianza al no tener en cuenta el soporte del consecuente. Reglas cuyo consecuente tiene un soporte muy alto pueden tener una confianza artificialmente alta. En ese caso, Lift tendrá un valor inferior a 1 indicando que la regla no es útil.

5.1.3. Una de las técnicas aplicables en reglas de asociación consiste en usar múltiples umbrales de soporte mínimo para considerar un itemset frecuente en lugar de un único mínimo global.

¿Qué utilidad e inconvenientes puede tener esta práctica? Ilustre la explicación con un ejemplo.

RESPUESTA: El uso de múltiples soportes es útil para evitar que los itemsets con algún ítem muy poco frecuente sean siempre eliminados (estos itemsets podrían tener información valiosa). El efecto que tiene sobre la poda a priori es que la regla de antimonotomía del soporte se deja de cumplir de forma general y hay que modificar la poda.

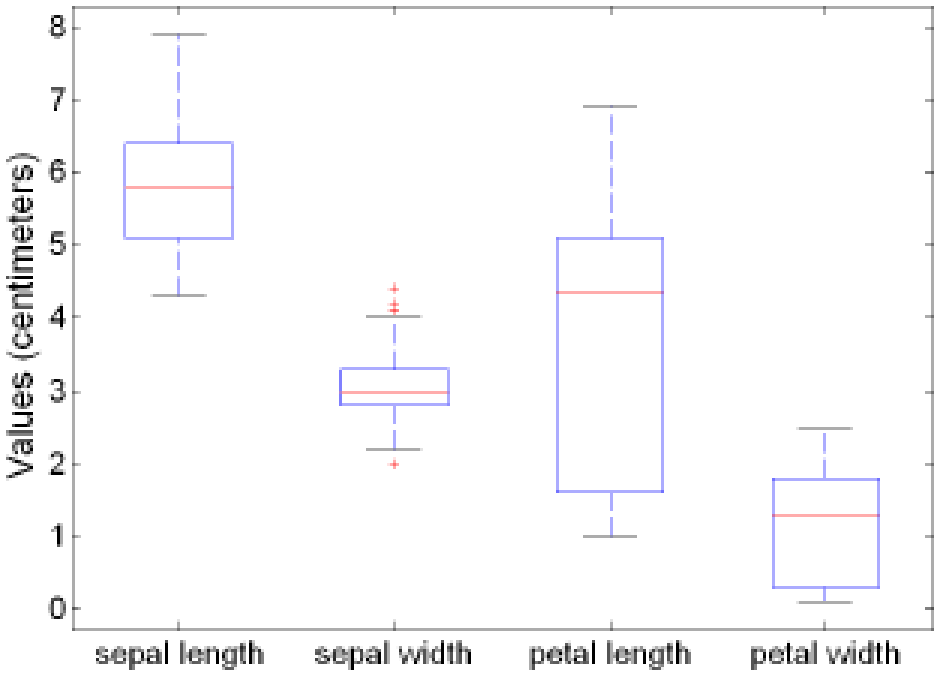
5.2. Enero 2017

5.2.1. Indique 2 formas de gestionar los valores perdidos en un conjunto de datos. Indica qué ventajas e inconvenientes ve en cada una de ellas.

RESPUESTA:

- Ignorar los valores perdidos: Se puede hacer a nivel de instancia, ignorando todas las instancias con al menos un valor perdido, o a nivel de variable, ignorando aquellas variables con al menos un valor perdido en una instancia. El problema es que podemos perder mucha información.
- Estimar los valores perdidos: Se estiman los valores perdidos usando modas, medianas, medias, estimaciones estadísticas o vecinos más cercanos. El problema es que introducimos ruido en la muestra.

5.2.2. Considere el box plot del dataset iris. ¿Qué información podemos obtener respecto al comportamiento de las variables?



RESPUESTA: Podemos observar que la variable petal lenght tiene una gran dispersión al igual que, en menor medida, las variables petal width y sepal length. Por el contra-

rio, sepal width tiene valores muy homogéneos entre los diferentes patrones. Respecto al comportamiento de las variables, de un gráfico box plot no podemos deducir nada, pues una variable más homogénea puede ser más discriminante que una variable con mayor dispersión.

5.2.3. Indique un aspecto positivo y otro negativo de un árbol de decisión, una SVM y el vecino más cercano

Árbol de decisión:

- Aspecto positivo: Capaces de tratar con problemas muy grandes, muy rápidos en clasificación, interpretables cuando son pequeños, buena relación coste/rendimiento.
- Aspecto negativo: Menor rendimiento que otros métodos, inestables.

SVM:

- Aspecto positivo: Pueden ser muy eficientes con conjuntos de datos con miles de variables, muy buen rendimiento, robustos ante la presencia de ruido, estables.
- Aspecto negativo: Muy costosos computacionalmente, muy sensibles a los parámetros de entrenamiento.

Vecino más cercano:

- Aspecto positivo: No necesitan entrenamiento, buen rendimiento, estables ante variaciones en el conjunto de instancias.
- Aspecto negativo: Necesitan almacenar el conjunto de entrenamiento completo por lo que tienen problemas de escalabilidad, inestables ante variaciones en el conjunto de variables.

5.2.4. En la construcción de reglas de asociación, ¿qué efecto tiene el uso de un soporte mínimo variable según los ítems en un itemset sobre el algoritmo Apriori?

RESPUESTA: El soporte pierde la propiedad de anti-monotomía y, por lo tanto, el algoritmo Apriori deja de ser aplicable porque está basado en dicha propiedad. Existen

diferentes modificaciones del algoritmo para poder seguir aplicandolo, aunque con menor efectividad.

5.2.5. Indique dos puntos fuertes del agrupamiento jerárquico con respecto al particional

RESPUESTA:

- No asume un número determinado de clústers en el conjunto de datos.
- El resultado es una taxonomía de las instancias que puede ser de mucha utilidad en muchas áreas de conocimiento.

5.3. Enero 2016

5.3.1. ¿Qué significado tienen desde el punto de vista intuitivo las medidas de error sensibilidad y especificidad para problemas de clasificación de dos clases?

RESPUESTA: La sensibilidad mide la capacidad que tiene un clasificador para no errar en la identificación de positivos clasificándolos como negativos. La especificidad mide la capacidad de no clasificar los datos erróneamente como positivos si son negativos.

5.3.2. ¿En qué consiste el sobreaprendizaje (overfitting) en la construcción de un clasificador? ¿Es posible evitarlo?

RESPUESTA: El sobreaprendizaje ocurre cuando un clasificador aprende muy bien el conjunto de entrenamiento a costa de perder su capacidad de generalización. No existen métodos para evitarlo de forma consistente aunque si hay técnicas para tratar de atenuar su efecto, como el uso de modelos más simples la detención prematura del entrenamiento mediante validación cruzada.

5.3.3. ¿Puedo resolver un problema de clasificación de N clases (N >2) si tengo un método de clasificación que solo puede distinguir entre dos clases?

RESPUESTA: Sí, se puede transformar el problema de N clases en M problemas de dos clases. Métodos conocidos son el one-vs.-one, el one-vs.all o los códigos ECOC.

5.3.4. Indique cómo llevaría a cabo la comparación de los métodos siguientes de clasificación

- Comparación de dos métodos sobre un conjunto de N problemas: Se aplicaría el test de Wilcoxon.
- Comparación de un método contra un serie de métodos estándar sobre un conjunto de N problemas para ver si es mejor que todos ellos: Primero aplicaríamos el test de Friedman o Iman-Davenport para comprobar si hay diferencias significativas. En caso afirmativo, se aplicaría el procedimiento de Holm para comparar nuestro método con cada uno de los métodos estándar paso a paso.

5.3.5. ¿Qué tipo de clústers tiende a generar un metodo de clustering particional como por ejemplo k-medias?

RESPUESTA: Genera normalmente clústers homogéneos y de forma globular, es por ello que funciona pobremente si nuestros clústers no corresponden a esta forma.

5.4. ENERO 2023

5.4.1. Cuestión 1

La medida de confianza de una regla de asociación puede dar lugar a evaluaciones erróneas de algunas reglas, esto es, evaluar una regla con un nivel alto de confianza cuando el hecho que representa no ocurre en la realidad. ¿En qué condiciones puede ocurrir este hecho? Indique un ejemplo ilustrativo. ¿Qué medida alternativa se puede usar para evitar este problema?

RESPUESTA: En la medida de confianza de una regla X $\rightarrow$Y el soporte del consecuente no se considera. Cuando dicho soporte es alto, la regla puede tener un valor de confianza

relativamente alto. Este problema se puede evitar aplicando medidas como lift.

5.4.2. Cuestión 2

¿Qué ventaja tiene el uso de $Gain_{ratio}$ con respecto al uso de Gain para obtener el mejor split en un nodo?.

RESPUESTA: La ganancia de información tiende a preferir las características con más categorías, pues tienden a tener una entropía más baja, lo que origina un sobre-aprendizaje de los datos de entrenamiento. $Gain_{ratio}$ mitiga este problema al penalizar las características por tener más categorías usando la fórmula de split info.

5.4.3. Cuestión 3

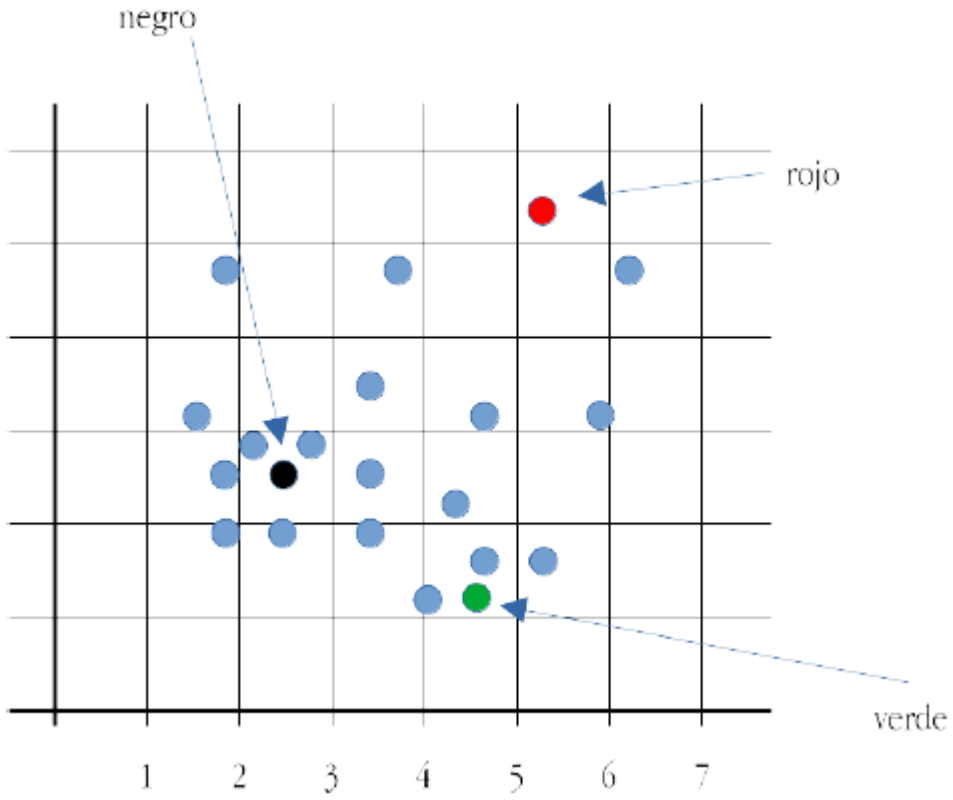
Los métodos de agrupaciones de clasificadores basados en bagging y boosting tienen un efecto diferente en el error de sesgo y varianza de los clasificadores. Explique dicha diferencia e indique la razón del comportamiento diferente. ¿Cuál de los dos métodos, bagging o boosting, tiene mayor potencial para mejorar el rendimiento de un clasificador?

RESPUESTA: Los métodos de bagging mejoran el factor de la varianza del error pues solo realizan un muestreo aleatorio que no cambia el sesgo del clasificador. Los métodos de boosting mejoran la varianza, por la misma razón que bagging, pero además mejoran el error de sesgo pues obligan al clasificador a centrarse en los patrones más difíciles.

El error de varianza es normalmente menor en el total del error, por eso boosting tiene mayor potencial de mejora del acierto.

5.4.4. Cuestión 4

Considere el concepto de puntos core, border y outlier del algoritmo DBSCAN. Indique de qué tipo serían los puntos negro, rojo y verde la siguiente figura con Eps=1 y MinPts=3.



RESPUESTA: El punto rojo es un outlier, pues no hay ningún otro punto dentro de su radio. Los puntos negro y verde son puntos core porque tienen en su radio al menos 3 puntos.