UNIVERSIDAD
DE GRANADA

Facultad de Ciencias

Escuela Técnica Superior de Ingeniería
Informática y Telecomunicaciones

Doble Grado de Ingeniería
Informático y Matemáticas

TRABAJO DE FIN DE GRADO

Estudio de la información mutua y el test delta como criterios para la selección de variables.

Presentado por:

Iñaki Madinabeitia Cabrera

Curso académico 2019-2020

Estudio de la información mutua y el test delta como criterios para la selección de variables.

Iñaki Madinabeitia Cabrera

Iñaki Madinabeitia Cabrera *Estudio de la información mutua y el test delta como criterios para la selección de variables.*.

Trabajo de fin de Grado. Curso académico 2019-2020.

Responsable de tutorización

Alberto Guillén Perales
Departamento de Arquitectura y Tecnología de Computadores

Doble Grado de Ingeniería Informático y Matemáticas

Facultad de Ciencias

Escuela Técnica Superior de Ingeniería Informática y Telecomunicaciones

Universidad de Granada

Declaración de originalidad

D./Dña. Iñaki Madinabeitia Cabrera

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2019-2020, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 5 de septiembre de 2019

Fdo: Iñaki Madinabeitia Cabrera

Dedicatoria (opcional)
Ver archivo `preliminares/dedicatoria.tex`

Índice general

Agradecimientos

Agradecimientos del libro (opcional, ver archivo `preliminares/agradecimiento.tex`).

Summary

An english summary of the project (around 800 and 1500 words are recommended).

File: `preliminares/summary.tex`

Introducción

De acuerdo on la comisión de grado, el TFG debe incluir una introducción en la que se describan claramente los objetivos previstos inicialmente en la propuesta de TFG, indicando si han sido o no alcanzados, los antecedentes importantes para el desarrollo, los resultados obtenidos, en su caso y las principales fuentes consultadas.

Ver archivo `preliminares/introduccion.tex`

Parte I.

Resumen.

Si el trabajo se divide en diferentes partes es posible incluir al inicio de cada una de ellas un breve resumen que indique el contenido de la misma. Esto es opcional.

Parte II.

Objetivos.

1. Objetivos

Los objetivos clave que se abordan en este proyecto son los siguientes:

- Análisis comparativo de estimadores de información mutua.
- Optimización de estimadores de información mutua.
- Idea e implementación de un cálculo óptimo del Test Delta en un espacio de $2^d - 1$ soluciones, donde d es la dimensión del problema.
- Análisis comparativo entre la información mutua y el Test Delta.
- Construcción de modelos que determinen el subconjunto de variables más adecuado para problemas de regresión.
- Propuesta de integración de la información mutua y el test Delta.
- Análisis de dicha propuesta.

Parte III.

Estado del arte.

2. Información Mutua

2.1. Resultados previos

Teorema 2.1. ***Teorema de Radon-Nikodym.*** *Sea (X, E) un espacio medible, con dos medidas σ-finitas μ y v. Si v es absolutamente continua respecto de μ, entonces existe una función medible $f : X \to [0, \infty[$ tal que para todo conjunto medible $A \subset X$:*

$$v(A) = \int_A f d\mu$$

Definición 2.1. La f definida en el Teorema 2.1 es única casi por doquier. A esta función se le denomina **la derivada de Radon-Nikodym**.

2.2. Entropía de Shannon y entropía conjunta

La idea básica de la entropía en teoría de la información es que un mensaje contiene mayor entropía cuanta más sorpresa genere su contenido. En otras palabras, si un evento es muy probable que ocurra, dicho evento tiene poca entropía pues si ocurriese no generaría ningún especial interés, ya que ocurrió lo que se esperaba.

Antes de preguntarnos cómo medir esta información sobre una variable aleatoria, vamos a preguntarnos cómo podemos medir la información que nos proporciona un suceso. En computación, resulta interesante expresar la información en unidades de bits, luego esto nos motiva a utilizar el logaritmo en base de 2. Sin embargo, ¿cuál es el parámetro en dicho logaritmo en base 2? Teniendo una probabilidad, junto con lo dicho en el anterior párrafo de que un suceso nos da más información cuanto menos probable es que ocurra (inversamente proporcional a la probabilidad del suceso), nos queda que:

Definición 2.2. Sea un suceso E y una probabilidad P (cuyo dominio contiene a E), definimos el **contenido de la información** del suceso bajo probabilidad P como: [McMo8]

$$I(E) := \log_2 \left(\frac{1}{P(E)} \right) = -\log_2 \left(P(E) \right)$$

Procedemos a establecer cómo medir la información que nos da una variable aleatoria, mediante la entropía de Shannon:

Definición 2.3. Sea una variable aleatoria X que toma valores x_i con $i = 1, 2, ..., n$, y sea P una distribución de probabilidad (cuyo dominio contiene a los valores de X), definimos la **entropía de Shannon**: [Sha8x]

$$H(X) := E[-\log_2(P(X))] = -E[\log_2(P(X))]$$

Vemos que en el caso discreto, siendo p la función masa de probabilidad de P (esto es, $p(x_i) = P[X = x_i]$), nos queda que:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i))$$

Habiendo definido la entropía de una variable aleatoria, podemos extender la definición a una pareja de variables aleatorias de forma sencilla: podemos tomar dicha pareja como un vector de variables aleatorias.

Definición 2.4. La **entropía conjunta** de dos variables aleatorias X e Y con una distribución conjunta P es definida como: [CT]

$$H(X, Y) := -E[\log_2(P(X, Y))]$$

Vemos que en el caso discreto, siendo p la función masa de probabilidad de P (esto es, $p(x, y) = P[X = x, Y = y]$), nos queda que:

$$H(X, Y) = -\sum_{x \in X}\sum_{y \in Y} p(x, y)\log_2(p(x, y))$$

Notación. De ahora en adelante, nos referiremos a log como $\log_2$.

2.3. Divergencia de Kullback-Leibler

En estadística, la divergencia de Kullback-Leibler entre dos distribuciones de probabilidad es una cantidad que mide cuán diferente es una distribución respecto de la otra [KL68].

Definición 2.5. Sean P y Q distribuciones de probabilidad que toman valores en el mismo espacio X, y además P es absolutamente continua respecto de Q ($\forall x \in X,\ Q(x) = 0 \implies P(x) = 0$), la **divergencia de Kullback-Leibler** es la esperanza de la diferencia logarítmica entre P y Q, donde se toma la esperanza usando las probabilidades de P. Es decir:

$$D_{KL}(P||Q) = \int_X \log\left(\frac{dP}{dQ}\right) dP$$

donde $\frac{dP}{dQ}$ es la derivada de Radon-Nikodym (Def. 2.1) de P respecto de Q.

Notación. Sea X una variable aleatoria y P una distribución de probabilidad, nos referimos con $P(x_i)$ a $P[X = x_i]$.

En el caso discreto, tenemos que la divergencia de Kullback-Leibler es:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Que equivalentemente es:

$$D_{KL}(P||Q) = - \sum_{x \in X} P(x) \log \left(\frac{Q(x)}{P(x)} \right)$$

Pues:

$$\sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) = \sum_{x \in X} P(x) \left(- \log \left(\frac{Q(x)}{P(x)} \right) \right)$$
$$= \sum_{x \in X} -P(x) \log \left(\frac{Q(x)}{P(x)} \right) = - \sum_{x \in X} P(x) \log \left(\frac{Q(x)}{P(x)} \right)$$

En el caso continuo, tenemos que:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

donde p y q son las funciones densidad de P y Q.

2.4. Definición y propiedades de la Información Mutua

La información mutua entre dos variables aleatorias es una medida que cuantifica la cantidad de información (en bits) que se obtiene de una variable aleatoria observando la otra.

Definición 2.6. Sean X e Y dos variables aleatorias cuya distribución conjunta es $P_{X,Y}$ y cuyas distribuciones marginales son P_X y P_Y respectivamente, se define la **información mutua** como:

$$I(X, Y) := D_{KL}(P_{X,Y}||P_X P_Y)$$

En el caso discreto, si X toma n valores e Y toma m valores, tenemos que

$$I(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_i) \log \left(\frac{P(x_i, y_i)}{P(x_i)P(y_i)} \right)$$

En el caso continuo, siendo p la función de densidad de P, tenemos que

$$I(X,Y) = \int_X \int_Y p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dxdy$$

La idea clave para entender la información mutua es la siguiente: la información mutua entre dos variables aleatorias es la intersección entre la entropía de una y la entropía de la otra. Es decir, $I(X,Y) = H(X) \cap H(Y)$. En la figura Figura 2.1 podemos visualizar esta idea.

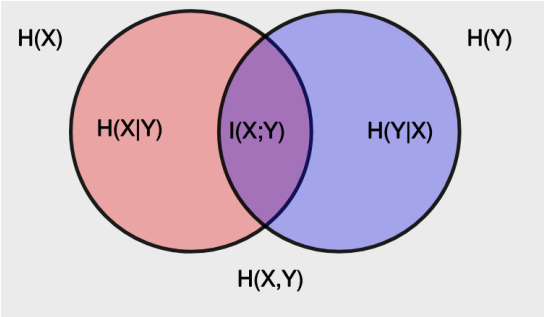

Figura 2.1.: Información mutua como intersección de $H(X) \cap H(Y)$ [Mac].

En definitiva, la información mutua mide el grado de dependencia estadística entre dos variables, luego es una medida que reduce la incertidumbre sobre el valor de una variable conociendo el valor de la otra variable.

Hay un importante resultado que nos dice que la información que X tiene acerca de Y no aumenta si se realiza una transformación determinista sobre X. Por tanto, el objetivo es **usar la información mutua para reducir la dimensión conservando la mayor cantidad de información posible.**

3. Test Delta

Una técnica que no usa la información mutua y que al igual que la información mutua sirve para realizar una selección de variables, que veremos más adelante, es el **Test Delta**.

3.1. Test Delta

El Test Delta es una técnica para estimar la varianza del ruido, o el error cuadrático medio (MSE), que se puede lograr sin sobreajustamiento [Jono4]. Es útil para evaluar la correlación no lineal entre dos variables aleatorias, que son las parejas de entradas y salida. Es una generalización del enfoque propuesto en [Jono4], que básicamente se basa en el hecho de que la esperanza condicional se acerca la varianza del ruido cuando la distancia entre los puntos de los datos tiende a cero. Veremos esto más adelante.

El Test Delta trabaja bajo la suposición de que si dos puntos están cerca en el espacio de entrada, entonces sus correspondientes puntos de salida deben estar cerca. Si no es así, asumimos que es debido al ruido [KWW].

El único requisito es que el sistema sea suave (que la transformación de entrada a salida sea continua y que la primera derivada parcial respecto del espacio de entrada esté acotada) [KWW].

3.1.1. Estimación de la varianza del ruido

Sea el conjunto de datos de la forma:

$$\{(x_i, y_i) t.q, 1 \leq i \leq N\}$$

donde x_i la entrada como vector en $\mathbb{R}^d$ y y_i la salida como escalar en $\mathbb{R}$, asumimos con sistema suave que los valores de salida siguen la siguiente forma:

$$y_i = f(x_{i1}, ..., x_{id}) + r_i$$

donde f es una función suave (derivadas continuas en cualquier orden) y r_i es un valor aleatorio que representa el ruido [KWW]. Los términos del ruido r_i asumimos que son independientes idénticamente distribuidos con media cero [ELL$^+$].

La estimación de la varianza del ruido es el estudio que trata de dar una estimación de la varianza del ruido dado algunos datos sin considerar aspectos específicos de la función f [ELL$^+$].

3.1.2. Vecino más cercano

El estimador de la varianza del ruido considerado está basado en una aproximación mediante el vecino más cercano. El NN (*nearest neighbour*, vecino más cercano) de un punto x está definido como el *único* punto que minimiza la distancia métrica respecto a x en el espacio de entrada [ELL+].

$$NN(i) := arg\min_{j \neq i} ||x_i - x_j||^2$$

Observamos que $NN(i)$ devuelve el índice j del x_j más cercano a x_i.

En este contexto, usamos la distancia euclídea como distancia métrica.

3.1.3. Definición de Test Delta

El Test Delta se basa en realizar una estimación del ruido entre pares de puntos de entrada y salida, y aplicar el Test Delta para obtener el conjunto de variables que minimice el ruido entre los puntos.

Definición 3.1. Sea X una variable que toma N valores en $\mathbb{R}^d$ (cada elemento es un vector de dimensión d), sea Y una variable que toma N valores en $\mathbb{R}$, donde N es el número de muestras y d la dimensión de la entrada, y sea $NN(i,k)$ el índice del k-ésimo vecino más cercano a x_i; se define el **Test Delta** como: [PP94]

$$\delta_{N,k} := \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - y_{NN(i,k)} \right)^2$$

Luego para el vecino más cercano, nos quedaría ($k = 1$):

$$\delta_N := \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - y_{NN(i)} \right)^2$$

Es decir, aproximamos la varianza del ruido considerando la diferencia de las salidas asociándolo con la cercanía de los puntos en el espacio de entrada [ELL+]. Es decir:

$$Var(r) \approx \delta_N$$

3.1.4. Demostración del Test Delta como buen estimador

Acotación.

Sean X_i $(1 \leq i \leq M)$ variables aleatorias independientes idénticamente distribuidas de acuerdo a una acotada probabilidad de densidad p (por ejemplo, $p \leq ||p||_\infty$), siendo M el número de muestras. Asumimos que el rango de las variables está incluida en un conjunto compacto C contenido en $\mathbb{R}^n$. Con $NN(i,k)$ denotamos el k-ésimo vecino más cercano de X_i, y $d_{i,k}$ es $||X_i - X_{NN(i,k)}||$ con la norma euclídea. Sean $B(x_0, r) := \{x \in \mathbb{R}^n \text{ t.q. } ||x - x_0|| < r\}$ bolas de vecindario en $\mathbb{R}^n$, definimos la función:

$$\omega_{x_0}(r) = \int_{B(x_o,r)} p(x)dx$$

que corresponde a la probabilidad de que un punto de C esté contenido en la bola $B(x_0, r)$.

Proposición 3.1. *Para cada $\alpha > 0$*

$$E[\omega_{X_i}(d_{i,k})^\alpha | X_i] = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)}$$

donde $\Gamma(\cdot)$ es la función Gamma y α se refiere al α-ésimo momento de las distribuciones de la distancia al vecino más cercano [Eva],[KSG].

La demostración está basada en las propiedades de la función Beta establecidas en [Eva], donde también se muestra que

$$\frac{\Gamma(M)}{\Gamma(M+\alpha)} = M^{-\alpha} + O(M^{-\alpha-1})$$

denotando el término constante

$$c(k, \alpha, n) := \frac{\Gamma(k + \frac{\alpha}{n})}{\Gamma(k)}$$

y V_n como la medida de Lebesgue de la bola unidad en $\mathbb{R}^n$ (su volumen).

Con todo esto, pasamos a demostrar la proposición que necesitamos.

Proposición 3.2. *Para la constante $c(k, \alpha, n) \geq 1$,*

$$E[d_{i,k}^\alpha] \geq c(k, \alpha, n) V_n^{-\alpha/n} ||p||_\infty^{-\alpha/n} \frac{\Gamma(M)}{\Gamma(M + \alpha/n)}$$

Demostración. Una manipulación algebraica de $E[d_{i,k}^{\alpha}]$ nos lleva a

$$E[d_{i,k}^{\alpha}] \geq L(X_i)^{-\alpha/n} E[\omega_{X_i}(d_{i,k})^{\alpha/n}|X_i]$$

donde

$$L(x) = \sup_{0<r<\infty} \omega_x(r)/r^n$$

Con la proposición 1,

$$E[\omega_{X_i}(d_{i,k})^{\alpha/n}|X_i] = c(k,\alpha,n)\frac{\Gamma(M)}{\Gamma(M+\alpha/n)}$$

De $L(X_i) \leq ||p||_{\infty} V_n$ sacamos que

$$E[(d_{i,k}^{\alpha})|X_i] \geq c(k,\alpha,n) V_n^{-\alpha/n} ||p||_{\infty}^{-\alpha/n} \frac{\Gamma(M)}{\Gamma(M+\alpha/n)}$$

El hecho de que $E[d_{i,1}]$ es del orden de $M^{-1/n}$ nos muestra que cuando la dimensión del espacio de entrada crece, la distancia media al vecino más cercano tiende lentamente a cero cuando el número de muestras incrementa.

3.1.5. Observación final

Esto nos deja que, cuando N tiende a infinito, δ_N tiende a la varianza del ruido en la salida [PP94]. Por tanto, cuanto más grande sea N, con mayor seguridad podremos afirmar que la estimación nos da resultados fiables [PP94].

Parte IV.

Problema de la selección de variables.

4. Problema de la selección de variables

La selección de variables (*input selection*) es uno de los mayores problemas en Aprendizaje Automático, especialmente cuando el número de características es relativamente grande comparado con el número de observaciones.

Matemáticamente hablando, un conjunto finito de variables es suficiente para extraer un modelo exacto de infinitas observaciones. En la práctica, no tenemos infinitas observaciones, por tanto el tamaño necesario de variables incrementa drásticamente con el número de observaciones. [RHJL]

4.1. Maldición de la dimensionalidad

En estadística, la maldición de la dimensionalidad se refiere a varios fenómenos que ocurren al analizar y organizar datos de grandes dimensiones que no suceden en dimensiones pequeñas.

Un ejemplo es que, si quisiéramos hallar cuántos puntos son suficientes para representar un cubo unidimensional (equivalentemente, un intervalo de longitud 1) para que la distancia entre cualesquiera dos puntos cercanos sea 10^{-2}, tendríamos que resolver esta simple ecuación (siendo $x > 0$ la cantidad de puntos):

$$\frac{1}{x} = 10^{-2}$$

luego bastan 10^2 puntos, pues $1/10^2 = 10^{-2}$. Sin embargo, si nos preguntamos lo mismo pero en el caso de un hipercubo de diez dimensiones, resolveríamos la misma ecuación por cada dimensión y entonces bastarían $10^{2^{10}} = 10^{20}$ puntos.

En aprendizaje automático se da el **fenómeno de Hughes** [Hug59], que nos dice que teniendo el número de muestras de entrenamiento fijo, el poder predictivo del clasificador o regresor primero tiene una relación directamente proporcional al número de dimensiones (aumenta cuando aumenta el número de dimensiones), pero luego se vuelve indirectamente proporcional (disminuye al aumentar el número de dimensiones).

4.2. Selección de variables

Ahora, añadimos la premisa fundamental para resolver la maldición de dimensionalidad seleccionando variables:

Premisa 4.1. Los datos contienen características que son redundantes o irrelevantes. Por tanto, dichas características pueden ser eliminadas sin que esto conlleve a una gran pérdida de información.

En aprendizaje automático y en estadística, la selección de variables consiste en escoger un subconjunto de características relevantes, evitando aquellas redundantes o irrelevantes. Además de evitar la *maldición de la dimensionalidad*, también encontramos estas propiedades:

- Simplifica modelos.
- Agiliza los períodos de entrenamiento.
- Favorece la generalización y evita el sobreentrenamiento.

4.3. Ejemplo de algoritmo que usa Información Mutua: mRMR

Mínima redundancia máxima relevancia, o también conocido como **mRMR**, es un algoritmo que selecciona un subconjunto de características que mejor caracteriza las propiedades estadísticas de una variable de clasificación objetivo (target), sujeto a las restricciones de que estas características sean tan diferentes entre sí como sea posible, pero marginalmente sean lo más similar posible a la variable de clasificación objetivo. Es decir, que estén muy alejadas unas características de las otras pero que guarden una alta correlación con la variable de clasificación. Esta correlación puede verse estimada con la dependencia estadística entre las variables. La información mutua puede ser usada para cuantificar dicha dependencia, pues mRMR puede verse como una aproximación de maximizar la dependencia entre las distribuciones conjuntas de las características seleccionadas y de la variable de clasificación [ALC10].

Para características continuas, el *F-statistic* puede ser usado para calcular la correlación de la clase (relevancia) y el coeficiente de correlación de Pearson puede ser usado para calcular la correlación entre las características (redundancia). Luego, las características pueden ser seleccionadas una por una mediante una búsqueda *greedy* que optimice la función objetivo. Dos tipos de funciones objetivo comúnmente usados son **MID (criterio de Mutual Information Difference)**, que representa la diferencia de la relevancia entre las características, y **MIQ (criterio de Mutual Information Quotient)** que representa el cociente de la redundancia entre las características [RGFO 9].

4.3.1. Extensiones

Seleccionar variables puede no ser una tarea trivial: por ejemplo en el estudio sobre expresiones genéticas se tiene el carácter temporal de los datos. Sin embargo, la mayoría de los enfoques de selección de características desarrollados para datos de microarrays no pueden manejar datos multivariantes temporales sin un aplanamiento previo de datos, lo que resulta en una pérdida de información temporal. Extendiendo mRMR, se propuso un enfoque temporal de mínima redundancia máxima relevancia (TMRMR) que tuvo buenos resultados [RGFO 9].

Parte V.

Estimadores de información mutua

5. Información mutua para dos variables

Para simplificación del modelo, de ahora en adelante vamos a tomar la definición de la información mutua (y por tanto de la entropía de Shannon) tomando p como la probabilidad uniforme, donde $p(x) = P[X = x]$, $p(x, y) = P[X = x, Y = y]$. Por tanto, sea X una variable aleatoria discreta que toma n valores, la probabilidad $p(x)$ donde $x \in X$ es

$$p(x) = \frac{K(x)}{N}$$

donde $K(x)$ es la aplicación que determina cuántas veces aparece x en X. Se podría ver como la intersección entre X y una variable que toma el elemento x n veces, o se podría ver como una operación lógica 'AND' entre el vector X y otro vector con el elemento x n veces.

6. Optimizaciones y propuestas

Se ha basado y optimizado las funciones disponibles en [MM] de 'single_entropy' (entropía de una variable), 'entropy' (entropía entre dos variables) y 'mutual_information' (información mutua entre dos variables).

6.1. Entropía para una variable

Como se ha explicado con anterioridad, al trabajar con la probabilidad uniforme debemos contar cuántas veces aparece x en la variable X. La optimización en este caso viene por parte de la misma estructura de datos. Se ha transformado X en un diccionario cuyas claves son los distintos elementos que tiene X, y cuyos valores son las veces que aparece cada clave en X. Por tanto, $p(x)$ son los valores de la clave x dividido por el número de elementos de X.

6.2. Entropía para dos variables

La entropía para dos variables se puede ver de la siguiente manera, tal y como se visualiza en [MM]: $\forall x \in X, \forall y \in Y, p(x,y)$ es el número de elementos que hay en la intersección (entendiéndose como operación lógica 'AND') entre dividido por el número de elementos que contienen las variables X e Y.

A nivel de implementación, la traducción de lo explicado acaba en dos bucles anidados que recorren X e Y, respectivamente. La propuesta de optimización en este caso se basa en la transformación de dos bucles anidados que recorren X e Y, a dos bucles anidados que recorren X y otro que como máximo recorre las posiciones donde X es x de forma factorial:

$\forall x \in X$, obtenemos los lugares donde aparece x en X. Ahora, si llamamos P_x a la variable que nos indica dichos lugares, $\forall p \in P_x$, mientras que P_x no esté vacío, seleccionamos la y en la posición p de Y, borramos la posición p de P_x y llevamos el contador a 1, y ahora recorremos el resto de posiciones: si se repite la y de la nueva posición con la y de la posición p, aumentamos el contador en 1 y borramos la posición de P_x, y sino seguimos recorriendo. Entonces aplicamos la fórmula de la entropía donde $p(x,y)$ es el contador dividido por el número de elementos que tiene X e Y (tras esto volvemos a la parte de 'mientras que P_x no esté vacío').

En Python, el código queda de la siguiente manera:

```
for x in set(X):
    posiciones_x = where(X[:,0]==x)[0].tolist()
    while len(posiciones_x) > 0:
        y = X[:,1][posiciones_x[0]]
        cont_y = 1
        del posiciones_x[0]
        # We get the first y_i, count it (uniform probability)
        # and we remove the position from posiciones_x.
        y = X[:,1][posiciones_x[0]]
        cont_y = 1
        del posiciones_x[0]
        # We iterate in posiciones_x in a specific way: if we remove a
        # position then we don't update i, else i += 1.
        # Note that len(posiciones_x) is being updated inside the loop.
        i = 0
        while i < len(posiciones_x):
            # We find y_j.
            z = X[:,1][posiciones_x[i]]
            # If y_i == y_j, then we count it and remove.
            if y == z:
                cont_y += 1
                del posiciones_x[i]
            # Else, we iterate to the next element.
            else:
                i += 1
        # p(x,y) with uniform probability.
        pxy = cont_y / n_rows
        if pxy > 0.0:
            summation += pxy * math.log(pxy, log_base)
if summation == 0.0:
    return summation
else:
    return - summation
```

En un caso con dos variables arbitrarias, esta propuesta sería con mucha probabilidad más ineficiente que la opción dada, ya que pasamos de un algoritmo cuadrático $O(n^2)$ a uno $O(n(m!))$, con $m < n$, y generalmente $n(m!) > n^2$ (para $n = 4$ y $m = 3$ ya se da). Sin embargo, no trabajamos con dos variables arbitrarias: cuando apliquemos la entropía entre dos variables, lo haremos sobre una variable X y la variable de clasificación Y, que usualmente sus elementos diferentes son pocos comparado al abanico de posibilidades de X. Es por tanto que dejamos implementado y propuesto este algoritmo.

6.3. Información mutua entre dos variables

Esta optimización es justo la unión de la primera optimización y la propuesta en el apartado anterior, ya que están basadas en $p(x)$ y $p(x,y)$, y ambas aparecen en la definición de información mutua.

7. Información mutua en un conjunto de variables

Con la información mutua entre dos variables definida, es de gran utilidad preguntarnos cuál es la información mutua en un conjunto de variables. Es decir, sea $X = \{X_1, ..., X_N\}$, con cada X_i siendo una variable que puede tomar n valores, y sea Y la variable de clasificación (sin pérdida de generalidad podemos agrupar X e Y y llamar X_N a Y), ¿qué información mutua hay entre X e Y?

Vamos a indicar las opciones que hemos valorado a lo largo del proyecto, y la respuesta final.

7.1. La suma

La primera opción que tomamos fue la suma:

$$I(X_1, X_2, ..., X_N) := \sum_{i=1}^{N} I(X_i, X_N)$$

Sin embargo, es un mal planteamiento: por definición, ni siquiera podemos volver a escribir la información mutua en base a entropías, y por la idea que vendrá más adelante de la selección de variables. Sobre esto último, más adelante seleccionaremos las variables que del conjunto de variables dé una información mutua mayor, pues serían las variables que más nos aporta información sobre Y sin conocerla. Imaginemos ahora un caso en el que Y dependa de dos variables de X, pero X está compuesto de 5 variables. Esta mala extensión de la información mutua multivariante nos llevaría a seleccionar un subconjunto de X como las variables que más información nos dan sobre Y sin poder justificar que realmente sea así. En la práctica, hemos hecho esta misma comprobación y cuantas más variables habían, mayor era la información mutua, y seleccionaba todo el conjunto X.

7.2. La media

$$I(X_1, X_2, ..., X_N) := \frac{1}{N} \sum_{i=1}^{N} I(X_i, X_N)$$

Esta opción es digna de mención pues, aunque sea una mala extensión con mucho parecido a lo anterior, en realidad el segundo criterio de descarte es contrario. Es decir, volvemos a no poder escribir la información mutua en base a entropías, pero ahora la selección de variables cambia drásticamente: al realizar la media, ahora la variable de X que devolverá

mayor información mutua (como la que más explica la variable Y) sí tendrá relación con Y, pero el subconjunto que dará mayor información mutua siempre consistirá de una sola variable, independientemente de si está basado o no de más variables. Es decir, si por ejemplo Y está basado en las variables X_0 y X_1, y $I(X_0, Y) > I(X_1, Y)$, como Y está basado en X_0 y X_1 se tiene que $I(X_i, Y) > I(X_j, Y)$ con $i = 0, 1$ y $j = 2, .., N$; y como $I(X_0, Y) > I(X_1, Y)$, se tiene que

$$I(X_0, Y) = \frac{I(X_0, Y)}{1} = \frac{2 * I(X_0, Y)}{2} = \frac{I(X_0, Y) + I(X_0, Y)}{2} > \frac{I(X_0, Y) + I(X_1, Y)}{2}$$

Luego siempre tendríamos que la variable X_0, la variable que da mayor información mutua en solitario, sería la seleccionada.

7.3. La extensión teórica más convincente

Volvamos a ver cómo podemos escribir la información mutua en base a entropías. Como hemos demostrado sobre

Parte VI.

Cálculo del Test Delta

Parte VII.

Construcción de modelos a determinar el subconjunto de variables más adecuado

Parte VIII.

Propuesta de combinación de Test Delta con Información Mutua.

8. Propuesta de combinación de Test Delta con Información Mutua

Sabiendo las ventajas que nos aporta la información mutua, en concreto la selección de variables y los algoritmos basados en ella para sacar las características que reflejen mínima redundancia y máxima relevancia, así como las ventajas del rápido cálculo del Test Delta como propia técnica de selección de variables, surge la siguiente idea.

Combinar la información mutua con el test Delta.

A priori, una técnica de selección de variables que escoja un subconjunto que refleje las propiedades estadísticas de la variable de clasificación, que entre ellas sean tan diferentes entre sí como sea posible pero que marginalmente sean lo más similar posible, y que además las variables seleccionadas minimizan el ruido entre la entrada y la salida, nos hace pensar que puede obtener mejores resultados que la Información Mutua y el Test Delta por separado.

Parte IX.

Implementación.

Parte X.

Experimentos.

Parte XI.

Análisis.

Parte XII.

Conclusiones.

A. Primer apéndice

Los apéndices son opcionales.

Archivo: `apendices/apendice01.tex`

Glosario

La inclusión de un glosario es opcional.

Archivo: `glosario.tex`

$\mathbb{R}$ Conjunto de números reales.

$\mathbb{C}$ Conjunto de números complejos.

$\mathbb{Z}$ Conjunto de números enteros.

Bibliografía

Las referencias se listan por orden alfabético. Aquellas referencias con más de un autor están ordenadas de acuerdo con el primer autor.

[ALC10] B. Auffarth, M. Lopez, and J. Cerquides. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. advances in data mining. applications and theoretical aspects. http://www.csc.kth.se/~auffarth/publications/redrel.pdf, 2010. [Citado en pág. 17]

[CT] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Hoboken, New Jersey: Wiley. [Citado en pág. 7]

[ELL+] Emil Eirola, Elia Liitiinen, Amaury Lendasse, Francesco Corona, and Michel Verleysen. *Using the Delta Test for Variable Selection*. European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning Bruges (Belgium). [Citado en págs. 10 and 11]

[Eva] D. Evans. *Data-derived estimates of noise for unknown smooth models using nearneighbour asymptotics*. PhD Tesis, Cardiff University. [Citado en pág. 12]

[Hug59] G.F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968, doi:10.1109/TIT.2005.844059. [Citado en pág. 16]

[Jon04] A. J. Jones. New tools in non-linear modeling and prediction. *Computational Management Science*, 1(2):109–149, 2004. [Citado en pág. 10]

[KL68] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951, doi:10.1214/aoms/1177729694, MR 0039968. [Citado en pág. 7]

[KSG] A. Kraskov, H. Stgbauer, and P. Grassberger. *Estimating mutual information*. Phys. Rev. E, 69:066138. [Citado en pág. 12]

[KWW] S. E. Kemp, I. D. Wilson, and J. A. Ware. A tutorial on the gamma test. *School Of Computing*, 6(1 and 2):67–75. [Citado en pág. 10]

[Mac] D.J.C. Mackay. *Information theory, inferences, and learning algorithms*. [Citado en pág. 9]

[McM08] David M. McMahon. *Quantum Computing Explained*. Hoboken, New Jersey: Wiley-Interscience, 2008. [Citado en pág. 6]

[MM] Roberto Maestre and Bojan Mihaljevic. Script to calculate mutual information between two random variables. https://github.com/rmaestre/Mutual-Information/blob/master/it_tool.py. [Citado en pág. 21]

[PP94] H. Pi and C. Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Computation*, 6(3):509–520, 1994. [Citado en págs. 11 and 13]

[RGFO 9] Milos Radovic, Mohamed Ghalwash, Nenad Filipovic, and Zoran Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(9), 2017, doi:10.1186/s12859-016-1423-9. [Citado en pág. 17]

[RHJL] Nima Reyhani, Jin Hao, Yongnan Ji, and Amaury Lendasse. *Mutual Information and Gamma Test for Input Selection*. Helsinki University of Technology - Neural Network Research Center. [Citado en pág. 16]

[Sha8x] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948, doi:10.1002/j.1538-7305.1948.tb01338.x. [Citado en pág. 7]