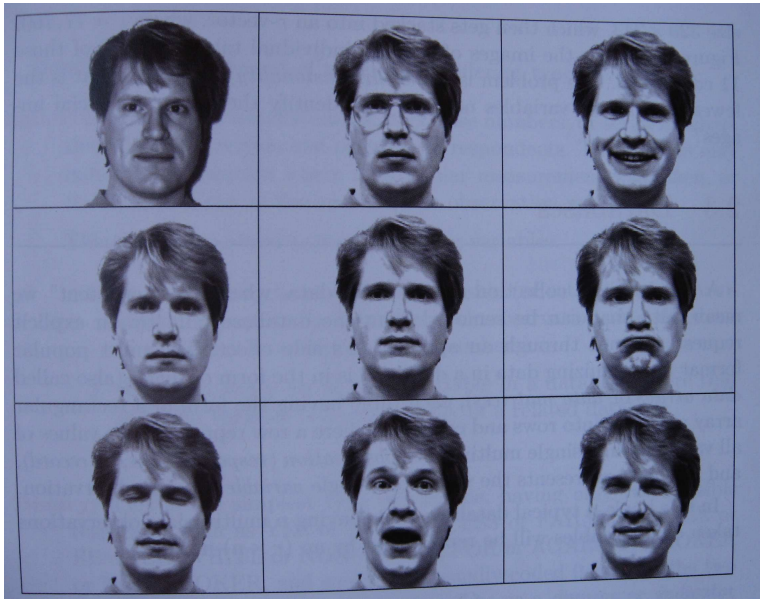
Reconocimiento Estadístico de Patrones, Parte 3

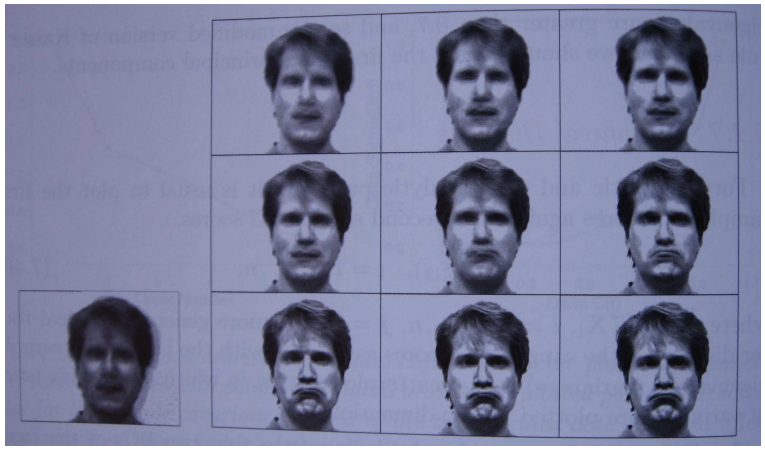
Johan Van Horebeek

horebeek@cimat.mx

Enero - Junio 2011

6.5 Comentarios finales





Existen muchos variantes; por ejemplo:

- nonnegative matrix factorization: coeficientes deben ser positivos
- local linear embedding
- local MDS

$$\sum_i \sum_{j \text{ vecino de } i} (d(x_i, x_j)^2 - d(x_i^*, x_j^*)^2)^2.$$

7. Visualizar, explorar y medir dependencias

7. Visualizar, explorar y medir dependencias

7.1 Visualización de la dependencia: regresión no paramétrica (caso continuo)

Nos limitamos aquí a una introducción más bien intuitiva.

Para un kernel simétrico K y los datos $\{(x_i, y_i)\}$, define la función que mapea x a:

$$\frac{\sum_i y_i K_h(x - x_i)}{\sum_i K_h(x - x_i)}. \tag{1}$$

Motivación de la forma (1):

Recuérdense que $E(Y - g(X))^2$ es mínimo para $g(x) = E(Y|X = x)$;

o sea: $E(Y|X = x)$ es desde cierto punto de vista (criterio) el *mejor* predictor para Y.

Por definición:

$$E(Y|X = x) = \frac{\int y f_{X,Y}(x, y)dy}{f_X(x)} = \frac{\int y f_{X,Y}(x, y)dy}{\int f_{X,Y}(x, y)dy}.$$

Para estimar $f_{X,Y}(x, y)$ usamos un estimador basado en un kernel bidimensional formado por el producto de dos kerneles unidimensionales.

Tarea: deriva que bajo estos supuestos se obtiene efectivamente (1)

7. Visualizar, explorar y medir dependencias

7.1 Visualización de la dependencia: regresión no paramétrica (caso continuo)

Nos limitamos aquí a una introducción más bien intuitiva.

Para un kernel simétrico K y los datos $\{x_i, y_i\}$, define la función que mapea x a:

$$\frac{\sum_i y_i K_h(x - x_i)}{\sum_i K_h(x - x_i)}.$$

Otra motivación de la forma (1):

Recuérdense que en regresión lineal (mínimos cuadrados) minimizamos $\sum_i (y_i - a - bx_i)^2$.

Un caso especial es minimizar $\sum_i (y_i - a)^2$

Fijamos x; para tomar en cuenta las distancias entre x y x_i minimizamos:

$$\sum_i (y_i - a)^2 K_h(x - x_i), \tag{2}$$

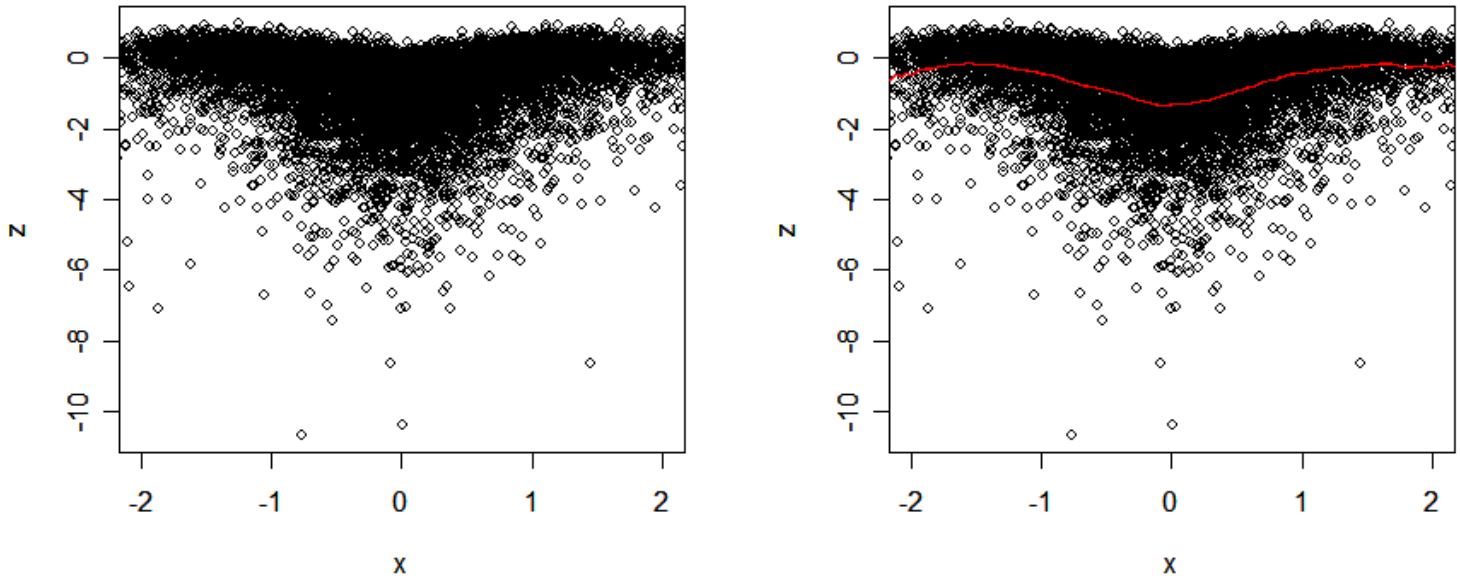
con K un kernel simétrico como antes.

Derivar (3) con respeto a a, resulta en:

$$2\sum_i (y_i - a) K_h(x - x_i) = 0 \quad \Leftrightarrow \quad a = \frac{\sum_i y_i K_h(x - x_i)}{\sum_i K_h(x - x_i)}$$

Observa: !a depende de x!

Ejemplo



Insertar demo `kernelregresion.r`

```
points(ksmooth(x,z,bandwidth=0.3),type="l",col="red")
```

Otra estimador es lowess (loess) donde minimizamos:

$$\sum_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_p x_i^p)^2 K_{h(x)}(x - x_i), \tag{3}$$

`loess()` o `plsmo()`.

7.2 Visualización de la dependencia: Mosaicplot (datos nominales)

Idea: compara contra lo que uno obtendría bajo independencia.

Tomamos primero caso 2D: (X, Y).

Define $n(x, y)$ el número de observaciones con $X = x, Y = y$ y

$P(x, y)$ la probabilidad basada en la frecuencia relativa: $n(x, y)/n$.

IDEA

(1) Con cada (x, y) asociamos un rectángulo;

un lado es proporcional a $P(X = x)$ y otro lado es proporcional a $P(Y = x|X = x)$

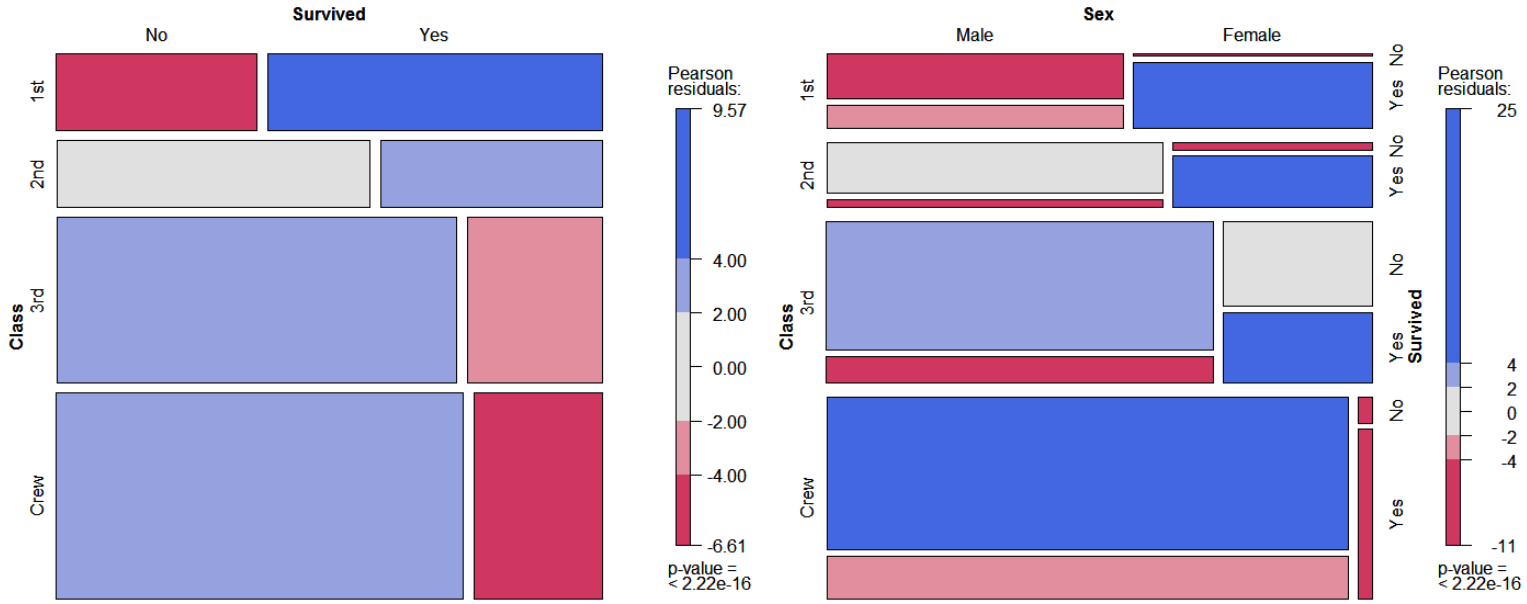
(o a $P(Y = y)$ y $P(X = x|Y = y)$ respectivamente).

(2) Coloreamos cada rectángulo en base de la diferencia normalizada entre $n(x, y)$ y el número de observaciones esperadas bajo independencia, $e(x, y)$,

$$\frac{(n(x, y) - e(x, y))^2}{e(x, y)}$$

¿Cómo calcular $e(x, y)$?

Para dimensiones mayores se subdivide cada rectángulo alternando con cortes horizontales y verticales.

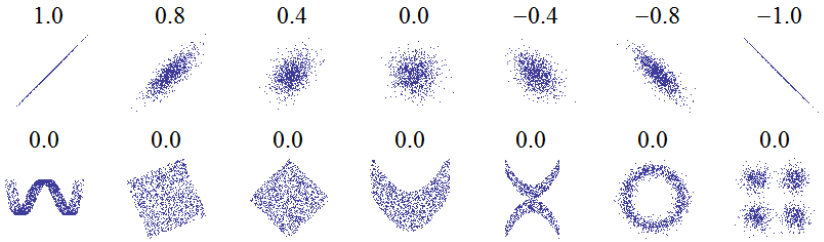



```
library(vcd); data(Titanic)
mosaic(~ Class + Survived, data = Titanic, shade = TRUE, legend =TRUE)
mosaic( ~ Class + Sex+Survived, data = Titanic, shade = TRUE, legend = TRUE)
```

7.3 Medir dependencia

A. Medidas de dependencias (suponiendo un orden):

(1) correlación (ev. aplicar primero transformación monótona)



(2) Spearman ρ: es la correlación basada en los rangos de los datos

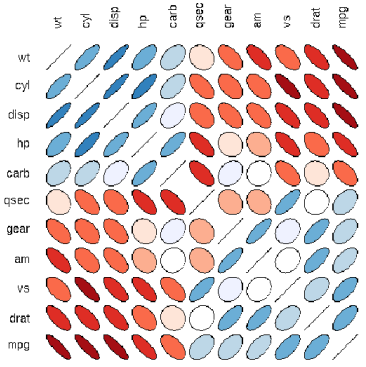
(3) Kendall τ (medida para datos ordinales)

B. Medidas de dependencias (datos nominales):

(1) medidas de reducción de incertidumbre

(2) power divergence statistics

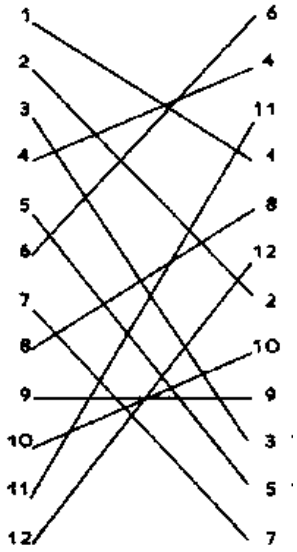
Aplicación: en lugar de observaciones visualizamos alguna de estas medidas de dependencia.



`library(ellipse);plotcorr(···)`

A. Medidas de dependencias (suponiendo un orden):

Motivación Kendall τ Para cada dato (x,y), calcula el rango de x y de y; conéctalos con una linea:



Sea c el número de cruces; si no hay empates:

$$\tau = 1 - \frac{2c}{n(n-1)/2}.$$

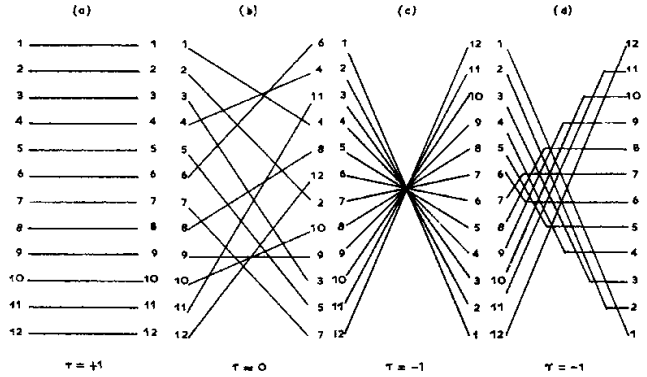
En general $\tau = \Pi_C - \Pi_D$ donde

Π_C es la probabilidad de que dos observaciones (x^1, y^1) y (x^2, y^2) están en concordancia:

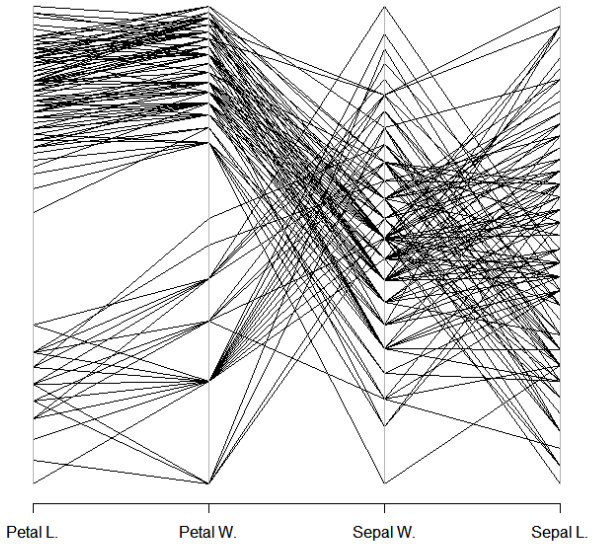
$$x^1 < x^2 \text{ y } y^1 < y^2, \text{ o } x^1 > x^2 \text{ y } y^1 > y^2$$

Π_D es la probabilidad de que dos observaciones (x^1, y^1) y (x^2, y^2) están en discordancia:

$$x^1 < x^2 \text{ y } y^1 > y^2, \text{ o } x^1 > x^2 \text{ y } y^1 < y^2$$



Uso en parallel coordinate plot:



Verifica que $\tau = 0$ si hay independencia.

TABLE 2.8 Cross-Classification of Job Satisfaction by Income

Income (dollars)	Job Satisfaction: Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
< 15,000	1	3	10	6
15,000–25,000	2	3	10	7
25,000–40,000	1	6	14	12
> 40,000	0	1	9	11

Source: 1996 General Social Survey, National Opinion Research Center.

Calcula número de pares en concordancia:

$$1 * (3 + 10 + 7 + ...) + 3 * (10 + 7 + ...) + ... + 14 * 11 = 1331$$

y en discordancia:

$$3 * (2 + 1 + 0) + ... + 12 * (0 + 1 + 9) = 849.$$

¿La diferencia positiva es "significativa"? $\rightarrow$ requerimos pruebas de hipótesis (basados en ciertos supuestos)

B. Medidas de dependencias (datos nominales):

(1) Medidas de reducción de incertidumbre

Define $V(X)$ medida de variabilidad (entropia de Gini, de Shannon, etc) y calcula:

$$\frac{V(X) - E_Y(V(X|Y))}{V(X)}.$$

Elección popular en Computación: Información mutua:

$H(X)+H(X)-H(X,Y) = H(X)-E(H(X|Y)) = H(Y)-E(H(Y|X))$, con H entropía de Shanon

(2) Power divergence statistics

Se mide la diferencia de lo que se observa con lo que uno espera bajo independencia.

Definir "distancia" entre distribuciones:

$$I^{\lambda}(P, P_0) = \frac{1}{\lambda(\lambda+1)} \sum_i P_i[(\frac{P_i}{P_{0,i}})^{\lambda} - 1] \quad \text{power divergence statistics}$$

Casos especiales: $\lambda = 1(Pearsontest), -0.5(Neymantest), 0(KullbackLeibler)$.

7.4 Reglas de asociación

Supongamos que $(X_1, \cdots, X_d)$ son variables binarias.

Para $A \subset \{1, \cdots, d\}$, denota $\prod_{i \in A} X_i$ como X_A.

Queremos buscar conjuntos A, B, tal que $X_A = 1 \Rightarrow X_B = 1$ con "alta probabilidad" (!por definir bien!).

- Buscamos todos los conjuntos C (=frequent item sets) tal que

$$P(X_C = 1) > \epsilon_1 \ \ (=\text{soporte}).$$

- Buscamos particiones de C en A y B tal que

$$P(X_B = 1 | X_A = 1) > \epsilon_2 \ \ (=\text{confianza/predicibilidad}).$$

y

$$\frac{P(X_B = 1, X_A = 1)}{P(X_B = 1)P(X_A = 1)} > \epsilon_3 \ \ (=\text{elevación}).$$

Si A, B cumplen con lo anterior, decimos que $A \Rightarrow B$ es una regla de asociación.

Se pueden obtener de manera eficiente los frequent item sets (insertar derivación algoritmo apriori).

Problema: encontrar estructuras en la inmensa cantidad de reglas de asociación.

Para considerar relaciones "negativas" de X_i, hay que anadir $(1 - X_i)$.

Hay que discretizar variables continuas. ?Cómo?

```
library(arules) ;  data("Adult"); rules <- apriori(Adult, parameter=

list(confidence = 0.5,supp=0.2),

      appearance = list(rhs = c("income=small",

      "income=large"),
      default="lhs")); inspect(rules)
```