UN MUNDO DE DATOS, UN MUNDO DE NÚMEROS

ANTOM LABRANHA
IES San Clemente
Departamento de Didácticas Aplicadas - USC

A computerización do mundo, xa o sabemos, está a trocar todas as nosas actividades e todas as nosa relacións en datos numéricos. É o *Big data*. Para algúns, un cambio radical no futuro da vida humana. Para os máis, un enigma case indescifrable. Aparentemente, só son números, só é ciencia e técnica. Coma desde Pitágoras. Mais é desexable encerrar a nosa vida en moreas de datos? Queremos que un algoritmo nos diga de quen nos debemos namorar? A través de dous exemplos sinxelos, intentarei recrear intelixencia artificial e minería de datos, amosando, grosso modo, como os procesos construtivos cos que se afronta o *Big data* involucran, non só competencias disciplinares e esforzos, máis tamén sensibilidades e emocións.

TOMANDO DECISIÓNS NA INCERTEZA

Hai dous conceptos, inversamente interrelacionados, que a Estatística manexa explicitamente e que as persoas conxugamos a cotío, mesmo non o fagamos conscientemente. Refírome a *precisión* e *confianza*. Por exemplo, cando lle botamos a idade a unha persoa: se dicimos que ten entre 50 e 70 anos, fácil que acertemos, pero iso non é moito dicir; porén, se dixésemos que ten entre 58 e 62 anos, seríamos bastante precisos, pero o risco de fallar tamén sería grande: gañamos precisión perdendo confianza, e viceversa.

En xeral tendemos a colocarnos nun valor puntual e oscilar nun entorno del: "debe andar polos 60, máis ou menos", adoitaremos dicir. O dito entorno concretámolo, por exemplo en ± 2, que será a marxe de erro admitida (ou [50, 60] o intervalo de aceptación).

En situacións tan comúns coma esa, semella que establecemos intuitivamente a *precisión da estimación que realizamos* e a *confianza na súa validez* (que son dous parámetros baixo os que se realizan os estudos mostrais). Terreo esvaradío ...o de apelar á "intuición". Hai uns anos, para traballarmos estes conceptos co alumnado da ESO, realizamos unha actividade introdutoria á que, buscando motivalo, démoslle xeito de campionato de "canastra múltiple":

Nunha caixa fanse furados circulares de tamaños sensiblemente diferentes. Cada participante dispón dunhas bólas (utilizamos 10) que intentará "encestar" dende a liña de lanzamento, elixindo libremente cantas bólas lanza a cada canastra, obtendo en caso de éxito 2 e 1 puntos, respectivamente na pequena ou na grande.

Para acertar no pequeno debemos ser máis precisos, hai menos marxe. Consegues maior puntuación, o que o fai apetecible o intento, pero é menor a probabilidade de acertar. No grande temos máis marxe, logo aumenta a probabilidade de acertar, é máis fácil e, en boa lóxica, asignamos menor puntuación.

Dispoño dun total de 10 bólas, que me convén facer?

Primeiro adestran e logo comezan as eliminatorias. Anoto para cada xogador/a como pensa inicialmente distribuír os seus lanzamentos. Observo que o fan segundo a habilidade mostrada na fase preparatoria, ou sexa, cada quen concédese unha marxe en función da propia confianza no éxito. Mais élles permitido cambiar sobre a marcha e, en xeral, ao longo dos intentos varían o plan inicial, segundo os puntos que vaian conseguindo.

Todo sucede no patio, sen cadernos nin calculadoras, dialogando. Cada quen improvisaba unha análise sinxela dos seus resultados, o que supón algún tipo de avaliación rudimentaria das respectivas probabilidades.Tampouco foran "convidados" a facer cálculos nin manexaban, aínda, o concepto matemático subxacente de esperanza matemática.

Podería unha máquina imitar esta maneira de proceder e tomar decisións en función de como lle vaia indo? Podería. Imos velo. Antes unhas palabriñas de ánimo:

O cálculo de probabilidades é o único modelo matemático a disposición de quen pretende entender o descoñecido e o incontrolable. Afortunadamente, este modelo é á vez moi potente e moi cómodo.
Mandelbrot (un dos creadores da xeometría fractal)

Os valores asignados para as probabilidades serían as frecuencias relativas: $P = \dfrac{n^{o}\ acertos}{n^{o}\ \text{int}\,entos}$.

Falaríamos de *p* e *g* para referirnos ás canastras pequena e grande, respectivamente.

Se lanzamos *x* veces á pequena (acertaríamos, por termo medio, *x·p,* a 2 puntos por éxito) e quedarían *10-x* lanzamentos á grande (acertaríamos, por termo medio, *(10-x)·g*, a 1 punto por acerto).

A esperanza matemática é unha estimación da puntuación que conseguiríamos:

$$E(x) = x\cdot p\cdot 2 + (10 - x)\cdot g\cdot 1$$

Simplificando[1] esa expresión: $E(x) = (2p - g)x + 10g$ (*), recoñécese que responde á ecuación dunha recta (recordemos aquilo de *y=mx+b*) na cal a inclinación vén indicada polo coeficiente da variable *x,* neste caso *2p–g*. Daquela:

i) i) Se $p > \frac{1}{2}g$ a pendente é positiva: a recta ascende ao aumentar *x*, logo a mellor decisión sería lanzar sempre á canastra pequena.

[1] A simplificación é un recurso técnico que, non alterando o contido matemático, permítenos comprender mellor as relacións obtidas e, se cadra, tirar máis doadamente novas conclusións. Vén sendo un tipo de resposta ao que a sabedoría popular refire como que "as árbores non nos deixan ver o bosque".

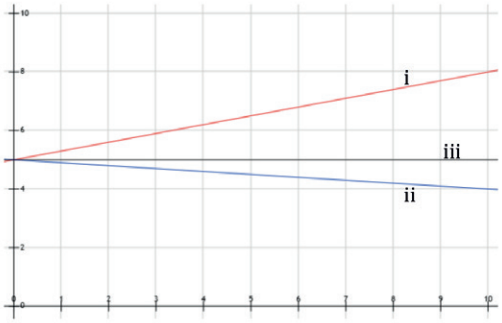
ii) ii) Se $p < \frac{1}{2}g$ a recta é descendente, diminúe ao aumentar *x*, logo convirían tirar sempre á grande.

iii) Se $p = \frac{1}{2}g$ anúlase o primeiro termo e queda *E*=10g, sexa como for como o reparto.

De igual xeito a como facían as/os estudantes, aínda que con máis tecnicismos, o robot recalcularía en cada lanzamento a correspondente probabilidade, reformulando a expresión (*) e, de se ver modificado o seu "estado" [i), ii) ou iii)], cambiaría a decisión.

Exemplo para $p = 0'4$ *e* $g = 0'5$ (i);

$p = 0'2$ *e* $g = 0'5$ (ii) ; $p = 0'25$ *e* $g = 0'5$ (iii)



Os valores iniciais de *p* e *q* poderíamos asignalos razoablemente, pero non é sequera necesario. Pode a propia máquina escollelos aleatoriamente (entre 0 e 1) pois, segundo fose aumentando a cantidade de datos, as frecuencias relativas tenderían cara uns valores abondo estables. A constatación de que iso é o que adoita acontecer en experiencias do máis diverso, levou ao matemático ucraíno Von Mises (XIX-XX) a definir a probabilidade como o límite das frecuencias relativas. Isto pertence ao ámbito do que demos en chamar "regularidade estatística" ou "lei dos grandes números".

Resultaranos fácil entender que aumentando o número de opcións (canastras e puntuacións adxudicables, no exemplo) a diversidade de "estados" posibles aumenta e, consecuentemente, o número de operacións necesarias para tomar a decisión conveniente, multiplícase. Tamén podemos alterar o número de ensaios (bólas a lanzar, no exemplo). A variabilidade medraría exponencialmente[2], demandando cada vez un maior número de datos, que axiña chegaría a ser enorme: o *Big data* agroma.

Porén, o valor da matemática non radica unicamente na posibilidade de resolver estes ou aqueles problemas complexos, senón tamén en que nos procesos de achádego de tales solucións emerxen modelos de resolución adaptables a grandes "familias" de situacións que inclúen fenómenos que inicialmente nin sequera relacionabamos entre si.

O caso da compra-venta de accións en bolsa, ben individual, ben a través dunha sociedade organizada para tal fin (fondos de investimento) ten unha natureza semellante ao xogo das canastras. Algo parecido sucedería coa distribución do gasto dun capítulo orzamentario, considerando o "retorno" en actividade produtiva derivada como unha virtual "puntuación conseguida". Tamén a experimentación coas proporcións moleculares dun fármaco composto, ou coa combinación pautada dun "cóctel" de fármacos preelaborados son susceptibles dun estudo desta natureza. E un longo etcétera.

MACÍAS O NAMORADO

"Moi logo, os datos masivos serán capaces de dicirnos se nos estamos namorando" (*Big data: a revolución dos datos masivos.* Mayer, V. e Cukier, K.*).* Gulosa predicción. Aplicaríamos para tal fin algún dos potentes algoritmos de análise de grupos (segmentación), que tratan de reunir

[2] Os termos "multiplícase" e "exponencialmente" úsanse aquí e no que segue en sentido coloquial, enfático.

aqueles elementos que nalgún sentido poidamos considerar semellantes (clúster), para o cal se aplica a idea común de que o serán tanto máis canto maior sexa a súa proximidade.

Moitos algoritmos seguen criterios xerárquicos de conectividade, dando lugar a *dendogramas* (representación mediante unha estrutura de árbore, en categorías e subcategorías sucesivas): o nº de pólas que conectan un dato con outro é o indicador de semellanza.

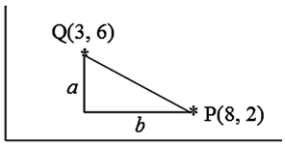
Noutra perspectiva, os algoritmos tipo *k-means* buscan clasificar os datos en *k* grupos, cada un ao redor dun dato medio (centroide) no que estarían aqueles datos que son máis próximos a el ca aos outros centroides. Teríamos, entón, que decidir que é o que imos analizar (as mensaxes de texto parecen axeitadas para namoramentos) e definir o que entendemos por "distancia".

Do proemio do Marqués de Santillana ao condestable Pedro de Portugal (1429-1466), conde de Barcelona, que fora recoñecido como rei de Aragón polo principado de Catalunya na loita contra o rei trastámara Juan II de Aragón, facemos un extracto que achega significativa noticia:

"Después de ellos vinieron Vasco Peres de Camões y Fernand Casquiço, y aquel grande enamorado Macías, del cual no se hallan sino cuatro canciones, pero ciertamente amorosas y de muy hermosas sentencias, conviene a saber: *Cativo da miña tristura, Amor cruel e brioso, Señora, en quien fiança y Provei de buscar mesura.*"	 Estatua do trobeiro Macías, en Padrón	Cativo da miña tristura, ya todos prenden espanto e preguntan, que ventura fay que me tormenta tanto; mays non sey no mundo, amigo, que mays de meu quebranto diga de esto que vos digo. (Cancioneiro de Baena)

No ámbito da teoría da información manéxase o concepto de entropía como "cantidade de información relevante" contida nunha mensaxe. Por simplificar, pensemos só nos 140 carácteres do Twitter, que diariamente almacena ducias de terabytes (1 tera =1000^4 bytes; máis de 1 billón de unidades de información): o *Big data* ameázanos.

Para isto do amor teríamos un conxunto de palabras e cadeas de alta relevancia (como as que abundan na *cantiga de amigo* do Macías) e outras de baixa cualificación[3]. Facemos unha avaliación, poñámoslle semanal, da entropía media a cada usuario obxecto da nosa observación. Esa sería a primeira coordenada, x, da persoa "seguida". A segunda, y, podería ser, por exemplo, a media da duración do intervalo de tempo que pasa entre chío e chío nas horas de actividade (estou pensando en certa ansiedade por "ligar", no sentido xenuíno da palabra, que arrastra cando a si o sentido coloquial).



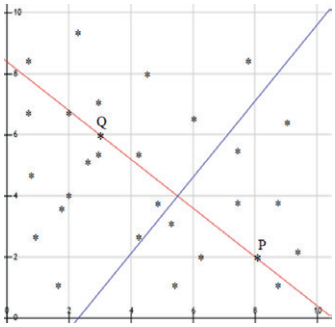
Cada persoa obxecto do seguimento aparecería representada como un punto e o conxunto formaría unha "nube". Por adoptar unha escala habitual, asignemos valores entre 0 e 10 para ambas as dúas variables (coordenadas). Quizais nos pareza adecuado o punto P(8, 2) para un prototipo de persoa namoradeira -sen esaxerar-, e o Q(3, 6) para quen non está especialmente ocupada nestas necesidades. O paso se-

[3] A codificación realízase cun dicionario de análise de contido (software *textpack*). Unha vez que teñamos perfiladas as unidades a rexistrar (substantivos, verbos, adxectivos...), cada vez que o dicionario encontre unha desas palabras codifícaa e outro programa específico asígnaselle o valor que teñamos decidido. O procedemento compleméntase cun programa de *desambiguación* (*key word in context*) que pon en relación os termos menos claros coas unidades do contexto no cal aparecen.

guinte é o de definir a distancia entre puntos, cuestión que adoitamos facer seguindo un camiño escolar típico: a distancia euclídea, que vén sendo a hipotenusa do triángulo rectángulo que se formaría. Lembrando a Pitágoras: $h^2 = a^2 + b^2$:

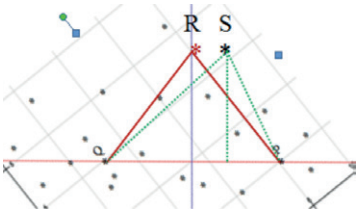
$$(d_{PQ})^2 = (6-2)^2 + (8-3)^2 = 4^2 + 5^2 = 41 \rightarrow d_{PQ} = \sqrt{41} = 6'40$$

Agora facemos dous grupos, por proximidade a P ou a Q, respectivamente, cun procedemento tamén escolar: a mediatriz do segmento que une P con Q (recta que pasa por dous puntos e perpendicular polo punto medio).



Os puntos da mediatriz equidistan de P e Q, que elixíramos como centroides; os da súa esquerda están máis próximos de Q e os da súa dereita, de P.

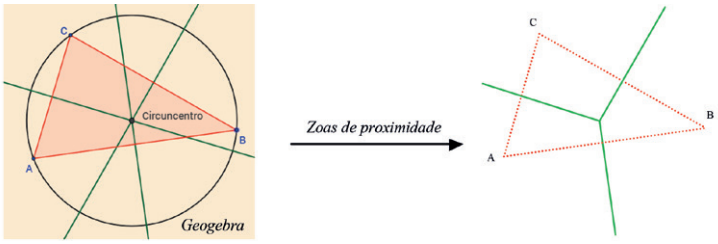
Isto resultará fácil de apreciar se xiramos a figura. Tomemos un punto R calquera da mediatriz: unindo con Q e P formamos dous triángulos rectángulos de iguais altura e base (xa que a mediatriz está trazada pola metade), logo son iguais as hipotenusas (distancias).



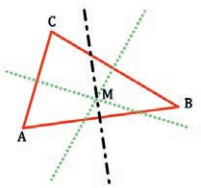
Collendo un punto calquera, por exemplo S, aumentamos a base do triángulo rectángulo que se formaría respecto de Q (maior hipotenusa, consecuentemente) e diminuímos a do que se formaría respecto de P (menor hipotenusa), logo está máis cerca de P ca de Q[4].

Se contemplásemos un terceiro estado da cuestión, xurdiría un novo centroide. Chamémoslles agora A, B e C; aplicaríamos o método a cada parella de puntos A e B, A e C, B e C, e formaríamos os tres grupos correspondentes. En xeral teremos *k* "estados", *k* puntos centroide. O número de operacións necesarias para formar os grupos convenientes, emparellándoos, multiplícase.

Merece a pena desenvolver o caso dos tres estados, pois o procedemento seguido condúcenos a un dos puntos notables do triángulo, que tamén se estuda na xeometría euclídea escolar: o **circuncentro** (centro da circunferencia circunscrita). Se os puntos de cada mediatriz equidistan dos extremos do respectivo segmento, o punto común a elas equidista dos tres extremos, logo é posible trazar unha circunferencia con centro este punto e radio a distancia común, que pasará por A, B e C.



[4] Ese mesmo esquema sérvenos para confirmar a proposición recíproca: se un punto equidista de Q e P estará na mediatriz do segmento do que estes son extremos (de non ser así estaríamos no caso do punto S e, trazando unha perpendicular ao segmento, formaríamos dous triángulos rectángulos de igual altura pero diferente base, polo que as hipotenusas -distancias- serían diferentes).



Abonda con falarmos do punto de corte de dúas calquera das tres mediatrices, que xa establecería a equidistancia aos tres vértices, polo que nova mediatriz tería que coincidir con elas no mesmo punto:

$d_{MA} = d_{MC} \quad e \quad d_{MC} = d_{MB} \quad \Rightarrow d_{MA} = d_{MB}$, logo M pertence á mediatriz do segmento AB.

Por outra banda, se o número de características ou condutas observadas aumenta, cada individuo terá unha "cualificación" ou coordenada para cada unha delas, polo que nos situaremos nun espazo de 3, 4, 5... en xeral, *n* dimensións[5]. Nesa tesitura non será fácil preestablecer bos centroides baseándonos nas nosas intuicións, polo cal adoita comezarse elixíndoos aleatoriamente e aplicar o método ata termos *k* rexións. Logo búscase o "centro xeométrico" dos puntos de cada unha desas rexións, que serán os centroides do seguinte paso: vólvese aplicar o método e configúranse outras rexións, que só coincidirán parcialmente coas anteriores –que xa se desbotan-. Reitérase o proceso ata que as rexións se estabilicen o suficiente (outra vez topamos con Von Mises). A variabilidade medra exponencialmente, procesando cada vez un maior número de datos, que axiña chegaría a ser inmenso: o *Big data* acósanos.

[5] Inténtase, nesa pluralidade, explicar as correlacións entre as variables observadas en función dun pequeno número doutras variables, non observadas directamente. É o que se denomina *análise factorial*. O procedemento responde á idea de "base" dun espazo vectorial que, en definitiva, é unha formalización dunha idea simplificadora moi común e humana: cando hai moita variabilidade, desexamos encontrar uns poucos elementos (*factores*) tales que, combinándoos axeitadamente, permitan explicar toda a diversidade que temos diante. Tecnicamente faise mediante combinacións lineares (coma nos sistemas de ecuacións ordinarios) admitindo unhas marxes de erro para cada individuo respecto da media do seu grupo.