Estadística avanzada: el problema del sobreajuste y el método de descripciones mínimas

Pablo Argibay

INTRODUCCIÓN: MODELOS EN ESTADÍSTICA

Existen varias y claras definiciones de lo que la estadística es. Sin embargo, para los fines de este trabajo, nos centraremos en el propósito de la estadística y consideraremos que este es extraer información útil de los datos observados. Nos quedamos así con dos conceptos de importancia a la hora de utilizar o valorar una técnica estadística en particular: el concepto de información relevante o útil para el operador y el de datos observados. En general, se acepta que dicha información útil viene en forma de "modelo", entendiendo que los modelos son códigos o métodos de descripción de los datos. Hacer inferencia estadística significa seleccionar un modelo que se ajuste a los datos. Por otra parte, la modelización estadística es el aprendizaje de reglas y restricciones en un conjunto de datos observados. Está claro que, en sentido amplio y moderno, los modelos pueden ser: distribuciones estadísticas, polinomios, cadenas de Markov, redes neurales y árboles de decisión, entre otros.

Como quiera que sea, es importante entender que los modelos no son la realidad, aunque se ajusten muy bien a ella. Y digo que es importante entenderlo ya que uno de los mayores problemas que en general tanto yo como otros hemos observado en investigación biomédica es la confusión del modelo (llamémoslo también teoría o hipótesis) con la realidad. Incluso en las ciencias más duras, la utilización del denominado método axiomático no conlleva el descubrimiento de verdades inmutables acerca de la realidad. El más perfecto aparato de diseño y medición no se encontrará en la vida real con que la suma de los ángulos interiores de un triángulo da 180°. La geometría euclidiana es maravillosa, pero solo un modelo de la realidad, útil por cierto. En estadística el tema es aún un poco más complejo: agrupamos las propiedades del mundo (datos) en distribuciones de probabilidad y, mediante heurísticas más o menos creíbles, hacemos inferencias acerca del mundo. Está más que claro que, si no tenemos bastante evidencia física del evento, el asumir distribuciones arbitrarias y hacer "deducciones" no conlleva nada útil, por más visos de "toma de decisiones racionales" que nos parezca que tengamos. En forma tradicional, cuando un fiel creyente de la estadística procede, asume los datos como provenientes de una población de una distribución probabilística, paramétricamente definida (esto último, lo de "paramétricamente definida" no es fundamental, ya que la problemática se extiende a los denominados métodos "no paramétricos"). El problema es que esto funciona si existe "ley" o regularidad, como ocurre en física cuántica. En nuestra realidad biomédica, en general, es bastante difícil discernir si la muestra proviene de alguna distribución particular en la población. El problema no es menor, ya que uno de los objetivos (bien grande) de la investigación científica es saber lo más que podamos acerca de la "maquinaria" que genera los datos observados, es decir, queremos saber la estructura del modelo.

Premisa fundamental de este trabajo: no podemos asumir a priori nada de la "maquinaria" que genera los datos. El objetivo sería intentar aprender los rasgos de regularidad en esos datos adjuntando a ellos modelos probabilísticos. Como antes hablamos de información útil, el objetivo en estos términos (información medida en bits[1]) sería: dado un conjunto de datos, averiguar los bits que:

- Tenemos de información aprendible
- Tenemos de complejidad
- Dejamos como ruido sin estructura aprendible con el modelo.

EL PROBLEMA DEL SOBREAJUSTE (*OVERFITTING*)

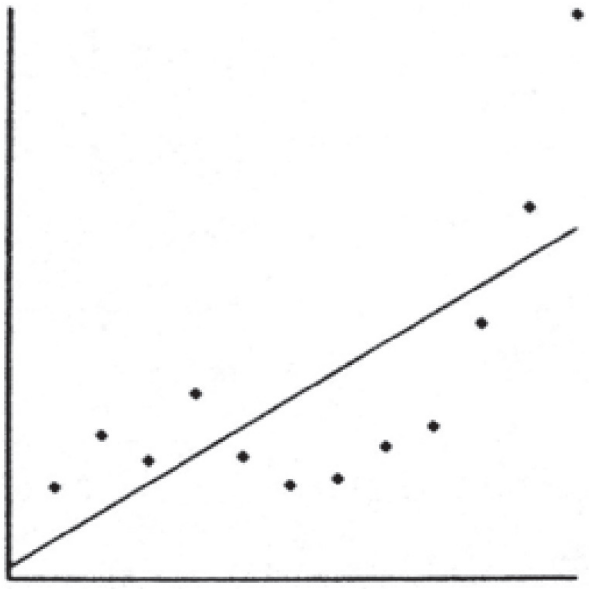
Consideremos una serie dispersa de puntos (y1, x1),…, (yn, xn) en un plano, a los cuales queremos ajustar un polinomio.[2] Buscamos ni más ni menos f(x). Tenemos varias opciones:

1. El bit (acrónimo de *binary digit* o dígito binario) es la unidad mínima de información empleada en informática, en cualquier dispositivo digital, o en la teoría de la información. Con él podemos representar dos valores cualesquiera, como verdadero o falso, abierto o cerrado, blanco o negro, norte o sur, masculino o femenino, rojo o azul, etc. Basta con asignar uno de esos valores al estado de "apagado" (0), y el otro al estado de "encendido" (http://es.wikipedia.org/wiki/Bit).
2. Un polinomio es una expresión algebraica compuesta de dos o más monomios. Tiene la forma: P(x) = an xn + an - 1 xn - 1 + an - 2 xn - 2 + ... + a1 x1 + a0. Siendo an, an -1 ... a1 , a0 números llamados coeficientes. El grado de un polinomio P(x) es el mayor exponente al que se encuentra elevada la variable x. Ejemplo de polinomio de grado 2: $P(x) = 2x^2 + 3x + 2$.

Laboratorio de Aprendizaje Biológico y Artificial (LABAL) – Instituto de Ciencias Básicas y Medicina Experimental (ICBME) – Hospital Italiano de Buenos Aires.
Correspondencia: pablo.argibay@hospitalitaliano.org.ar

1. Buscar un polinomio muy simple, como por ejemplo un modelo de regresión lineal (Fig.1).

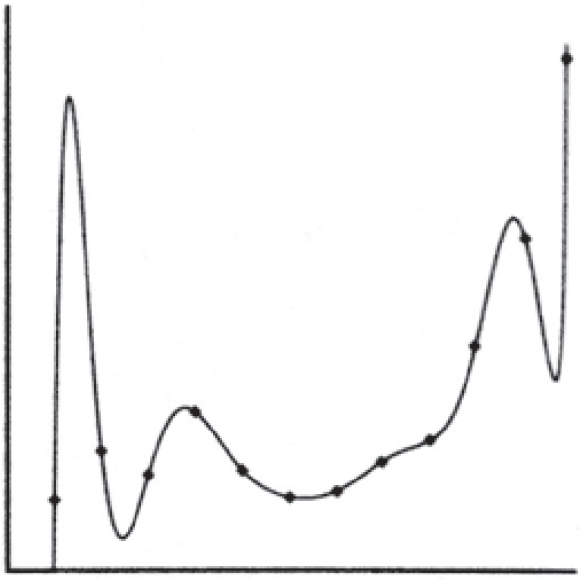
Figura 1. Ajuste a los puntos de un modelo de regresión lineal.



Es el más sencillo de los modelos y, sin embargo, no se ajusta bien, ni es factible que pueda predecir algo en el futuro.

2. Buscar un polinomio complejo de grado cercano a (n-1) (Fig. 2).

Figura 2. Ajuste a los puntos de un polinomio complejo de grado n-1.



Este modelo es más complejo y se ajusta a los datos; sin embargo, en algunos lugares se "escapa" y es poco factible que prediga algo en el futuro: está sobreajustado.

Se define la paradoja del sobreajuste como: los modelos complejos contienen más información sobre los datos de la muestra, pero menos información sobre los datos futuros (predicción).

El problema del sobreajuste en estadística no es menor y lleva a algunos problemas graves en investigación:

• Algunas relaciones que parecen estadísticamente significativas son solamente ruido.

• La complejidad del modelo estadístico es muy grande para la cantidad de datos que tenemos.

• El modelo en general no es replicable y predice mal.

Daremos un ejemplo.

Primero definiremos dos variedades de error en modelización estadística:

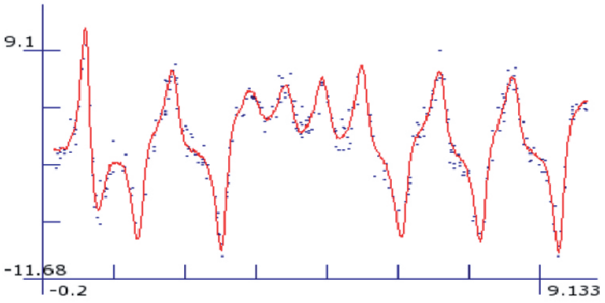
• Error de entrenamiento: aplicable a la muestra observada.

• Error de generalización: aplicable a futuras muestras.

Es obvio que nuestro fin como investigadores es minimizar el segundo.

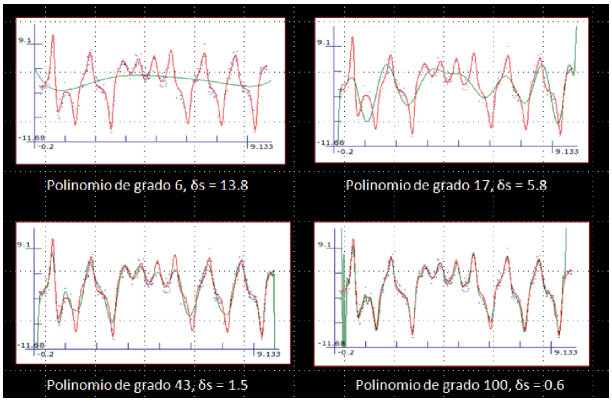
Tomemos el caso de una muestra de 300 puntos en el plano bidimensional (Fig. 3):

Figura 3. Conjunto de 300 puntos en un plano bidimensional. Este tipo de gráficos es más que común en investigación biomédica.



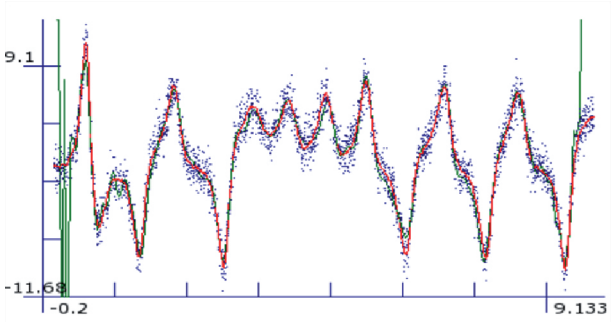
Veamos de ajustar diversos polinomios y tomemos como una medida relativa de error el desvío típico (Fig. 4):

Figura 4. Diferentes polinomios de grados 6, 17, 43 y 100 con sus correspondientes desvíos típicos.



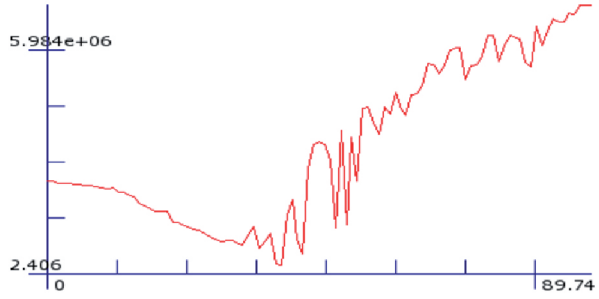
Es casi obvio que el bajo grado de error del polinomio de grado 100 amerita su utilización en las predicciones. Sin embargo, veamos qué ocurre cuando en realidad observamos con más discriminación y vemos 3000 puntos con el ajuste del polinomio de grado 100 (Fig. 5).

Figura 5. Curva de ajuste a una distribución de 3000 puntos.



Observamos que, si bien la curva se ajusta en algunos puntos, en otros francamente se "escapa". El desvío típico en este caso es de 10^{12} con el que parecía el mejor polinomio en la muestra. Incluso si graficáramos el error en función del grado de polinomio tendríamos el siguiente gráfico (Fig. 6):

Figura 6. Gráfico del error en función de la complejidad de un polinomio.



Lo que observamos es que, mientras un polinomio de grado 43 tendría un desvío típico de 2.7, el polinomio de grado 100 tiene un error inmenso: 1.000.000.000.000.

En lo que sigue describiremos, en forma prácticamente informal, una metodología capaz de seleccionar un modelo que se encuentre entre la mejor bondad de ajuste de los datos observados y la complejidad o riqueza del modelo.

DESCRIPCIONES MÍNIMAS (*MINIMUM DESCRIPTION LENGTH*): UNA OPCIÓN EN INFERENCIA ESTADÍSTICA

En la década de 1960, tres investigadores preocupados por el problema de la complejidad, la información y el azar llegan a conclusiones similares. Se trata de tal vez el más conocido de los investigadores en estadística, el ruso Andréi Kolmogórov (1903-1987); de un argentino genial, Gregory Chaitin (1947-), entonces un joven estudiante, y de Ray Solomonoff (1926-2009), un norteamericano experto en inteligencia artificial. Estos investigadores trabajaron en conceptos derivados de la teoría de la computación más que de la teoría de la probabilidad. El interés de sus trabajos se centra en la información de un objeto individual. Para ellos la información o complejidad de una cadena binaria (en esencia, todo puede ser codificado como cadena binaria de 1s y 0s) es equivalente a la longitud del programa más corto con el que una computadora puede generar la cadena. Crean la Teoría algorítmica de la información (TAI), con gran impacto −entre otras cosas− sobre el campo de la modelización estadística. La TAI es una aproximación computacional a la estadística que conceptualmente:

- No trata los datos como una muestra desde algún tipo de distribución.
- Considera los modelos como programas de computadora que describen o codifican los datos.
- Entiende que el programa más corto es una medida de la complejidad de la cadena de datos.

Escapa a los fines de este trabajo, sin embargo debido a consideraciones matemáticas y computacionales: este último punto, el de la medida de la complejidad tiene un gran problema: no es computable. En otras palabras no se puede resolver a través de un algoritmo.

Introdujimos en este trabajo la TAI ya que, si bien en este aspecto (no computabilidad) no es demasiado atractiva, ha derivado en metodologías de potencia elevada en modelización estadística.

LA NAVAJA DE OCCAM[3]

Para Guillermo de Occam, monje franciscano que estudió en Oxford, el conocimiento (científico sobre todo) debería ser austero y lo más cercano a la experiencia, evitando la creación indiscriminada de conceptos abstractos no verificables directamente por la experiencia. La economía en el conocimiento de los objetos para explicar se conoce como navaja de Occam. Se ha extendido este concepto filosófico al campo de la ciencia, sugiriéndose que la aplicación de la navaja de Occam implica que, cuando dos teorías en igualdad de condiciones tienen las mismas consecuencias (explican lo mismo), la teoría más simple tiene más probabilidades de ser correcta que la compleja. Si bien intuitivamente parece que las ideas de Occam son adecuadas en la elección de modelos, hipótesis y teorías, surgen diversos detractores del principio básico de Occam y de sus consecuencias cuando se formaliza dicho principio. Sin embargo, es indudable que en la actividad científica, la navaja de Occam guía la

3. "Pluralitas non est ponenda sine necessitate" (La pluralidad no se debe postular sin necesidad.).

actitud de los investigadores en relación con la aceptación de teorías rivales. Desde una aproximación estadística al tema podríamos aceptar que la explicación más simple y suficiente (abarcadora) es la más probable, pero no necesariamente la verdadera. Aunque, claro, los investigadores hace tiempo que tenemos problemas con un concepto absoluto de la "verdad", cuando de teorías científicas se trata. Como quiera que sea, la navaja de Occam es una buena heurística que puede guiar la toma de decisiones acerca de la elección frente a teorías rivales.

En general, creo que se ha hecho un mal ataque al principio de Occam confundiendo creencias u opiniones con hipótesis científicas. Está de más decir que la navaja de Occam es una heurística que se utiliza para elegir entre hipótesis o teorías que ya cumplen con una serie de criterios epistemológicos como la falsabilidad,[4] la capacidad explicativa, la capacidad predictiva, la concordancia con otras teorías, etc. Si dos teorías son equivalentes en estos aspectos, entonces sí podría ser mejor elegir la más simple o la que no contiene un montón de entidades innecesarias. En el problema de la selección de modelos y el problema de sobreajuste, este tema es crucial.

Sin embargo, más allá de la guía que significa el principio de Occam, necesitamos dirimir el tema de la selección de modelos de una manera formalizable y cuantificable.

EL MÉTODO DE DESCRIPCIONES DE LONGITUD MÍNIMA O *MINIMUM DESCRIPTION LENGTH* (MDL)

El primer concepto interesante del método MDL es que los modelos de los que hemos venido hablando son considerados códigos o métodos de descripción de los datos. En este sentido, la labor de la inferencia estadística sería seleccionar el modelo que lleva al menor tamaño de código de los datos. En otras palabras: un modelo es bueno si nos permite describir los datos observados con una cadena de código corto. Sencillamente, el requerimiento de MDL es que las hipótesis (modelos) se especifiquen formalmente como códigos. Pero… ¿qué es un código y cómo se lo relaciona con las probabilidades?

EL CONCEPTO DE CÓDIGO

- Sea A un alfabeto con un conjunto finito de símbolos.
- Sea A^n el conjunto de todas las cadenas de longitud n.
- Sea $A^* = Un = 0\ A^n$ la unión de las A^n.
- Sea la primera A° una cadena vacía = λ.
- Sea $x = a_1, \ldots, an \varepsilon A^*$ un mensaje, cadena de datos o secuencia.
- Un código es un mapa 1 a 1 desde A* a B*, siendo este el conjunto de todas las cadenas finitas binarias.

Para decirlo simplemente, un código es una descripción, básicamente una descripción de un sistema de símbolos en otra. Tal vez los ejemplos más comunes son el sistema Morse en telegrafía y el ASCII (American Standard Code for Information Interchange), en el manejo de símbolos por la computadora. Lo interesante en el abordaje MDL es considerar a los modelos como códigos o descripciones de los datos observados. Esto −como veremos− tiene algunas ventajas, en particular el uso de códigos particulares (códigos prefijo[5]), los cuales no serán tratados en este trabajo. Por otra parte, la relación necesaria se establece entre "longitud de códigos" y "medidas de probabilidad", relación comúnmente tratada en teoría de códigos.

La MDL fue primariamente presentada por Jorma J. Rissanen (1932-) en su libro *Stochastic Complexity in Statistical Inquiry*, editado por World Scientific en 1989 y que, a pesar de la modificación o enriquecimiento de algunos de los conceptos iniciales, tiene aún plena vigencia para introducirse en el tema. Inicialmente para Rissanen con un código aritmético se puede codificar cualquier conjunto de datos modelizados en cualquier modalidad estadística, además de una forma completamente mecánica y uniforme. Por otra parte, se plantea que la bondad de cualquier modelo se puede medir en términos de compresión de los datos. En forma muy simplificada podemos decir que la mejor explicación (modelo), dado un conjunto limitado de observables, es la que permite la mayor compresión de los datos. Cuanto más comprimimos los datos, más aprendemos acerca de la regularidad subyacente en los datos. Finalmente, la relación entre probabilidades y códigos viene dada de la siguiente manera:

Siendo P(x) una distribución sobre un alfabeto que asigna una probabilidad positiva para cada símbolo, entonces: para cada distribución P de probabilidad, existe un código único y decodificable tal que, para todos los resultados x, la longitud de código de x (L(x)), es igual a [–log p(x)], siendo p la función masa de P.

Volviendo al problema inicial del ajuste de polinomios o más ampliamente a la cuestión fundamental: ¿cómo decidimos acerca de explicaciones (modelos o teorías), que compiten acerca de los datos, dado un número limitado de observaciones? La MDL plantea que cualquier regularidad en los datos puede usarse para comprimir los datos. Es decir, describirlos usando menos símbolos que el número de símbolos necesario para describirlos literalmente.

4. Falsabilidad o refutabilidad: según Karl Popper, si conseguimos demostrar mediante la experiencia que un enunciado observable es falso, se sigue deductivamente que la proposición universal es falsa.

5. Un código prefijo es un código, típicamente un código de longitud variable, con la "propiedad de prefijo": ninguna palabra de código es prefijo de cualquier otra palabra de código del conjunto. Un código con las palabras de código {0, 10, 11} tiene la propiedad de prefijo; un código {0, 1, 10, 11} no la tiene, porque "1" es prefijo tanto de "10" como de "11". http://es.wikipedia.org/wiki/C%C3%B3digo_prefijo

MDL, IDEA BÁSICA

1. El objetivo de la inferencia estadística puede ser considerado como una búsqueda de regularidad en los datos.
2. La "regularidad" se puede identificar como habilidad para comprimir.

La MDL combina 1 y 2, viendo el aprendizaje como compresión de datos: dado un grupo de hipótesis H (modelos) y un conjunto de datos D, debemos tratar de encontrar hipótesis o combinación de hipótesis en H que más compriman a D. Para evitar confusiones, consideramos como "hipótesis puntual" a una distribución simple de probabilidades y específicamente como "modelo" a un conjunto de distribución de probabilidades o funciones con la misma forma funcional (p. ej., polinomios de grado 2).

Vayamos a un ejemplo simple tomado de F. Tusell y a su desarrollo, para entender la visión más elemental de MDL.

Ejemplo: tomemos el lanzamiento de una moneda 100 veces con una probabilidad de "caras" igual a un parámetro determinado = q, y obviamente una probabilidad de + = 1-q. Desde el punto de vista de cantidad de información, describir el resultado de un experimento de lanzar cien monedas al aire requiere 100 bits. Utilizamos "0" para codificar "+" y "1" para codificar "cara". Estos 100 bits son exactamente la cantidad de información necesaria para singularizar una cadena binaria de longitud 100 de entre las 2100 posibles cuando no hay nada que haga a algunas de ellas más plausibles que otras.

¿Cómo lo podemos hacer mejor?

Podríamos hacer una codificación de $X^{\rightarrow}$, la serie de resultados $(X_1,\ldots,X_{100})$, con un código de longitud $\lceil -\log p(X^{\rightarrow}|\theta) \rceil$, si conociéramos q. Dado que no conocemos q, debemos darle un valor a q, para posteriormente codificarlo. Luego podemos codificar $X^{\rightarrow}$.

Este agregado del número de bits necesario para codificar q da lugar a una versión inicial bastante preliminar y con algunos problemas pero que cumple en esencia con MDL.

Para la codificación de q, en principio se pueden dar dos situaciones:

- Que tengamos información a priori sobre los parámetros, que se pueda traducir a una distribución a priori con densidad $\pi(\theta^{\rightarrow})$.
- Que $\theta^{\rightarrow}$ sea un número real que requeriría infinitos bits para fijar con exactitud. Sin embargo, podemos trabajar con una versión truncada.

Primero debemos fijar "q" dígitos binarios y definir una precisión $\delta = 2^{-q}$.

Con k parámetros tendríamos:

$$-\log_2 [((]) \prod_{i=1}^{k} \delta i$$

La versión cruda de MDL propone tomar el modelo que minimiza la longitud total de código, la necesaria para los datos $X^{\rightarrow}$, más la necesaria para los parámetros. En síntesis:

$$\begin{aligned} MDL &= -\log_2 P(\vec{x}|\vec{\theta}) + l(\vec{\theta}) \\ &= -\log_2 P(\vec{x}|\vec{\theta}) - \log_2 \pi(\vec{\theta}) - \sum_{i=1}^{k} \log_2 \delta_i. \end{aligned}$$

Donde $l(\theta^{\rightarrow})$ es la longitud de código necesaria para transmitir el o los parámetros empleados.

Siguiendo con el ejemplo:

Más arriba hablamos de trabajar con una versión truncada de nuestro $\theta^{\rightarrow}$. Arbitrariamente trunquemos en 8 bits, $\delta = 2\text{-}8 \approx 0{,}003906$. Consideremos un Θ_*, que sea el conjunto de valores que puede adoptar el parámetro truncado. Por otra parte si tenemos una distribución a priori uniforme $\pi(\theta)$ para $0 \leq \theta \leq 1$, con nuestra distribución a priori $\pi(\theta) = 1$. MDL daría el valor:

$$MDL = \min_{\theta \in \Theta_*} \left\{ -\log_2 \theta^{60}(1-\theta)^{40} - \log_2 \pi(\theta) - \log_2 \delta \right\}$$

teniendo en cuenta 100 lanzamientos con sesenta caras y cuarenta cruces.

Con un δ constante minimicemos el primer término:

Con el truncamiento tenemos (153 + ½)/256 = 0.599609 y (154+½)/256 = 0.603516, puntos medios de intervalos de longitud 1=256 en que se subdivide [0; 1] cuando se emplea precisión $\delta = 2^{-8}$. Nuestra longitud de descripción queda en 97.0951 + 8 = 105.0951 bits.

Aun un modelo binomial con cien parámetros, si bien asegura el resultado, requiere muchos más parámetros y su coste de codificación.

El ejemplo es elemental, no realza la potencia de MDL, pero nos da una idea aproximada del método, aunque −claro− en su versión más limitada o cruda.

CONCLUSIONES

En relación con el problema del sobreajuste, la MDL provee un método eficiente que se sitúa entre la selección de la mejor bondad de ajuste y la complejidad de los modelos involucrados. En relación con la selección de polinomios o cualquier tipo de hipótesis ajustadas a los datos y la selección de modelos, una concepción simple aunque informal de MDL es que hacemos inferencias interesantes en la medida en que somos capaces de comprimir los datos.

BIBLIOGRAFÍA

- Grünwald PD. The minimum description length principle. Cambridge, Mass: The MIT Press; 2007. (Adaptive computation and machine learning series).
- Grünwald PD, Myung IJ, Pitt, MA. Advances in minimum description length: theory and applications. Cambridge, Mass: MIT Press; 2005.
- Nannen V. The paradox of overfitting [Tesis, Internet]. Amsterdam: Dutch National Research Institute for Mathematics and Informatics; 2003. [Consulta: 13/10/2011]. Disponible en: http://volker.nannen.com/work/mdl/.
- Rissanen J. Information and complexity in statistical modeling. New York: Springer; 2010.
- Rissanen J. Stochastic complexity in statistical inquiry. River Edge, NJ: World Scientific; 1989. (Series in computer science).
- Tusell F. Complejidad estocástica [Internet]. [Consulta: 13/10/2011]. Disponible en: http://www.et.bs.ehu.es/~etptupaf/pub/papiros/complex.pdf.