comparación y predicción de clases y tests de permutaciones para los niveles de significación (*http://linus.nci.nih.gov/BRB-ArrayTools.html*).

REFERENCIAS

Affymetrix 2002. Statistical algorithms description document (MAS v5.0).

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society,* B, 57:289-300.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. 2002. A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics,* 18:105S-110S.

Hardin, J. 2005. Microarray data from a statistician´s point of view, *STATS,* 42:4-13.

Harrington, C.A., Rosenow, C. and Retief, J. 2000. Monitoring gene expression using DNA microarrays, *Current opinion in Microbiology*, 3:285-291.

Lockhart, D., Dong, H., Bryne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology,* 14:1675-1680.

Schena, M., Shalon, D., Davis, R., and Brown, P. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science,* 270:467-470.

Storey, J. 2002. A direct approach to false discovery rates, *Journal of the Royal Statistical Society,* B, 64:479-498.

2. ARTÍCULOS DE INVESTIGACIÓN OPERATIVA

MÁQUINAS DE VECTOR DE APOYO: PROBLEMAS DE PROGRAMACIÓN MATEMÁTICA EN CLASIFICACIÓN

Emilio Carrizosa y Belén Martín-Barragán
Dpto. Estadística e Investigación Operativa. Universidad de Sevilla
ecarrizosa@us.es,belmart@us.es

1. Introducción

En la última década, la capacidad de almacenamiento de información digital se ha duplicado cada nueve meses. Crece, por tanto, a una velocidad muy superior a la prevista por la ley de Moore para el crecimiento de la capacidad de cálculo, [18, 25], provocando la aparición de las denominadas *fosas de datos*, [18]: datos que son almacenados y descansan en paz, sin que nadie los reclame o los recuerde.

La constatación de la existencia de tales fosas de datos, y la consiguiente pérdida de oportunidades de avance en el conocimiento o de negocio, está provocando un enorme interés por el desarrollo de técnicas que, complementando a las previamente existentes, permitan obtener información desconocida y potencialmente útil de datos provinientes de campos tan diversos como la Bioinformática (expresión genética,…), gestión de clientes (fuga de clientes, análisis de la cesta de la compra,…), la banca (valoración de riesgo en créditos, detección de uso fraudulento de tarjetas de crédito, …), Internet (clasificación de páginas web, filtrado de correo indeseado, …), [1, 2, 3, 16, 19, 20, 22, 35].

Hablamos, usando una denominación de moda en los medios científicos, y, en particular, en las líneas editoriales de algunas de las revistas de más alto índice de impacto en nuestra área de conocimiento, de la

Minería de Datos. Las referencias [2, 8, 22, 23, 34] pueden servir de introducción al tema.

Examinando, por ejemplo, las distintas opciones del software de código abierto `Weka`, [33], descrito en [34], se observa que uno de los pilares de la Minería de Datos, aunque bastante anterior a ésta, es la *Clasificación*. Encontramos junto a procedimientos bien conocidos en la comunidad estadística, como la regresión logística, los árboles de clasificación, los modelos bayesianos o las redes de neuronas artificiales, otros más recientes, como el que nos ocupa en estas líneas: las *Máquinas de Vector de Apoyo* (en inglés, Support Vector Machines), que ha saltado del mundo del Aprendizaje Estadístico, [12, 31, 32] al de las aplicaciones pasando por el de la Programación Matemática. Véase [4, 5, 6, 11, 26, 27, 29, 30, 36] para otros métodos de clasificación que, como las Máquinas de Vector de Apoyo, usan técnicas avanzadas de Programación Matemática.

2. El problema de clasificación

Tenemos un conjunto de objetos Ω. Cada objeto $u \in \Omega$ tiene dos componentes $u = (x^u,\ c^u)$, donde $x^u \in ¡^{\,p}$ representa el vector de variables predictoras, y $c^u \in C$ es la clase a la que pertenece u. Por simplicidad en la exposición, supondremos el caso binario, $C = \{-1,\ 1\}$.

Se dispone de un conjunto no vacío de objetos $I \subset \Omega$, la *muestra de aprendizaje.* El objetivo es predecir, a partir de I, la clase c^v a la que pertenece un objeto $v \in \Omega$ conociendo solo x^v. Para ello se buscan $\omega \in ¡^{\,p}$, $\beta \in ¡$, se construye la función de evaluación f,

$$f(x) = \omega^T x + \beta, \qquad (1)$$

y con ésta, la *regla lineal de clasificación* que clasifica en el grupo 1 a aquellos $x \in ¡^{\,p}$ con $f(x) > 0$ y en el grupo -1 a los x con $f(x) < 0$. Los x con $f(x) = 0$ serán clasificados siguiendo alguna regla predeterminada.

La primera pregunta que nos hacemos es si existen o no ω, β tales que la correspondiente regla lineal clasifique correctamente el 100% de los individuos de I,

$$y^u\left(\omega^T x^u + \beta\right) > 0 \quad \forall u \in I. \qquad (2)$$

Cuando el sistema (2) sea factible, diremos que I es *separable linealmente.* Es fácil comprobar (usando, por ejemplo, resultados básicos de dualidad en Programación Lineal), que la separabilidad lineal de I es equivalente a que los cierres convexos de los conjuntos

$\{x^u : u \in I,\ c^u = 1\}$, $\{x^u : u \in I,\ c^u = -1\}$

sean disjuntos. Esta condición puede comprobarse numéricamente en tiempo polinómico en el cardinal de *I* y la dimensión *p* de los datos.

2.1. El caso separable.

Cualquier $(\omega,\ \beta)$ solución de (2) satisface que $\omega \neq 0$. En particular, $(\omega,\ \beta)$ genera un hiperplano, $\{x \in ¡^{\,p} : \omega^T x + \beta = 0\}$, de modo que el semiespacio $\{x \in ¡^{\,p} : \omega^T x + \beta > 0\}$ contiene al conjunto $\{x^u : u \in I,\ c^u = 1\}$, y el semiespacio $\{x \in ¡^{\,p} : \omega^T x + \beta < 0\}$ contiene al conjunto $\{x^u : u \in I,\ c^u = -1\}$.

Cuando *I* es linealmente separable, el sistema (2) tiene infinitas soluciones, que generan infinitos hiperplanos distintos. ¿Cómo elegimos una de estas soluciones? La calidad de la clasificación, *sobre la muestra de aprendizaje*, es idéntica: todas clasifican correctamente el 100% de *I.* Sin embargo, no todas parecen igualmente razonables. En la Figura 1 podemos ver dos hiperplanos que separan los grupos de *I* (círculos y cuadrados). Intuitivamente, podemos pensar que el hiperplano representado por un trazo grueso es más conveniente que el de trazo fino. En particular, este último asigna al objeto representado con '?' la clase cuadrado, cuando parece mucho más verosímil que pertenezca a la clase de los círculos.

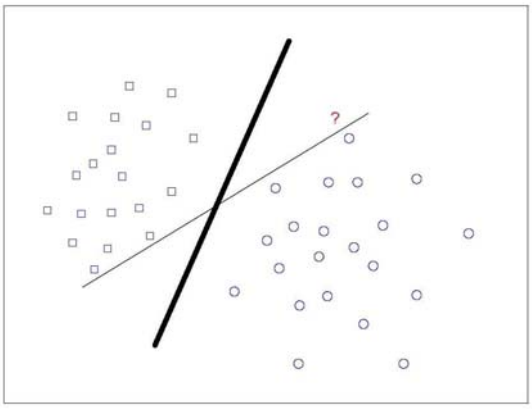

Figura 1. ¿Dos reglas que clasifican igual de bien?

El ejemplo anterior nos indica intuitivamente la conveniencia de elegir un hiperplano que esté *alejado* de las dos clases. Las Máquinas de Vector de Apoyo se basan precisamente en este principio, como a continuación se describe. Se fija una norma $\|\cdot\|$ en $\mathbb{R}^p$ para medir las distancias (usualmente la euclídea). Para un objeto $u \in I$, la distancia entre x^u y el semiespacio en el que quedará clasificado incorrectamente viene dada por

$$\rho^u(\omega, \beta) = \max\left\{\frac{y^u(\omega^T x^u + \beta)}{\|\omega\|^o}, 0\right\}, \quad (3)$$

e.g. [7], donde $\|\cdot\|^o$ denota la norma dual a $\|\cdot\|$.

Se define el *margen* en la muestra de aprendizaje I como el mínimo ρ^u:

$$\rho^I(\omega, \beta) = \min_{u \in I} \rho^u(\omega, \beta). \quad (4)$$

El clasificador buscado es aquél que no sólo clasifique correctamente a todos los objetos de I, sino que tenga margen máximo. Geométricamente, la búsqueda del clasificador de máximo margen puede verse como un problema de Localización, [8], pues el problema es equivalente a construir la banda de máxima anchura (las distancias medidas con la norma $\|\cdot\|$) que deja un grupo a cada lado, como se muestra en las Figuras (2)-(3).

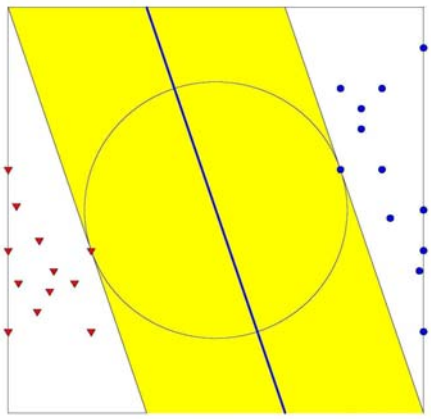

Figura 2. Máximo margen (norma l_2)

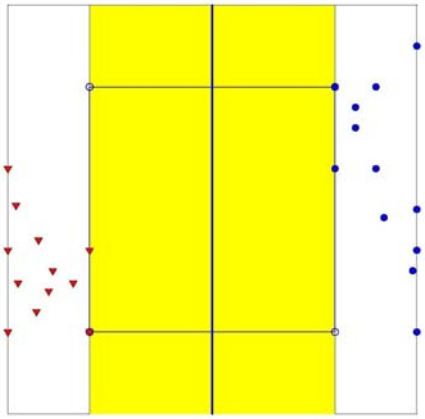

Figura 3. Máximo margen (norma l_∞)

Usando la homogeneidad de la función margen, el problema de maximización del margen puede ser formulado como el siguiente problema convexo con restricciones lineales:

$$\begin{aligned} \min \quad & \|\omega\|^o \\ \text{s.a.:} \quad & y^u\left(\omega^T x^u + \beta\right) \geq 1 \quad \forall u \in I \\ & \omega \in \mathbb{R}^p, \quad \beta \in \mathbb{R}. \end{aligned} \quad (5)$$

Si, para medir las distancias hemos usado, como en el ejemplo de la Figura 3, una norma $\|\cdot\|$ poliédrica, (i.e., cuya bola unidad es un poliedro) su dual $\|\cdot\|^o$ también es poliédrica, y por tanto (5) puede reformularse como un problema de Programación Lineal, resoluble, incluso para grandes bases de datos, con optimizadores comerciales como CPLEX, [21]. El caso más estudiado en la literatura, no es, sin embargo, el que tiene como $\|\cdot\|$ una norma poliédrica, sino la euclídea. Entonces (5) es equivalente al

siguiente problema cuadrático convexo con restricciones lineales:

$$\begin{array}{ll} \min & \omega^T\omega \\ \text{s.a.:} & y^u\left(\omega^T x^u+\beta\right)\geq 1 \quad \forall u\in I \\ & \omega\in \mathbb{R}^p,\ \beta\in\mathbb{R}, \end{array} \tag{6}$$

que puede resolverse, por ejemplo, usando planos de corte, [28].

2.2. El caso no separable.

En la Sección 2.1 hemos supuesto que I era linealmente separable. Si no es el caso, el problema (5) es infactible, por lo que deben aplicarse enfoques alternativos. Uno de estos enfoques consiste en aplicar a los datos, como preprocesamiento, una transformación $\phi:\mathbb{R}^p\to F$, donde F es un espacio vectorial de mayor dimensión (posiblemente infinita), de manera que, en el nuevo espacio, la muestra de aprendizaje $\hat{I}=\{(\phi(x^u),\ c^u): u\in I\}$ sea linealmente separable, [10, 14, 15, 17, 24]. Conseguido esto, se buscan $\omega\in F$, $\beta\in\mathbb{R}$, y se construye la regla de clasificación, que estaría basada en la función f,

$$f(x)=\omega^T\phi(x)+\beta, \tag{7}$$

que asigna, como es habitual, al grupo 1 si $f(x)>0$, y al grupo -1 si $f(x)<0$. Esta regla es lineal sobre los datos transformados, pero no lineal en el espacio original $\mathbb{R}^p$. El problema de maximización del margen es

$$\begin{array}{ll} \min & \|\omega\|^2 \\ \text{s.a.:} & y^u\left(\omega^T\phi(x^u)+\beta\right)\geq 1 \quad \forall u\in I \\ & \omega\in F,\ \beta\in\mathbb{R}. \end{array} \tag{8}$$

Para el caso en que $\|\cdot\|$ sea poliédrica y F tenga dimensión grande (pero finita), (8) se escribe como un problema lineal de gran tamaño, para cuya resolución son especialmente convenientes técnicas de generación de columnas, permitiendo al mismo tiempo hacer selección automática de variables, [10].

Si, en cambio, usamos la norma euclídea para medir las distancias en el espacio transformado, (8) es un problema cuadrático convexo cuyo dual es

$$\begin{array}{ll} \max & \sum_{u\in I}\lambda^u-\frac{1}{2}\sum_{u,\ v\in I}\lambda^u\lambda^v y^u y^v\phi(x^u)^T\phi(x^v) \\ \text{s.a.:} & \sum_{u\in I}y^u\lambda^u=0 \\ & \lambda^u\geq 0,\forall u\in I \end{array} \tag{9}$$

Definiendo el *núcleo*

$K:(x,\ y)\in\mathbb{R}^p\times\mathbb{R}^p\to\phi(x)^T\phi(y)\in\mathbb{R}$,

(9) se convierte en

$$\begin{array}{ll} \max & \sum_{u\in I}\lambda^u-\frac{1}{2}\sum_{u,\ v\in I}\lambda^u\lambda^v y^u y^v K(x^u,\ x^v) \\ \text{s.a.:} & \sum_{u\in I}y^u\lambda^u=0 \\ & \lambda^u\geq 0,\forall u\in I. \end{array} \tag{10}$$

Para poder resolver (10), ni siquiera es necesario conocer ϕ, sino un algoritmo de evaluación del núcleo K que induce.

El problema de maximización resultante es cóncavo cuadrático, con tantas variables como elementos en I, y con una única restricción, lineal, junto a las de no negatividad. La dimensión de este problema es, por tanto, independiente de la dimensión p de los datos del problema original y de la dimensión de F. Esto hace de (10) una formulación especialmente atractiva en aplicaciones con no demasiados datos, pero de alta dimensionalidad, como las de, por ejemplo, [16, 35]. Para más detalles, véase, por ejemplo [13, 24].

Una estrategia alternativa (y a veces complementaria) para abordar el caso no separable, es la que se basa en la maximización del *margen débil*, [12, 13, 24], en la que, partiendo del problema infactible (6), se perturban sus restricciones para hacerlo factible, introduciendo una penalización en el objetivo para controlar la perturbación introducida. Así se obtiene el problema (siempre factible)

$$\begin{array}{ll} \min & \omega^T\omega+C(\|\xi\|_r)^r \\ \text{s.a.:} & y^u\left(\omega^T x^u+\beta\right)+\xi^u\geq 1, \quad \forall u\in I \\ & \omega\in\mathbb{R}^p,\ \beta\in\mathbb{R},\ \xi\in\mathbb{R}^{|I|}, \end{array} \tag{11}$$

donde $\|g\|_r$ denota la norma l_r y $C>0$ es una constante que se usa para equilibrar la perturbación ξ y el margen en los puntos correctamente clasificados, usualmente elegida por técnicas de validación cruzada.

Terminamos el análisis comentando que, en una gran variedad de aplicaciones, la importancia del error cometido al clasificar incorrectamente un objeto depende fuertemente del grupo al que éste pertenece: los costes asociados a los falsos positivos y a los falsos negativos pueden ser muy distintos, y, como en el caso del diagnostico de enfermedades, puede ser difícil cuantificar esa importancia asignando costes. En tal caso podemos plantear el problema biobjetivo de maximización simultánea del margen en cada uno de los dos grupos. Como se prueba en [9], para el caso euclídeo, las soluciones eficientes resultan ser hiperplanos paralelos a la solución del problema de máximo margen clásico (8).

Fijando el ω obtenido al resolver (8), y dejando variar β, se obtienen las distintas soluciones eficientes, que dan distintos niveles de compromiso entre los falsos positivos y los falsos negativos sobre la muestra de aprendizaje *I*. Esto se ilustra en la Figura 4, en la que aparecen en línea gruesa los distintos compromisos así obtenidos entre falsos positivos y falsos negativos en *I*, siendo éstos una guía a los que obtendríamos sobre Ω, representados en trazo fino.

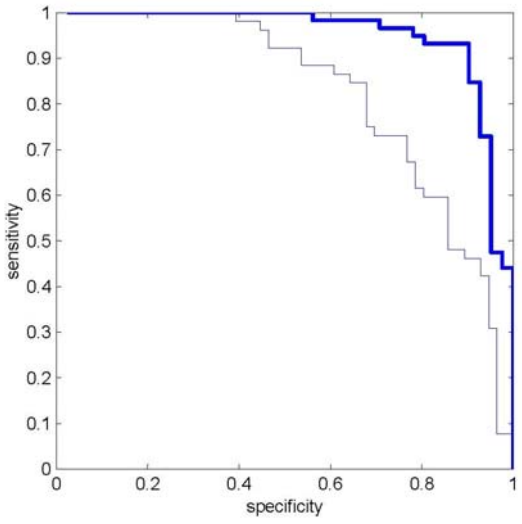

Figura 4. Clasificadores eficientes

3. Conclusiones

La construcción de reglas de clasificación basadas en la maximización del margen está mostrando ser extraordinariamente eficaz en diversos campos aplicados de la Minería de Datos.

A pesar de los grandes avances obtenidos en los últimos años, son aún muchos los aspectos (de modelado, de tipo numérico, algorítmico) por explorar.

Con estas líneas esperamos haber despertado la curiosidad por una técnica de la que no hemos explicado ni el origen de su exótico nombre (por cierto, el término inglés Support Vector Machines no debe traducirse como "¡Apoye las máquinas vectoriales!"), que goza de creciente aceptación entre los usuarios de la Minería de Datos, y, esperemos que cada vez más, de los estadísticos y los investigadores de operaciones españoles.

Agradecimientos

El trabajo ha sido parcialmente subvencionado por el Ministerio de Ciencia y Tecnología, a través de los proyectos BFM2002-04525-C02-02 y BFM2002-11282-E, y por el Plan Andaluz de Investigación, proyecto FQM-329

Referencias

[1] Alexe, S., Blackstone, E., Hammer, P., Ishwaran, H., Lauer, M. y Pothier Snader, C.E. Coronary risk prediction by logical analysis of data. *Annals of Operations Research*, 119:15-42, 2003.

[2] Apte, C. The big (data) dig. *OR/MS Today*, February 2003.

[3] Apte, C., Liu, B., Pednault, E.P.D. y Smyth, P. Business applications of Data Mining. *Communications of the ACM*, 45:49-53, 2002.

[4] Bennet, K.P. y Mangasarian, O.L. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23-24, 1992.

[5] Bradley, P., Mangasarian, O. y Musicant, D. Optimization methods in massive datasets. En Abello, J., Pardalos, P.M., and Resende, M.G.C.,

editores, *Handbook of Massive Datasets*, pag. 439-472. Kluwer Academic Pub., 2002.

[6] Bradley, P.S., Fayyad, U.M. y Mangasarian, O.L. Mathematical programming for data mining: formulations and challenges. INFORMS Journal on Computing, 11(3):217-238, 1999.

[7] Carrizosa, E. y Fliege, J. Generalized goal programming: Polynomial methods and applications. Mathematical Programming, 93:281-303, 2002.

[8] Carrizosa, E. y Martín-Barragán, B. Problemas de clasificación: una mirada desde la localización. En Avances en localización de servicios y sus aplicaciones. B. Pelegrín (Ed.), pp. 249-276. Servicio de Publicaciones de la Universidad de Murcia, 2005.

[9] Carrizosa, E. y Martín-Barragán, B. Two-group classification via a biobjective margin maximization model. Por aparecer en *European Journal of Operational Research*.

[10] Carrizosa, E., Martín-Barragán, B. y Romero-Morales, M.D. A Biobjective Model to Select Features With Good Classification Quality and Low Cost. Proceedings of the Fourth IEEE International Conference on Data Mining. IEEE Publications, 2004. Pag. 339-342.

[11] Carrizosa, E. y Plastria, F. Optimal expected-distance separating halfspace. Report MOSI/7, Vrije Universiteit Brussel, 2004.

[12] Cortes, C. y Vapnik, V. Support-vector network. Machine Learning, 1:113-141, 1995.

[13] Cristianini, N. y Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambridge University Press, 2000.

[14] Demiriz, A., Bennett, K.P. y Shawe-Taylor, J. Linear programming boosting via column generation. Machine Learning, 46(1):225-254, 2002.

[15] Duarte Silva, A.P. y Stam, A. Second order mathematical programming formulations for discriminant analysis. *European Journal of Operational Research*, 72:4-22, 1994.

[16] Efron, B., Tibshirani, R., Storey, J. y Tusher, V. Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association, 96:1151-1160, 2001.

[17] Falk, J.E. y Karlov, V.E. Robust separation of finite sets via quadratics. Computers and Operations Research, 28:537-561, 2001.

[18] Fayyad, U. y Uthurusamy, R. Evolving data mining into solutions for insight. Communications of the ACM, 45:28-31, 2002.

[19] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. y Lander, E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439):531-537, 1999.

[20] Han, J., Altman, R.B., Kumar, V., Mannila, H. y Pregibon, D. Emerging scientific applications in Data Mining. Communications of the ACM, 45:54-58, 2002.

[21] ILOG CPLEX 8.1 User's Manual. http://www.pcs.cnu.edu/~riedl/software/cplex81/doc/userman/onlinedoc/

[22] Hand, H., Mannila, H. y Smyth, P. Principles of Data Mining. MIT Press, 2001.

[23] Hastie, T., Tibshirani, R., y Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction Springer, 2001.

[24] Herbrich, R. Learning Theory Classifiers. Theory and Algorithms. MIT Press, 2002.

[25] Informe de Intel sobre la ley de Moore. http://www.intel.com/research/silicon/mooreslaw.htm

[26] Mangasarian, O.L. Mathematical programming in data mining. Data Mining and Knowledge Discovery, 42(1):183-201, 1997.

[27] Piramuthu, S. Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research, 156:483-494, 2004.

[28] Platt, J. Fast training of support vector machines using sequential minimal optimization. En *Advances in Kernel*

Methods - Support Vector Learning B. Schlkopf, C. J. C. Burges, y A. J. Smola (Eds.), pp. 185-208. MIT Press, 1999.

[29] Rubinov, A.M., Bagirovand, A.M., Soukhoroukova, N.V. y Yearwood, J. Unsupervised and supervised data classification via nonsmooth and global optimization. TOP, 11(1):1-93, 2003.

[30] Stam, A. Nontraditional approaches to statisical classification: Some perspectives on l_p -norm methods. Annals of Operations Research, 74:1-36, 1997.

[31] Vapnik, V. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[32] Vapnik, V. Statistical Learning Theory. Wiley, 1998.

[33] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

[34] Witten, I.H., y Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[35] Xie, D., Singh, S.B., Fluder, E.M. y Schlick, T. Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases. Mathematical Programming, 95:161-185, 2003.

[36] Zopounidis, C. y Doumpos, M. Multicriteria classification and sorting methods. European Journal of Operational Research, 138:229-246, 2002.

3. ARTÍCULOS DE APLICACIÓN

ESTUDIO DE FUNCIONALIDAD EN CENTROS DE FITNESS O GIMNASIOS

Arturo Alvear González
Estudiante de Estadística
Universidad de Salamanca

INTRODUCCIÓN

Son bien conocidas las ventajas que conlleva la práctica de cualquier disciplina deportiva: mejor calidad de vida, reducción de estrés, mejora física y de la autoestima, prevención de enfermedades, etc. En los últimos años, la incorporación de nuevas disciplinas deportivas, como por ejemplo aeróbic o determinadas modalidades orientales, a nuestra práctica deportiva han contribuido a reducir la monotonía y a aumentar la versatilidad a la hora de ejercitarse.

La finalidad que se persigue en este estudio es la de tratar de probar estadísticamente algunas de las ventajas que se consiguen en los centros de fitness o en gimnasios, así como estudiar las posibles relaciones, estadísticamente significativas, entre las variables de interés consideradas en este tipo de centros.

Algunos estudios anteriores, fundamentalmente llevados a cabo por los propios preparadores y deportistas, ya han puesto de manifiesto que aunque en tales centros no se consiguen milagros, si que ayudan a mejorar nuestra calidad de vida y nuestra salud, siendo su utilidad de gran importancia para personas mayores (véase en este sentido el libro de Beraldo y Pollet (1995)).

POBLACIÓN CONSIDERADA Y VARIABLES ANALIZADAS

El estudio se ha realizado en base a los datos suministrados por un gimnasio de capacidad media (360 sujetos, 73% hombres y 27% mujeres) localizado en la provincia de Burgos, siendo nuestras conclusiones extrapolables a centros de características similares al considerado.