ALGORITHM FOR LINEAR PATTERN SEPARATION

Dr.prof. V. Moraru, M. Rusu,
Technical University of Moldova

INTRODUCTION

The linear separation problem of data sets is an important concept in the data analysis. This theory is widely used and applied in many fields such as pattern recognition [1], decision making [2], disease diagnosis [3], biometrics [4], automatic document processing [5] and others.

For example, there are two sets of objects (attributes)

$$A = \{a^1, a^2, ..., a^m\}$$

$$B = \{b^1, b^2, ..., b^k\}$$

where a^i and $b^i \in R^n$ for $\forall i = 1, 2, ..., m$, $\forall j = 1, 2, ..., k$ and $A \bigcap B = \varnothing$.

The aim of the linear separation is to build a decision function of the form:

$$f(x) = w^T x - w_0$$

which divides the space into two subsets such

$$f(a^i) < 0 \text{ şi } f(b^i) > 0$$

where $\forall i \in A$ and $\forall j \in B$.

Here w is a column vector of R^n, and w_0 is a scalar: $w_0 \in R$. The symbol „*T*” means transposition operation, in this case all vectors are column vectors.

Separation (classification) can be formulated as a quadratic programming problem. In this paper we will discuss three models of linear separation. For one of these models we will introduce an effective procedure for numerically solving.

1. THE MODEL OF MAXIMUM SEPARATION [6]

We choose a hyperplane

$$w^T x - w_0 = 0 \tag{1}$$

which maximizes the minimum distance from any point of the sets A and B. The distance from the point $z \in R^n$ to the hyperplane (1) is equal to

$$d(z) = \frac{f(z)}{\| w \|} = \frac{w^T z - w_0}{\| w \|}.$$

Here and below $\| \, . \, \|$ is the Euclidean norm. So

$$d(a^i) = \frac{w_0 - w^T a^i}{\| w \|},$$

$$d(b^i) = \frac{w^T b^i - w_0}{\| w \|},$$

$$d_{\min} = \min\{d(a^1), d(a^2), ..., d(a^m), d(b^1), ..., d(b^k)\}.$$

The problem consists in maximizing the size $d_{\min}$ which is equivalent to the following problem:

$$\left. \begin{array}{l} \delta \to \min \\ \text{subject to} \\ w_0 - w^T a^i \geq \delta \| w \|, i = 1,2,...,m \\ w^T b^i - w_0 \geq \delta \| w \|, i = 1,2,...,k. \end{array} \right\} \tag{2}$$

The problem (2) is a nonlinear problem (non-convex) towards the unknowns

$$\delta \in R, \ w_0 \in R, \ w \in R^n.$$

We impose the following restriction: $\| w \| = 1$. Then, introduce the variables

$$u = \frac{w}{\delta} \in R^n \text{ and } v = \frac{w_0}{\delta} \in R,$$

problem (2) is equivalent to the following problem:

$$\left. \begin{array}{l} \frac{1}{2} u^T u \to \min \\ \text{subject to} \\ [b^i]^T u - v \geq 1, i = 1,2,...,k, \\ -[a^i]^T u + \delta \geq 1, i = 1,2,...,m, \\ u \in R^n, v \in R. \end{array} \right\} \tag{3}$$

The problem (3) is a convex quadratic programming problem with $(n + 1)$ variables and $(m + k)$ linear constraints. If $u^* \in R^n$ and $v^* \in R$ is an optimal solution of problem (3), then the solution of (2) is:

$$w^* = \frac{u^*}{\| u^* \|}, w_0^* = \frac{v^*}{\| u^* \|}, \delta^* = \frac{1}{\| u^* \|}.$$

The vector w^* is perpendicular to the considered hyperplane and has a length equal to 1, and the size w_0^* is shortest distance between the hyperplane and origin of the coordinate system.

2. SUPPORT VECTOR MACHINES (SVM) [7]

In this method the elements of set A are labeled with $t = -1$, and the elements of set B with $t = 1$. In other words,

$$t(x) = \begin{cases} -1, \text{if } f(x) < 0, \\ +1, \text{if } f(x) > 0, \end{cases}$$

i.e.

$$\left.\begin{aligned} w^T x - w_0 < 0, \text{if } t(x) = -1, \\ w^T x - w_0 > 0, \text{if } t(x) = +1 \end{aligned}\right\} \qquad (4)$$

We note that the hyperplane (1) does not change if w and w_0 are multiplied by the same positive constant. It is convenient to choose this constant such bellow

$$\left.\begin{aligned} w^T a^i - w_0 = -1, i = 1,2,...,m, \\ w^T b^i - w_0 = +1, i = 1,2,...,k. \end{aligned}\right\}$$

Thus, taking into consideration the (4) we can write

$$t(x^i)(w^T x^i - w_0) \geq 1, \forall x^i \in A \cup B.$$

Determination of the optimal separating hyperplane is reduced to solving the problem:

$$\left.\begin{aligned} &\frac{1}{2} w^T w \to \min \\ &\text{subject to} \\ &t(x^i)(w^T x^i - w_0) \geq 1, \forall x^i \in A \cup B. \end{aligned}\right\} \qquad (5)$$

The problem (5) is similar to the problem (3). The constraints of (5) ensure that in the optimal solution w^*, w_0^* we have:

$$f(x^i) = \begin{cases} +1, \text{for } t(x^i) = 1, \\ -1, \text{for } t(x^i) = -1. \end{cases}$$

3. REFORMULATING THE PROBLEM IN THE TERMS OF THE CONVEX QUADRATIC PROGRAMMING

Let the convex hulls of the sets A and B:

$$conv(A) = \left\{ x{:}x = \sum_{i=1}^{m} \alpha_i a^i, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0, i = 1,2,...,m \right\}$$

$$conv(B) = \left\{ y{:}y = \sum_{i=1}^{k} \beta_i b^i, \sum_{i=1}^{k} \beta_i = 1, \beta_i \geq 0, i = 1,2,...,k \right\}$$

Then the problem of optimal separation of the sets A and B can be formulated as:

$$\left.\begin{aligned} &\frac{1}{2} \| x - y \|^2 = \frac{1}{2}(x - y)^T (x - y) \to \min \\ &\text{subject to} \\ &\sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0, i = 1,2,...,m, \\ &\sum_{i=1}^{k} \beta_i = 1, \beta_i \geq 0, i = 1,2,...,k. \end{aligned}\right\} \qquad (6)$$

Let the matrices $U_{m\times m}$, $V_{k\times k}$ and $Z_{m\times k}$ defined in the following way:

$$U_{m\times m} = \begin{pmatrix} (a^1,a^1) & (a^1,a^2) & ... & (a^1,a^m) \\ (a^2,a^1) & (a^2,a^2) & ... & (a^2,a^m) \\ ... & ... & ... & ... \\ (a^m,a^1) & (a^m,a^2) & ... & (a^m,a^m) \end{pmatrix}$$

$$V_{k\times k} = \begin{pmatrix} (b^1,b^1) & (b^1,b^2) & ... & (b^1,b^k) \\ (b^2,b^1) & (b^2,b^2) & ... & (b^2,b^k) \\ ... & ... & ... & ... \\ (b^k,b^1) & (b^k,b^2) & ... & (b^k,b^k) \end{pmatrix}$$

$$Z_{m\times k} = \begin{pmatrix} (a^1,b^1) & (a^1,b^2) & ... & (a^1,b^k) \\ (a^2,b^1) & (a^2,b^2) & ... & (a^2,b^k) \\ ... & ... & ... & ... \\ (a^m,b^1) & (a^m,b^2) & ... & (a^m,b^k) \end{pmatrix}$$

Also note

$$\gamma = (\alpha_1, \alpha_2, ..., \alpha_m, \beta_1, \beta_2, ..., \beta_k)^T \in R^{m+k},$$

$$Q = \begin{pmatrix} U_{m\times m} & -Z_{m\times k} \\ -Z^T_{k\times m} & V_{k\times k} \end{pmatrix},$$

$$B = \begin{pmatrix} 1 \ 1 \ ... \ 1 & 0 \ 0 \ ... \ 0 \\ \underbrace{0 \ 0 \ ... \ 0}_{m \ ori} & \underbrace{1 \ 1 \ ... \ 1}_{k \ ori} \end{pmatrix},$$

$$e = \begin{pmatrix} 1 & 1^T \end{pmatrix}.$$

The matrices $U_{m\times m}$, $V_{k\times k}$ and Q are positive semidefined.

With these notations the problem (6) becomes:

$$\left.\begin{array}{l}\frac{1}{2}\gamma^T Q\gamma \to \min \\ \text{subject to} \\ B\gamma = e, \\ \gamma \geq 0.\end{array}\right\} \quad (7)$$

Note that the objective function depends only on the scalar product of the vectors a^i and b^j. The problem (7) is a convex programming problem and therefore it has a global optimal solution.

4. SVM AS A PROBLEM TO SOLVE A SYSTEM OF EQUATIONS

Using Kuhn-Tucker theorem, it demonstrates that the dual problem (5) is:

$$\left.\begin{array}{l}\frac{1}{2}\gamma^T Q\gamma - \sum_{i=1}^{m+k}\gamma_i \to \min \\ \text{subject to} \\ \sum_{i=1}^{m+k} t(x_i)\gamma_i = 0, \\ \gamma_i \geq 0, \forall i = 1,2,\ldots,m+k.\end{array}\right\} \quad (8)$$

The vectors a^i, b^i *which* $\gamma_i > 0$ are called support vectors.

It is noted that the problems (7) and (8) give one and the same results. In the following we will show how the problem (7), which is equivalent to the problem (8) can be reduced to solving a quadratic equation system.

If $\gamma^* \in R^{m+k}$ is an optimal solution of the problem (7) then there is $\lambda^* \in R^2$ such that [8]:

$$\left.\begin{array}{l}B\gamma^* = e, \\ \Gamma^*\left(Q\gamma^* + B^T\lambda^*\right) = 0, \\ \gamma_i^* = 0 \Rightarrow \left[Q\gamma^* + B^T\lambda^*\right]_i \geq 0, \forall i.\end{array}\right\}$$

where $\Gamma = Diag(\gamma)$ is a diagonal matrix:

$$\Gamma = \begin{pmatrix}\gamma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \gamma_{m+k}\end{pmatrix},$$

and the notation $[c]_i$ here and further mean the component i of the vector c.

Note the $G = Diag\left(Q\gamma + B^T\lambda\right)$ and

$$F(\gamma,\lambda) = \begin{pmatrix}\Gamma\left(Q\gamma + B^T\lambda\right) \\ B\gamma - e\end{pmatrix}.$$

Such the problem (7) is reduced to solve the system of equations and inequalities:

$$\left.\begin{array}{l}F(\gamma,\lambda) = 0, \\ \gamma \geq 0, G \geq 0.\end{array}\right\}$$

Theorem. ***For*** $\forall\gamma \in D = \{\gamma : B\gamma = e, \gamma \geq 0\}$ ***the Jacobian matrix***

$$\nabla F = \begin{pmatrix}G + \Gamma Q & \Gamma B^T \\ B & 0\end{pmatrix}$$

is non-singular.

Demonstration of this theorem is analogous of proof the **Theorem 1** from [9].

We define now the functions $p, q : R \to R_+$:

$$p(x) = x^2 \max(0,x) = \frac{1}{2}\left(x^3 + |x|x^2\right),$$

$$q(x) = -x^2 \min(0,x) = -\frac{1}{2}\left(x^3 - |x|x^2\right).$$

The functions $p(x)$ and $q(x) = p(-x)$ are twice continuously differentiable:

$$p'(x) = \frac{3}{2}\left(x^2 + |x|x\right), p''(x) = 3\left(x + |x|\right),$$

$$q'(x) = \frac{3}{2}\left(-x^2 + |x|x\right), q''(x) = 3\left(-x + |x|\right).$$

It is easily established that:

1. $\begin{cases}p(x)q(x) = 0, p'(x)q'(x) = 0, \\ p''(x)q''(x) = 0, \forall x \in R.\end{cases}$

2. $\begin{cases}p(x) + q(x) > 0, p'(x) + q'(x) \neq 0, \\ p''(x) + q''(x) > 0, \forall x \neq 0.\end{cases}$

3. $\begin{cases}p(x) = 0 \text{ if and only if } p'(x) = 0, \\ q(x) = 0 \text{ if and only if } q'(x) = 0.\end{cases}$

Considering this and introducing auxiliary variables $\eta_1, \eta_2, \ldots, \eta_m$, $\mu_1, \mu_2, \ldots, \mu_k$, λ_1 and λ_2 the problem (7) can be reduced to solve a system of $2(m + k + 1)$ equations with same number of unknowns [9]:

$$\left.\begin{array}{l} p(\eta_i) = \alpha_i . i = 1,2,\ldots,m, \\ q(\eta_i) = \left[U_{m\times m}\alpha - Z_{m\times k}\beta + \lambda^1\right]_i, \\ i = 1,2,\ldots,m, \\ p(\mu_i) = \beta_i, i = 1,2,\ldots,k, \\ q(\mu_i) = \left[-Z_{m\times k}\alpha + V_{k\times k}\beta + \lambda^2\right]_i, \\ i = 1,2,\ldots,k, \\ B\gamma = e. \end{array}\right\} \quad (9)$$

This was noted as follow:

$$\lambda^1 = (\lambda_1, \lambda_1, \ldots, \lambda_1)^T \in R^m,$$
$$\lambda^2 = (\lambda_2, \lambda_2, \ldots, \lambda_2)^T \in R^k.$$

The system (9) may be reduced to the $(m+n+2)$ equations with the $(m+n+2)$ unknowns, replacing the vectors α and β using functions p and q :

$$\alpha = (p(\eta_1), p(\eta_2), \ldots, p(\eta_m))^T,$$
$$\beta = (p(\mu_1), p(\mu_2), \ldots, p(\mu_k))^T.$$

The best-known method for solving nonlinear systems of equations (9) is the Newton method [10]. Newton's method has very attractive theoretical and practical properties, because of its fast convergence: under the nonsingularity of the Jacobian matrix it will converge locally superlinearly.

Let be

$$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_m^*)^T$$

and

$$\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_k^*)^T$$

the optimal solution of problem (8). When the decision function is given by:

$$f(x) = \frac{2}{\left\|x^* - y^*\right\|^2}\left[(x^* - y^*)x - \left\|x^*\right\|^2 + \left\|y^*\right\|^2\right],$$

where

$$x^* = \sum_{i=1}^{m} \alpha_i^* a^i,\ y^* = \sum_{i=1}^{k} \beta_i^* b^i.$$

5. CONCLUSIONS

In this paper we presented an overview of mathematical problem separating two data sets. The classification methods are based on search for an optimal hyperplane which separates the considered data. A special place is occupied by SVM introduced in 1995 by Vladimir Vapnik and discussed in the literature by many researchers [11].

Here was introduced reformulation of Kuhn-Tucker optimality conditions in an equivalent system of smooth linear equations (cubic). The system of equations can be solved efficiently using Newton method. In the neighborhood of the optimal solution γ^* the rate of convergence of Newton's sequence is superlinear. The numerical examples clearly show that the proposed method is promising.

Bibliography

1. ***Mangasarian, O. L.*** *Linear and nonlinear separation of patterns by linear programming. Operations Research, Vol. 13, Nr. 3, pag. 434-452, 1965.*

2. ***Viveros, M. S., Nearhos, J. P.,*** *and* ***Rothman, M. J.***, *Applying Data Mining Techniques to a Health Insurance Information System. Proc. of the 22th International Conference on Very Large Data Bases (VLDB '96), pag. 286-294, 1996.*

3. ***Mangasarian, O. L., Street, W. N.*** *and* ***Wolberg, W. H.*** *Breast Cancer Diagnosis and Prognosis via Linear Programming, Operations Research, Vol. 43, No. 4, pag. 570-577, 1995.*

4. ***Ferreira, C.*** *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Springer Verlag, 478 pag., 2006.*

5. ***Merrikh-Bayat, F., Babaie-Zadeh, M.,*** *and* ***Jutten, C.*** *Linear-quadratic blind source separating structure for removing show-through in scanned documents, International Journal on Document Analysis and Recognition (IJDAR), Vol. 14, No.4, pag. 319-333, 2011.*

6. ***Freund, R. M.*** *Pattern Classification, and Quadratic Problems, Massachusetts Institute of Technology, 2004.*

7. ***Vapnik, V.*** *The nature of statistical learning theory, Springer-Verlag, New York, 1995.*

8. ***Gill, P. E., Murray, W.,Wright, M. H***. *Practical Optimization, Academic Press, London,1981.*

9. ***Moraru, V.,*** *A Smooth Newton Method for Nonlinear Programming Problems with Inequality Constraint, Computer Science Journal of Moldova, Vol. 19, Nr. 3 (57), pag. 333-353, 2011.*

10. ***Ypma, T. J.*** *Historical development of the Newton-Raphson method, SIAM Review, Vol. 37, No. 4, pag. 531-551, 1995.*

11. ***Byun, H., Lee, S. W.,*** *Applications of Support Vector Machines for Pattern Recognition: A Survey, Pattern Recognition with Support Vector Machines, Vol. 2388, pag. 213-236, 2002.*

Recommended for publication: 11.03.2013.