

## MSV

- ) Modelo lineal de regresión/clasificación de la fase del aprendizaje estadístico
- ) En clasificación binaria tenemos  $\mathcal{L} = \{(x_i, \hat{y}_i), \dots, (x_m, \hat{y}_m)\}$  conjunto de datos con  $x_i \in \mathbb{R}^n, \hat{y}_i \in \{-1, 1\} \quad \forall i=1, \dots, m$  la MSV busca el hiperplano que clasifique los datos en  $\mathcal{L}$  (separar datos etiquetados con "1" de los de "-1") de modo que se maximice el margen que rodea al hiperplano y se minimice los errores de clasificación
- ) Modelo en una MSV:

$$y(x|w_0, w) = w^T x + w_0 \text{ con } w, x \in \mathbb{R}^n, w_0 \in \mathbb{R}$$

Se pretende realizar  $m$  clasificaciones de acuerdo a la regla:

$$\operatorname{signo}(y(x_i|w_0, w)) = \begin{cases} 1 & \text{si } y(x_i|w_0, w) \geq 0 \\ -1 & \text{e.o.c.} \end{cases} \quad \forall i=1, \dots, m$$

$$\text{Notación: } y_i = y(x_i|w_0, w) \quad \forall i=1, \dots, m$$

## Datos linearmente separables

Si los  $m$  datos son separables, se tiene:  $\hat{y}_i y_i > 0 \quad \forall i=1, 2, \dots, m$  y existe

$$\mathcal{P} = \{x \in \mathbb{R}^n \mid y(x|w_0, w) = 0\} \text{ de dimensión } n-1 \text{ (infinitos P)}$$

Para el margen denotado como  $M_{w_0, w}$  calculamos:  $\min_{i \in \{1, \dots, m\}} \{d(\mathcal{P}, x_i)\}$

$$\text{como } d(\mathcal{P}, x_i) = \frac{|y(x_i|w_0, w)|}{\|w\|_2} = \frac{\hat{y}_i y_i}{\|w\|_2} \xrightarrow{\text{por def. de } \hat{y}_i \text{ y } y_i} \forall i=1, \dots, m$$

$$\therefore \min_{i \in \{1, \dots, m\}} \{d(\mathcal{P}, x_i)\} = \min_{i \in \{1, \dots, m\}} \left\{ \frac{\hat{y}_i y_i}{\|w\|_2} \right\}$$

Por la suposición de datos linealmente separables se tiene

$$\hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

$$\min_{\|w\|=1} \left\{ d(R, x_i) \right\} = \min_{i \in \{1, \dots, m\}} \left\{ \frac{1}{\|w\|_2} \right\} \text{ sujeto a: } \hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

$$\text{Consideramos } R_1 = \{x \in \mathbb{R}^n \mid y(x|w_0, w) = 1\}$$

$$R_{-1} = \{x \in \mathbb{R}^n \mid y(x|w_0, w) = -1\}$$

$$\text{Entonces } d(R^*, R_1) = d(R^*, R_{-1}) = \frac{1}{\|w\|_2} \quad \text{⊗}$$

En la MSV  $R^*$  es el hiperplano que satisface:

$$R^* = \max_{w_0, w} \{ M_{w_0, w} \}$$

Por  $\text{⊗}$  el margen,  $M_{w_0, w}$  es  $\frac{2}{\|w\|_2}$  y el problema en SVM es:

$$\max_{w_0, w} \left\{ \frac{2}{\|w\|_2} \right\} \text{ s.a. } \hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

$$\text{que es equivalente a: } \min_{w_0, w} \left\{ \frac{\|w\|_2^2}{2} \right\} \text{ s.a. } \hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

$$\text{.....} \quad \min_{w_0, w} \left\{ \frac{\|w\|_2^2}{2} \right\} \text{ s.a. } \hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

Obs

- ) Por la sup. de datos linealmente separables  $\exists R^*$  solución del problema en SVM
- ) Al conjunto de restricciones en las que se rompe la igualdad ( $\hat{y}_i y_i = 1$ ) se les llaman activas

- El problema en SVM es un problema cuadrático
- El problema en SVM involucra restricciones de desigualdad y los métodos que resuelven este tipo de problemas buscan aquellas restricciones que son activas en la solución mediante un proceso iterativo. Otros métodos son los de puntos interiores (PI)
  - Desventaja al usar PI para resolver los problemas de optimización en MSV:

La llamada formulación primal:

$$\min_{(w, b) \in \mathbb{R}^{n+1}} \left\{ \frac{\|w\|^2}{2} \right\} \rightarrow \text{optimización sobre } n+1 \text{ variables}$$

s.a.  $\hat{y}_i^T w + b \geq 1 \quad \forall i = 1, \dots, m$

escribir con el número de restricciones ( $m$ )

La formulación dual (más adelante su derivación):

$$\min_{\lambda \in \mathbb{R}^m} \frac{1}{2} \lambda^T \hat{Y} X^T X \hat{Y} \lambda - \lambda^T e \rightarrow \text{optimización sobre } m \text{ variables}$$

s.a.  $e^T \hat{Y} \lambda = 0$   
 $\lambda \geq 0$

tiene una matriz Hessiana densa:  $\hat{Y} X^T X \hat{Y}$ , construir  $X^T X$  son  $\mathcal{O}(m^2 n)$  operaciones que es lo más costoso en el producto  $\hat{Y} X^T X \hat{Y}$  pues  $\hat{Y}$  es diagonal si se utilizan PI para la formulación primal o dual el costo computacional es  $\mathcal{O}(m^3)$  por ello se han preferido otros métodos

(ver Libro de Nocedal-Wright para solución de programas cuadráticos)

## Formulación dual:

Consideremos la formulación primal de MSV  $\textcircled{X}$ :

$$\begin{array}{c} \min_{(w, z) \in \mathbb{R}^{n+1}} \\ \left\{ \frac{\|w\|^2}{2} \right\} \\ \text{s.a. } \hat{y}_i^T w \geq 1 \quad \forall i=1..m \end{array}$$

y la función Lagrangiana:

$$L(z, \lambda) = \frac{1}{2} z^T z - \lambda^T (Az - e) \text{ con } \lambda \geq 0$$

$$z = \begin{bmatrix} w \\ w_0 \end{bmatrix}, A = \begin{bmatrix} \hat{y} & x^T & \hat{y}_e \end{bmatrix}, e = (1)_i^m, \quad i=1..m$$

$$x = [x_1 \dots x_m] \in \mathbb{R}^{n \times m}, \quad \hat{y} = \text{diag}(\hat{y}_i)_{i=1}^m$$

Las condiciones de optimalidad (Karush-Kuhn-Tucker) para el problema anterior son:

$$\nabla_z L(z, \lambda) = z - A^T \lambda = 0 \quad (1)$$

$$A z - e \geq 0$$

$$\lambda \geq 0$$

$$\lambda^T (A z - e) = 0 \quad (\text{complementariedad})$$

De (1) considerando la asignación de variables anterior, se tiene:

$$w = x^T \lambda$$

$$e^T \lambda = 0$$

$$\text{por lo que } L(z, \lambda) = \frac{1}{2} \lambda^T \hat{y}^T x^T x \hat{y} \lambda - \lambda^T (A z - e)$$

$$= -\frac{1}{2} \lambda^T \hat{y}^T x^T x \hat{y} \lambda + \lambda^T e$$

$$\begin{array}{c} \min_{\lambda \in \mathbb{R}^m} \\ \frac{1}{2} \lambda^T \hat{y}^T x^T x \hat{y} \lambda - \lambda^T e \\ \text{s.a. } e^T \hat{y} \lambda = 0 \\ \lambda \geq 0 \end{array}$$

∴ El problema dual de  $\textcircled{X}$  es:

.) Estimamos  $w$  a partir de  $w = x \hat{\lambda}$

.) Para estimar  $w_0$ :

Sea  $S$  el conjunto de restricciones activas en  $Az - b \geq 0$   
 con  $\lambda_i > 0 \forall i \in S$  (por complementariedad  $\lambda_i > 0$ )

Se tiene:

$$\hat{y}_i y_{i-1} = \hat{y}_i (w^T x_i + w_0) - 1 = 0 \text{ para } i \in S$$

$$\therefore y_i - \hat{y}_i = w^T x_i + w_0 - \hat{y}_i = 0 \text{ pues } \hat{y}_i^2 = 1$$

$$\begin{aligned} \text{Por otro lado: } y(x|w_0, w) &= w^T x + w_0 \\ &= \lambda^T \hat{y} X^T x + w_0 \text{ por (1)} \\ &= \sum_{i=1}^m (\lambda_i \hat{y}_i x_i^T x) + w_0 \\ &= \sum_{i \in S} (\lambda_i \hat{y}_i x_i^T x) + w_0 \end{aligned}$$

$$\therefore \hat{w}_0 = \frac{1}{N_S} \sum_{j \in S} \left( \hat{y}_j - \sum_{i \in S} \lambda_i \hat{y}_i x_i^T x_j \right)$$

con  $N_S$  número de vectores de soporte

.) Los vectores de soporte contribuyen a la clasificación de  $y_j, \forall j=1, \dots, m$

$$\text{pues: } y_j = y(x_j | \hat{w}_0, \hat{w})$$

$$= \hat{w}^T x_j + \hat{w}_0$$

$$= \sum_{i \in S} (\lambda_i \hat{y}_i x_j^T x_i) + \hat{w}_0$$

y se clasifica de acuerdo al signo de  $y_j$

## Datos no linearmente separables

Sea  $\{z_i\}_{i=1..m}$  el problema primal de MSV es:

$$\min \frac{1}{2} \|w\|^2 + C e^T z$$

$$\text{s.a. } z_i > 0 \quad \text{(4)}$$

$$y_i \hat{y}_i \geq 1 - z_i \quad \forall i = 1, \dots, m$$

$$\text{donde } z = (z_i)_{i=1}^m, C > 0$$

Obs

•  $z_i$  son variables que permiten datos mal clasificados:

+  $z_i = 0$  representa el dato "i" bien clasificado

+  $z_i \in (0, 1]$  " " " " " pero dentro del margen

+  $z_i > 1$  dato mal clasificado

•  $C e^T z$  es una rotación superior de los mal clasificados y  $C$  ayuda a controlar el número máximo. También, se agrega al modelo para controlar un mal ajuste o sobreajuste del modelo a los datos

Formulación Dual Sea la función Lagrangiana:

$$\mathcal{L}(z, \lambda) = \frac{1}{2} z^T H z + d^T z - \lambda^T (A z - g) \text{ con } \lambda > 0$$

$$z = \begin{bmatrix} w \\ N_0 \\ z \end{bmatrix}, H = \begin{bmatrix} I_m & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, d = \begin{bmatrix} 0 \\ 0 \\ C e \end{bmatrix}$$

$$g = \begin{bmatrix} e \\ 0 \end{bmatrix}, A = \begin{bmatrix} \hat{Y} X^T & \hat{Y} e & I_m \\ 0 & 0 & I_m \end{bmatrix}, \lambda = \begin{bmatrix} \lambda \\ \gamma \end{bmatrix}$$

Donde:  $\lambda_g$  corresponde a  $\hat{y}_i y_i \geq 1 - \xi_i$   
 $\lambda_\xi$  " " "  $\xi_i \geq 0$

Las condiciones de optimidad (Karush-Kuhn-Tucker) para el problema anterior son:

$$\nabla_z \mathcal{L}(z, \lambda) = H_z + d - A^T \lambda = 0 \quad (1)$$

$$\begin{aligned} A_z - g &> 0 \\ \lambda &> 0 \\ \lambda^T (A_z - g) &= 0 \text{ (complementariedad)} \end{aligned}$$

De (1) se obtiene:  $w = X^T \lambda_g$

$$e^T \lambda_g = 0$$

$$C e - \lambda_g - \lambda_\xi = 0$$

Como  $\lambda > 0$  de la última ecuación se tiene:  $0 \leq \lambda_g \leq C$

y también:  $C e^T \xi = \lambda_g^T \xi$  por complementariedad  
La función lagrangiana se rescribe:

$$\begin{aligned} \mathcal{L}(z, \lambda) &= \frac{1}{2} \lambda_g^T X^T X \lambda_g + C e^T \xi - \lambda^T (A z - g) \\ &= -\frac{1}{2} \lambda_g^T X^T X \lambda_g + \lambda^T e \end{aligned}$$

y el problema dual es:

$$\min \frac{1}{2} \lambda_g^T X^T X \lambda_g - \lambda^T e$$

$$\text{s.a. } C^T \lambda_g = 0$$

$$0 \leq \lambda_g \leq C$$

Obs

- ) Estimamos  $W$  a partir de  $W = \hat{y}^T \lambda_g$
- ) La estimación de  $W_0$  es similar al desarrollo de datos linealmente separables:

$$\hat{W}_0 = \frac{1}{N_M} \sum_{K \in M} \left( \hat{y}_K - \sum_{i \in S} (\lambda_g(i) \hat{y}_i x_i^T x_K) \right)$$

donde:  $\lambda_g(i)$  es la  $i$ -ésima entrada de  $\lambda_g$

$M$  es el conjunto de datos que cumplen:  $0 < \lambda_g(K) < C$  y

$N_M$  es  $\#(M)$

- ) Los vectores de soporte contribuyen a la clasificación de  $y_j, \forall j=1..,m$

pus:

$$\begin{aligned} y_j &= y(x_j | \hat{w}_0, \hat{w}) \\ &= \hat{w}^T x_j + \hat{w}_0 \\ &= \sum_{i \in S} (\lambda_g(i) \hat{y}_i x_i^T x_j) + \hat{w}_0 \end{aligned}$$

y se clasifica de acuerdo al signo de  $y_j$

Puntos interiores con función de barrera logarítmica

Consideremos la formulación dual de la MSV.

$$\min \frac{1}{2} \lambda_g^T \hat{y}^T x^T x \hat{y} \lambda_g - \lambda_g^T e$$

$$\text{s.a. } e^T \hat{y} \lambda_g = 0$$

$$0 \leq \lambda_g \leq C$$

y hacemos las siguientes asignaciones de variables.

$$y = w$$

$$x = \lambda$$

$$A = X \hat{y}$$

$$b = \hat{y} e$$

Tenemos:

$$\min_{x,y} \frac{y^T y}{2} - e^T x$$

s.t.

$$b^T x = 0$$

$$Ax - y = 0$$

$$\begin{aligned} x - Ce &\geq 0 \\ x &\geq 0 \end{aligned}$$

Utilizando la función de barrera logarítmica reescribimos el problema anterior como:

$$\min_{x,y} \frac{y^T y}{2} - e^T x - \mu \left[ \sum_{i=1}^m \log(x_i) + \log(c_i - x_i) \right]$$

s.a.

$$A_y y + A_x x = 0$$

dónde

$$A_y = \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad A_x = \begin{bmatrix} -A \\ -b^T \end{bmatrix}$$

La función Lagrangiana es:

$$L(x, y, \lambda) = \frac{1}{2} y^T y - e^T x - \mu \left[ \sum_{i=1}^m \log x_i + \log(c_i - x_i) \right] - \lambda^T (A_y y + A_x x)$$

$L(x, y, \lambda)$  es una función convexa por lo que las condiciones de KKT son necesarias y suficientes:

$$\nabla L(x, y, \lambda) = \begin{bmatrix} y - A_y^T \lambda \\ -e - \mu (X^{-1} - (I - X)^{-1}) e - A_x^T \lambda \\ A_y y + A_x x \end{bmatrix} = 0$$

con  $X = \text{diag}(x_i)_{i=1}^m$ ,  $U = \text{diag}(c e_i)_{i=1}^m$

Si definimos  $s = M X^{-1} e$ ,  $v = M(U-X)^{-1} e$  tenemos:

$$\nabla \mathcal{L}(x, y, \lambda, s, v) = \begin{bmatrix} y - A_y^T \lambda \\ -e - s + v - A_x^T \lambda \\ A_y y + A_x x \\ X s - M e \\ (U - X) v - M e \end{bmatrix} = 0$$

Para resolver las ecuaciones anteriores utilizamos el método de Newton y encontrar el vector d del siguiente sistema

$$J(x, y, \lambda, s, v)d = -F(x, y, \lambda, s, v)$$

$$\text{con } F(x, y, \lambda, s, v) = \nabla \mathcal{L}(x, y, \lambda, s, v)$$

y J jacobiana de F

Escribiendo el sistema de manera explícita:

$$\begin{bmatrix} I_n & 0 & -A_y^T & 0 & 0 \\ 0 & 0 & -A_x^T & -I_m & I_m \\ 0 & S & 0 & \Sigma & 0 \\ 0 & -V & 0 & 0 & U - X \\ -A_y & -A_x & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} dy \\ dx \\ d\lambda \\ ds \\ dv \end{bmatrix} = - \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{bmatrix}$$

$$\text{con } S = \text{diag}(s_i)_{i=1}^m, V = \text{diag}(v_i)_{i=1}^m$$

Utilizando los bloques  $F_4, F_5$  del sistema anterior se tiene:

$$ds = -X^{-1}(F_3 + S dx)$$

$$d_V = -(U - X)^{-1}(F_4 - Vd_X)$$

y se pueden eliminar  $d_S, d_V$  resultando en el sistema:

$$Kd = \begin{bmatrix} I_n & 0 & -A_y^T \\ 0 & \textcircled{4}_x^{-1} & -A_x^T \\ -A_y & -A_x & 0 \end{bmatrix} \begin{bmatrix} d_y \\ d_x \\ d_\lambda \end{bmatrix} = -\begin{bmatrix} r_{C_y} \\ r_{C_x} \\ r_b \end{bmatrix}$$

donde:  $\textcircled{4}_x^{-1} = X^{-1}S + (U - X)V$

$$r_{C_y} = F_1$$

$$r_{C_{x1}} = V - A_x^T \lambda - e - s$$

$$r_{C_x} = r_{C_{x1}} + X^{-1}F_3 - (U - X)^{-1}F_4$$

$$r_b = F_3$$

Obs o) El sistema anterior es simétrico de dimensión  $2n+m+1$

y tiene una solución única que se obtiene haciendo  $\mu \rightarrow 0$

Este último se realiza en cada iteración con la actualización

$$\mu = \frac{x^T s + ((e - x)^T v)}{2m}$$

o) Si  $K = \begin{bmatrix} H & -\hat{A}^T \\ -\hat{A} & 0 \end{bmatrix}$  con  $H = \begin{bmatrix} I_n & 0 \\ 0 & \textcircled{4}_x^{-1} \end{bmatrix}, \hat{A} = [A_y \quad A_x]$

entonces  $H$  es diagonal y podemos partir a  $K$  como:

$$K = \begin{bmatrix} H_1 & A_1^T \\ H_2 & A_2^T \\ \vdots & \vdots \\ \cdots H_p & A_p^T \\ A_1 \quad A_2 \quad \cdots A_p & 0 \end{bmatrix} \quad \text{donde: } H_i, A_i \text{ resultan de}$$

partir a  $H, -\hat{A}$  de igual forma entre  $p$  procesadores.  $H_i$  es diagonal  $H_i = I, \dots, p$

Las entradas que no se identifican son iguales a cero

Si  $A$  es una matriz indefinida, posee una factorización de la forma:

$$K = LDL^T$$

También entraña a entradas se tiene:

$$H_i = L_i D_i L_i^T \quad \therefore D_i = H_i, L_i = I$$

$$L_{A_i} = A_i L_A^{-1} D_i^{-1} = A_i + I_i^{-1} \quad \dots \quad . \quad . \quad . \quad (1)$$

Si a la suma  $-\sum_{i=1}^p \Delta_i H_i^{-1} \tilde{A}_i^T$  la definimos como a la

matriz C entradas

$$C = L_c D_c L_c^T \quad \dots \quad (2)$$

## Trabajo en párrafo:

(1) y los productos  $A_i H_i^{-1} A_i^T$  se calculan en cada procesador

En un solo procesador se calcula  $C_y(?)$

Una vez que se tiene la representación  $K = LDL^T$  resolvemos el sistema:

$$LDL^T \begin{bmatrix} dz \\ d\lambda \end{bmatrix} = - \begin{bmatrix} r_c \\ r_b \end{bmatrix} \text{ con } z = \begin{bmatrix} y \\ x \end{bmatrix}, r_c = \begin{bmatrix} rc_y \\ rc_x \end{bmatrix}$$

con L triangular inferior, D diagonal

$$d\lambda'' = L_c^{-1} \left( r_b - \sum_{i=1}^p L_{Ai} r_{ci} \right) \dots (1)$$

$$d\lambda' = D_c^{-1} d\lambda'' \dots (2)$$

$$d\lambda = L_c^{-T} d\lambda' \dots (3)$$

$$dz_i' = D_i^{-1} r_{ci} \dots (4)$$

$$dz_i = dz_i' - L_{Ai}^T d\lambda \dots (5)$$

Los vectores  $d\tau$ ,  $d\lambda'$ ,  $d\lambda''$  se utilizan para cálculos intermedios

Trabajo en paralelo:

El producto  $L_{Ai} r_{ci}$  de (1) se realiza en cada procesador

El procesador que tiene almacenado  $L_c, D_c$  realiza (1), (2), (3) para calcular  $d\lambda'', d\lambda', d\lambda$  y distribuye  $d\lambda$  a cada procesador

El cálculo de  $dz_i'$ ,  $dz_i$  se realiza en cada procesador

El procesador que tiene almacenado  $d\lambda$  reúne  $d\tau$

Obs  $\Rightarrow$  Las variables  $x, u-x, s, v$  están sujetas a la restricción de no negatividad. Por ello las reglas de actualización son:

$$x^{(k+1)} = x^{(k)} + \alpha dx \quad \text{donde } \alpha = 0.995 \min(\alpha_x, \alpha_{u-x}, \alpha_s, \alpha_v)$$

$$y^{(k+1)} = y^{(k)} + \alpha dy$$

$$\lambda^{(k+1)} = \lambda^{(k)} + \alpha d\lambda$$

$$s^{(k+1)} = s^{(k)} + \alpha ds$$

$$v^{(k+1)} = v^{(k)} + \alpha dv$$

$$\text{con } \alpha_x = \min \left\{ \frac{-x_i^{(k)}}{dx(i)}, : dx(i) < 0 \right\}$$

análogo para  $\alpha_s, \alpha_v$  y

$$\alpha_{u-x} = \min \left\{ \frac{(c_{xi} - x_i)^{(k)}}{dx(i)}, : dx(i) > 0 \right\}$$

con  $d_{x(i)}$  i-ésima componente del vector  $d_x$  y  $x^{(k)}$  vector en  
la k-ésima iteración