

MSV

-) Modelo lineal de regresión/clasificación del marco 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

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

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

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

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

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

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

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

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

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

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

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

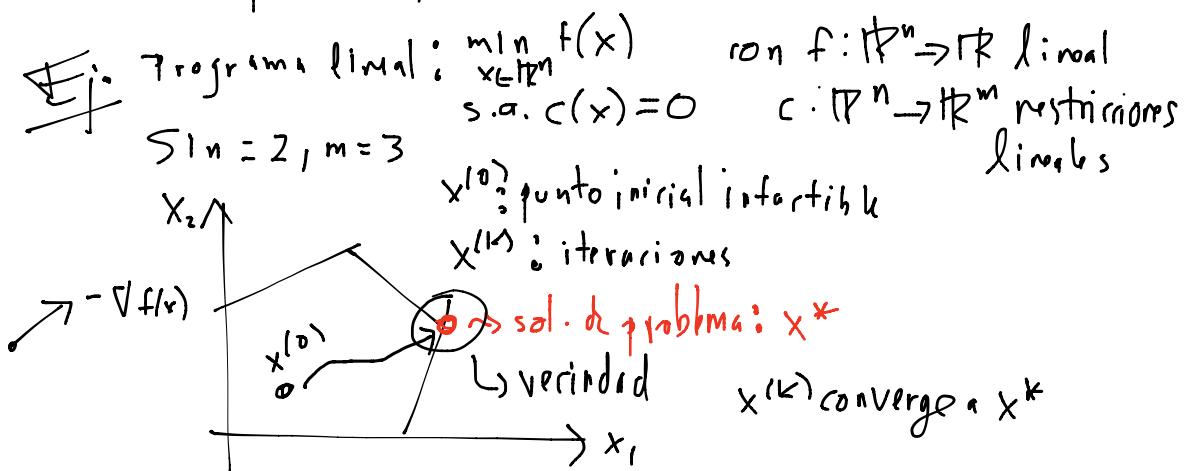
$$\text{que es equivalente a: } \min_{w_0, w} \left\{ \frac{\|w\|_2^2}{2} \right\} \text{ s.a. } \hat{y}_i y_i \geq 1 \quad \forall i = 1, \dots, m$$

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

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 se conoce como un problema cuadrático:
 -) optimizar una función objetivo cuadrática sujeta a restricciones lineales
 -) Si en un problema cuadrático la Hessiana de la función objetivo es positivo definido o semidefinida el problema cuadrático se llama convexo
 -) Si el problema cuadrático es convexo entonces
 -) El mínimo local es un mínimo global
 -) Las condiciones de optimalidad (KKT) son necesarias y suficientes
 -) 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.
 -) Puntos interiores (PI):
 - >> Son una clase de métodos que resuelven programas lineales y cuadráticos perturbando las condiciones de optimalidad relacionadas con la complementariedad a través de barreras logarítmicas y mantienen la división entre restricciones activas/inactivas el mayor número de iteraciones posibles hasta convergencia
 - >> Entre los métodos por PI los llamados primales-dobles son los que han recibido mayor atención por su eficiencia en tiempo de ejecución y otras características



v) Complejidad polinomial en problemas lineales:

Ventajas de PI:

- es un método que compite con simplex (complejidad exponencial)
- v) Convergencia a partir de puntos iniciales infactibles dentro/fuera de la región

v) A medida que el tamaño del problema crece (en número de variables) el número de iteraciones crece lentamente

Desventajas de PI

d) Cálculo de la siguiente iteración es costoso computacionalmente hablando (resolver un sistema de ecuaciones lineales)

>) En el contexto de la SVM:

>>) La llamada formulación primal:

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

$$\text{s.a. } \hat{y}_i^T y_i \geq 1 \quad \forall i = 1 \dots m$$

escribir con el número de restricciones (m)

>>) La formulación dual (más adelante su derivación):

$$\begin{aligned} \min_{\lambda \in \mathbb{R}^m} & \frac{1}{2} \lambda^T \hat{Y} X^T X \hat{Y} \lambda - \lambda^T e \\ \text{s.a.} & \hat{e}^T \hat{Y} \lambda = 0 \\ & \lambda \geq 0 \end{aligned} \rightarrow \text{optimización sobre } m \text{ variables}$$

tiene una matriz Hessiana densa: $\hat{Y} X^T X \hat{Y}$, construir $X^T X$ son $O(m^2 n)$ operaciones que es lo más costoso en el producto $\hat{Y} X^T X \hat{Y}$ pues \hat{Y} es diagonal

d) 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)

En lo siguiente suponemos $m \gg n$

