Introducción a las Máquinas de Soporte Vectorial

Estimación de funciones

El problema general que intentaremos abordar aquí es el siguiente:

Mediante algún proceso observacional/experimental se colectaron un conjunto de pares de datos (nuestra base de hechos) $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ donde $\boldsymbol{x}_i = \{x_i^1, \ldots., x_i^p\} \in \mathcal{R}^p$ es un vector **aleatorio** de entradas del sistema bajo estudio e y_i la salida, univariada, correspondiente a dicha entrada. Denominaremos a este conjunto *F* como "conjunto de entrenamiento" o nuestra base de hechos (Base de datos).

El objetivo es intentar entender y modelar el sistema que, ante una entrada $\boldsymbol{x}_i$ produce una salida y_i, esto quiere decir que existe una *f* tal que $y_i = f(\boldsymbol{x}_i)$, pero esta función *f* es desconocida. Esto implica encontrar una función $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta})$, con $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$ (espacio de parámetros) de manera tal que minimizamos el siguiente funcional

$$R(\boldsymbol{\beta}) = \int L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right) d\boldsymbol{F}(\boldsymbol{x}, y) \quad (1.1)$$

Donde L() se denomina función de pérdida y dependiendo del tipo de variable de respuesta (y_i) podemos definir al menos dos tipos básicos de problemas a resolver:

1. Reconocimiento de patrones (Clasificación): En este caso la salida del sistema desconocido puede tomar solo 2 valores $y_i = \{0,1\}$ y sea $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta})$, con $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$ un conjunto de funciones "indicadoras" que solo puede tomar los valores 0 y 1. Consideremos la siguiente función de pérdida:
   $$L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right) = \begin{cases} 0 \; si \; y = \hat{f}(\boldsymbol{x}, \boldsymbol{\beta}) \\ 1 \; si \; y \neq \hat{f}(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases} \quad (1.2)$$
   Para esta función de pérdida, el función 1.1 determina la probabilidad de error de clasificación.
2. Regresión (Predicción, aproximación de funciones)
   Sea $y_i \in \mathcal{R}$ y sea $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta})$ un conjunto de funciones reales que contiene la siguiente función de regresión $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta_0}) = \int y dF(y|\boldsymbol{x})$
   Se sabe que dicha función de regresión es la que minimiza el funcional 1.1 con la siguiente función de pérdida $L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right) = \left(y - \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right)^2 \quad (1.3)$

En términos generales podemos describir el problema de la siguiente manera: Sea la medida de probabilidad $\boldsymbol{F}(\boldsymbol{z})$ definida en el espacio $\boldsymbol{Z}$. Considere el conjunto de funciones $Q(\boldsymbol{z}, \boldsymbol{\beta})$ con $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$. El objetivo es minimizar el funcional de riesgo $R(\boldsymbol{\beta}) = \int Q(\boldsymbol{z}, \boldsymbol{\beta}) d\boldsymbol{F}(\boldsymbol{z})$ con $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$. donde $\boldsymbol{F}(\boldsymbol{z})$ es desconocida pero contamos con un conjunto de muestras i.i.d $\mathbf{z_1 \ldots . z_n}$.

Minimización del riesgo empírico (MRE)

Con el objeto de minimizar el funcional 1.3 bajo una función de distribución $\boldsymbol{F(z)}$ desconocida, podemos aplicar el siguiente principio inductivo:

i. El funcional de riesgo $R(\boldsymbol{\beta})$ se reemplaza por el funcional de riesgo empírico $R_{emp}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} Q(\boldsymbol{z_i}, \boldsymbol{\beta})$ (1.4) construido en base al conjunto de entrenamiento *F*.
ii. Se aproxima la función $Q(\boldsymbol{z}, \boldsymbol{\beta_0})$ que minimiza el riesgo 1.1 mediante la función $Q(\boldsymbol{z}, \boldsymbol{\beta})$ que minimiza el riesgo empírico.

Este concepto, MRE, es bastante general pudiendo considerarse a los métodos de mínimos cuadrados (MC) y de máxima verosimilitud (MV) realizaciones del MRE bajo funciones de pérdida específicas (la 1.2) en el primer caso en $L\big(p(\boldsymbol{x}, \beta)\big) = -\log\big(p(\boldsymbol{x}, \beta)\big)$. Con "p" función de densidad de probabilidad (conocida en el caso de MV)

Operar mediante el principio de MRE tiene la ventaja de que:

1. No depende de $\boldsymbol{F(z)}$
2. En teoría es posible minimizar 1.4 con respecto a $\boldsymbol{\beta}$

Sea $\widehat{\boldsymbol{\beta}}$ el vector de parámetros que minimiza $R_{emp}(\boldsymbol{\beta})$ a través de $\hat{f}\big(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}\big)$ y sea $\boldsymbol{\beta_0}$ el vector de parámetros que minimiza a $R(\boldsymbol{\beta})$ a través de $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta_0})$ con $\boldsymbol{\beta_0}$ **y** $\widehat{\boldsymbol{\beta}} \in \boldsymbol{\mathcal{B}}$ entonces:

¿Bajo qué condiciones podemos decir que $\hat{f}\big(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}\big)$ esta "cerca" de la función deseada $\hat{f}(\boldsymbol{x}, \boldsymbol{\beta_0})$ medida a través de la discrepancia entre $R_{emp}(\boldsymbol{\beta})$ y $R(\boldsymbol{\beta})$?

Dado $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ "fijo" **,** el riesgo funcional $R(\boldsymbol{\beta}^*)$ determina la "Esperanza Matemática" de una variable aleatoria $Q(\beta^*) = L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right)$, mientras que el riesgo empírico $R_{emp}(\boldsymbol{\beta}^*)$ es la media muestral de $Q(\beta^*)$, y por la ley de los grandes números tenemos que

$R_{emp}(\boldsymbol{\beta}^*) \xrightarrow[N\to\infty]{} R(\boldsymbol{\beta})$ (1.5)

Lo que constituye la razón teórica para el uso del funcional empírico 1.4.

Sin embargo, la sola convergencia del valor esperado no implica que el mínimo de $R_{emp}(\boldsymbol{\beta})$ sea el mínimo de $R(\boldsymbol{\beta})$.

Para que esto suceda debemos imponer ciertas restricciones:

1. Que el $R_{emp}(\boldsymbol{\beta})$ aproxime uniformemente a $R(\boldsymbol{\beta})$ con una precisión $\boldsymbol{\varepsilon}\ tal\ que \left|R_{emp}\big(\widehat{\boldsymbol{\beta}}\big) - R(\boldsymbol{\beta_0})\right| < 2\varepsilon \Rightarrow \forall\, \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}\ y\ \boldsymbol{\varepsilon} > 0$ se tiene que $P\big(Sup_{\beta}\left|R(\beta) - R_{emp}(\beta)\right| > \varepsilon\big) \xrightarrow[N\to\infty]{} 0$

Entonces, podemos ahora definir formalmente el principio Minimización del Riesgo Empírico [Vapnik]:

1. En lugar del funcional de riesgo $R(\boldsymbol{\beta})$, construimos el funcional de riesgo empírico como

$R_{emp}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(\boldsymbol{x_i}, \boldsymbol{\beta})\right)$ sobre la base de muestras i.i.d $\{\boldsymbol{x}_i, y_i\}$, con $i=1,2,...N$

2. Sea $\widehat{\boldsymbol{\beta}}$ el vector de parámetros que minimiza el funcional de riesgo empírico $R_{emp}(\widehat{\boldsymbol{\beta}})$ sobre el espacio de parámetros $\boldsymbol{\mathcal{B}}$. Luego el $R_{emp}(\widehat{\boldsymbol{\beta}})$ converge en probabilidad al valor mínimo posible del funcional $R(\boldsymbol{\beta})$ verdadero, con $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$, a medida que el tamaño N de muestras del conjunto S se hace infinitamente grande, dado que el funcional de riesgo empírico $R_{emp}(\widehat{\boldsymbol{\beta}})$ converge uniformemente a $R(\boldsymbol{\beta})$.
3. Definimos convergencia uniforme mediante $P\left(Sup_{\beta}\left|R(\beta) - R_{emp}(\beta)\right| > \varepsilon\right) \xrightarrow[N\to\infty]{} 0$ Como una condición necesaria y suficiente para la consistencia del principio de minimización del riesgo empírico.

El principio de MRE en clasificación

En este caso tenemos que la salida "*y*" solo puede tomar dos valores posibles {0,1}. Por lo tanto la función de pérdida será

$$L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right) = \begin{cases} 0 \; si \; y = \hat{f}(\boldsymbol{x}, \boldsymbol{\beta}) \\ 1 \; si \; y \neq \hat{f}(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases} (1.6)$$

es decir que el valor 1 de *L* implica un "error", por ende

$$R_{emp}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}(\boldsymbol{x_i}, \boldsymbol{\beta})\right)$$

es el error de entrenamiento o frecuencia de ocurrencia de un error (error cometido sobre el conjunto de entradas utilizado para encontrar el parámetro $\widehat{\boldsymbol{\beta}}$) y el funcional

$$R(\boldsymbol{\beta}) = \int L\left(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\beta})\right) d\boldsymbol{F}(\boldsymbol{x}, y)$$

es la probabilidad de error de clasificación, denotada por $P(\beta)$. Luego por la ley de los grandes números tenemos que la frecuencia empírica de ocurrencia de un evento (error) converge casi seguramente a la $P(error)$ a medida que N→∞. Como vimos más arriba, tenemos que con probabilidad α se cumple que $Sup_{\beta}\left|R(\beta) - R_{emp}(\beta)\right| \geq \varepsilon \Rightarrow$ con p=1-α

$$P(\beta) = R(\beta) < R_{emp}(\beta) + \varepsilon\left(N, VC, \alpha, R_{emp}(\beta)\right) (1.7)$$

Esto nos quiere decir que, si utilizamos un algoritmo que encuentre el $\widehat{\boldsymbol{\beta}}$ utilizando el principio MRE, el error de dicha función esta acotado por el error empírico y un término que depende de la cantidad de muestras N, de la precisión α y de un parámetro "*VC*" conocido como dimensión de Vapnik-Chervonenkis (dimensión VC).

La dimensión VC juega un rol fundamental en la teoría del aprendizaje estadístico de donde se deriva el algoritmo de Máquinas de Soporte Vectorial. La dimensión VC es un concepto puramente combinatorial y no geométrico. La dimensión VC es un número que indica la cantidad de veces que un conjunto de datos *F* puede partirse en dos conjuntos disjuntos ($\mathcal{L}_0$ y $\mathcal{L}_1$) de manera tal que para cada una de estas particiones se satisfaga la siguiente ecuación:

$$f(\boldsymbol{x}, \boldsymbol{\beta}) = \begin{cases} 0 \; si \; \boldsymbol{x} \in \mathcal{L}_0 \\ 1 \; si \; \boldsymbol{x} \in \mathcal{L}_1 \end{cases} (1.8)$$

sin errores.

La ecuación 1.7 junto con la definición de la dimensión VC nos provee de varios conceptos teóricos que nos ayudan a encontrar una mejor solución en pos de minimizar la *P(error)*.

En busca de la solución, el hiperplano óptimo para patrones separables linealmente.

Supongamos que contamos con un conjunto de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ donde $\boldsymbol{x}_i$ es la i-ésima muestra con $y_i \in \{-1,1\}$ como su salida deseada (que sea 0 o -1 es indistinto, pero continuaremos con -1 por simpleza del análisis que sigue). Por el momento asumiremos que los patrones representados por $y_i = +1$ y los representados por $y_i = -1$ son "linealmente separables". Es decir que existe la función f que satisface la ecuación 1.8. Entonces, si son "linealmente separables" el hiperplano que separa ambas clases o categorías esta dado por

$$\boldsymbol{\beta^T x} + b = 0 \text{ (1.9)}$$

lo que es equivalente a escribir

$$\begin{matrix} \boldsymbol{\beta^T x_i} + b \geq 0 \ \ para\ y_i = +1 \\ \boldsymbol{\beta^T x_i} + b < 0 \ \ para\ y_i = -1 \end{matrix} \text{ (1.10)}$$

Para un vector de parámetros $\boldsymbol{\beta}$ y sesgo "*b*" dados, la separación entre el hiperplano definido por 1.9 y el dato más cercano se denomina "Margen de separación ρ". Denominaremos a hiperplano óptimo aquella solución para la cual ρ es máximo.

El objetivo ahora es encontrar el vector de parámetros $\widehat{\boldsymbol{\beta}}$ que defina ese hiperplano óptimo, dado el conjunto de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. En particular llamaremos "vectores de soporte" a aquellos puntos $(\boldsymbol{x}_i, y_i)$ para los cuales se satisface la igualdad de la siguiente ecuación:

$$\begin{matrix} \boldsymbol{\beta^T x_i} + b \geq 1 \ \ para\ y_i = +1 \\ \boldsymbol{\beta^T x_i} + b \leq -1 \ \ para\ y_i = -1 \end{matrix} \text{ (1.11)}$$

En este contexto, aquellos vectores que están exactamente a la distancia ρ del hiperplano satisfacen la igualdad en la 1.11. Si consideramos el vector de soporte $\boldsymbol{x}^{(s)}$ para el cual tenemos que $y^{(s)} = \mp 1$, por la 1.11 tenemos que

$$g\left(\boldsymbol{x}^{(s)}\right) = \beta_0^T \boldsymbol{x}^{(s)} \mp b = \mp 1 \text{ para } y^{(s)} = \mp 1.$$

Se puede demostrar que el margen ρ queda definido como $\rho = \frac{g(\boldsymbol{x}^{(s)})}{\|\beta_0\|}$, con lo cual quedan determinadas las condiciones necesarias para la búsqueda del vector de parámetros óptimo de la siguiente manera:

Dado el conjunto de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, encontrar los valores óptimos del vector de parámetros $\boldsymbol{\beta}$ y sesgo "*b*" tal que satisfagan las siguientes condiciones

$$y_i(\beta^T \boldsymbol{x_i} + b) \geq 1 \text{ para } i=1,2...N$$

y con el vector de parámetros $\boldsymbol{\beta}$ minimizando la función de costo o pérdida

$$\varphi(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta}\,.$$

de las condiciones anteriores podemos observar las siguientes características:

- La función de costo $\varphi(\boldsymbol{\beta})$ es convexa en $\boldsymbol{\beta}$.
- Las condiciones son lineales en $\boldsymbol{\beta}$.

Este problema de optimización puede resolverse mediante la aplicación de los multiplicadores de Lagrange, lo cual implica reescribir las condiciones de la siguiente manera:

$$J(\boldsymbol{\beta}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta} - \sum_{i=1}^{N}\alpha_i[y_i(\boldsymbol{\beta}^T\boldsymbol{x_i} + b) - 1]\,,\ \ \alpha \geq 0 \text{ (1.12)}$$

La función *J* tiene forma de silla de montar que debe ser minimizada con respecto a $\boldsymbol{\beta}$ y *b* y maximizada con respecto a $\boldsymbol{\alpha}$. La solución de esto lleva a las siguientes ecuaciones que representan los valores óptimos de los parámetros:

$$\boldsymbol{\beta} = \sum_{i=1}^{N}\alpha_i y_i\,\boldsymbol{x}_i \text{ (1.13)}$$

La solución para $\boldsymbol{\beta}$ parece implicar la sumatoria sobre todos los vectores existentes en el conjunto de entrenamiento, sin embargo solamente los coeficientes de Lagrange $\boldsymbol{\alpha}$ que satisfacen $\alpha_i[y_i(\boldsymbol{\beta}^T\boldsymbol{x_i} + b) - 1] = 0$ son los únicos que toman valores distintos de cero, reduciendo al 1.13 a

$$\boldsymbol{\beta} = \sum_{i \in VS}\alpha_i y_i \boldsymbol{x}_i$$

donde VS denota el subconjunto de $\boldsymbol{F}$ que satisface la igualdad de la 1.11.

Esta representación, conocida como "primal" tiene también una representación dual de la siguiente forma:

Dado el conjunto de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, encuentre los multiplicadores de Lagrange $\{\alpha_i\}_{i=1}^{N}$ que maximizan la función objectivo

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\,\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

sujeto a las condiciones

1. $\sum_{i=1}^{N}\alpha_i y_i = 0$
2. $\alpha_i \geq 0$ para *i=1,2,...,N*

Ahora bien, ¿Cómo se relaciona las Máquinas de Soporte Vectorial con lo expuesto sobre la teoría de la minimización del riesgo empírico?

Mediante el siguiente Teorema [Vapnik, 1995, 1998]

Sea D el diámetro de la "bola" más pequeña que contiene a todos los vectores de entrada $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. El conjunto de hiperplanos óptimo descripto por la ecuación $\boldsymbol{\beta_0^T x} + b_0 = 0$ tiene una dimensión VC "*h*" restringida superiormente por

$$h \leq min\left\{\left[\frac{D^2}{\rho^2}\right], m_0\right\} + 1$$

donde [.] refiere al valor entero mínimo que es mayor o igual al valor encerrado entre los corchetes, ρ es el margen de separación que es igual a $1/\|\boldsymbol{\beta_0}\|$, y m_0 es la dimensión del espacio de entradas.

Ahora bien, hasta el momento solo hemos abordado la situación "ideal" de que los datos sean linealmente separables, pero en la realidad eso no sucede muy a menudo y en general existe cierto grado de solapamiento entre las distintas clases a clasificar, más aún cuando la dimensión del espacio de entradas es grande. Sin embargo, cuando este es el caso, aun podemos encontrar un hiperplano óptimo que minimice la probabilidad del error de clasificación promedio sobre el conjunto de entrenamiento.

En este contexto decimos que el margen de separación entre las clases es "blando" si existen o permitimos puntos $(\boldsymbol{x}_i, y_i)$ que violen la condición $y_i(\beta^T \boldsymbol{x_i} + b) \geq 1$. La violación a esta condición puede devenir de dos situaciones:

- El punto $(\boldsymbol{x}_i, y_i)$ cae dentro de la región de separación
- El punto $(\boldsymbol{x}_i, y_i)$ cae del lado opuesto al que debería según el valor de y_i, es decir del lado equivocado con respecto a la superficie de decisión.

Para formalizar el tratamiento de estas dos situaciones, introducimos un nuevo conjunto de variables (variables de holgura o amplitud) $\{\xi_i\}_{i=1}^N$ en la definición del hiperplano de separación según:

$$y_i(\beta^T \boldsymbol{x_i} + b) \geq 1 - \xi_i$$

Las variables de holgura ξ_i miden la desviación de cada punto con respecto a la condición ideal. Cuando $0 \leq \xi_i \leq 1$ el dato cae dentro de la región de separación (dentro del margen) pero del lado correcto con respecto al hiperplano. Mientras que si $\xi_i > 1$ el punto cae en el lugar equivocado del plano de decisión. Para encontrar ahora el vector de parámetros que optimiza el margen, reescribimos el funcional de costo como sigue:

$$\varphi(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\beta^T\beta} + \boldsymbol{C}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\xi_i}$$

La constante **C** > 0 cumple el rol de parámetro de "regularización", que controla el compromiso entre la complejidad de de la MSV y el número de vectores de soporte no-separables. Este parámetro C es usualmente seleccionado por el usuario.

La forma dual del problema de optimización queda ahora expresada de la siguiente forma:

Dado el conjunto de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, encuentre los multiplicadores de Lagrange $\{\alpha_i\}_{i=1}^N$ que maximizan la función objectivo

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

sujeto a las condiciones

1. $\sum_{i=1}^N \alpha_i y_i = 0$
2. $0 \leq \alpha_i \leq C$ para *i=1,2,...,N*

donde C es un parámetro positivo especificado por el usuario.

Como se puede ver, en la formulación dual no aparecen las "variables de holgura", siendo la solución prácticamente la misma que en el caso de margen rígido, excepto por una pequeña pero importante diferencia, ahora los α_s tienen una cota superior *C*.
El enlace entre la solución primal y dual se mantiene como

$$\boldsymbol{\beta} = \sum_{i \in VS} \alpha_i y_i \boldsymbol{x}_i$$

Solo que ahora tenemos que si $\alpha_i < C$, entonces $\xi_i = 0$ => $\boldsymbol{x_i}$ está del lado correcto del hiperplano, en cambio si $\alpha_i = C$, entonces $\xi_i > 0$ => $\boldsymbol{x_i}$ está dentro del margen si $0 < \xi_i < 1$ o del lado incorrecto del hiperplano $\xi_i \geq 1$.

El principio de MRE en regresión

En este caso, la variable $y \in \mathcal{R}$. En el abordaje clásico e incluso en métodos basados en redes neuronales artificiales, la función de pérdida es la cuadrática i.e $\left(y - f(\boldsymbol{x}, \boldsymbol{\beta})\right)^2$. Sin embargo, este estimador de riesgo es muy sensible a valores infrecuentes ("outliers") como a distribuciones asimétricas. Para abordar este problema, se requieren estimadores robustos e insensibles a pequeños cambios en el modelo.
Como una extensión del modelo basado en SVM que se ajusta a los criterios del MRE, Vapnik propuso en 1995 la siguiente función de pérdida:

$$L_\varepsilon^p(f(\boldsymbol{x}, \boldsymbol{\beta}), y) = \begin{cases} |f(\boldsymbol{x}, \boldsymbol{\beta}) - y|^p - \epsilon, & para\ |f(\boldsymbol{x}, \boldsymbol{\beta}) - y|^p \geq \varepsilon \\ 0 & en\ cc \end{cases} \quad with\ p = 1{,}2$$

donde ε es el parámetro de tolerancia. Esta función de pérdida L_ε^p se denomina $\varepsilon - insensible$ de norma *p*.
Ahora podemos suponer que estamos ante el siguiente problema:

$$y = f(\boldsymbol{x}) + \nu$$

donde la función $f(\boldsymbol{x})$ (posiblemente no lineal) está definida por la esperanza condicional $E(Y|\boldsymbol{x})$ donde $\boldsymbol{Y}$ es una variable aleatoria con realizaciones "y". El ruido aditivo "ν" es estadísticamente independiente de "$\boldsymbol{x}$" y tanto la función *f()* como "ν" son desconocidos. Todo lo que tenemos disponible es el conjunto de datos de entrenamiento $F = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$
Otra vez aquí, el problema consiste en minimizar el riesgo empírico

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} L_\epsilon(y_i, \boldsymbol{x_i})$$

Y para el caso de L_ε^1 podemos formular el problema de la siguiente manera:

$$y_i - \boldsymbol{\beta}^T \boldsymbol{x_i} \leq \epsilon - \xi_i$$
$$\boldsymbol{\beta}^T \boldsymbol{x_i} - y_i \leq \epsilon - \xi_i'$$

donde $\{\xi_i \geq 0\}_{i=1}^N$ son variables de holgura. Este problema de optimización es equivalente a minimizar el siguiente funcional de costo

$$\Phi(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi'}) = C\left(\sum_{i=1}^{N} (\boldsymbol{\xi_i} + \boldsymbol{\xi_i'})\right) + \frac{1}{2}\boldsymbol{\beta^T \beta}$$

Sujeto a

$$\left((\boldsymbol{\beta} \cdot \boldsymbol{x}_i) + b\right) - y_i \leq \varepsilon + \xi_i$$
$$y_i - \left((\boldsymbol{\beta} \cdot \boldsymbol{x}_i) + b\right) \geq \varepsilon + \xi_i'$$

$$\xi_i, \xi_i^{'} \geq 0, i = 1,2 \ldots, N$$

donde la constante C es un parámetro especificado por el usuario.

Mediante la aplicación de los multiplicadores de Lagrange, este problema de optimización se puede escribir de la siguiente manera:

$$\mathrm{B}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} y_i \cdot \alpha_i - \boldsymbol{\varepsilon} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\alpha_i| - \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i,j=1}}^{\boldsymbol{N}} \alpha_i \alpha_j \boldsymbol{x}_i \boldsymbol{x}_j$$

Sujeto a

$$\sum_i \alpha_i = 0, -C \leq \alpha_i \leq C, i = 1..N$$

Como en el caso anterior tenemos que solo unos pocos coeficientes de Lagrange son distintos de cero y definen a los vectores de soporte, por lo que podemos escribir el vector

$$\boldsymbol{\beta} = \sum_{i \in SV} (\alpha_i)\, \boldsymbol{x}_i$$

Luego la ecuación de predicción queda como

$$y_j = \boldsymbol{\beta} \boldsymbol{x_j} = \sum_{i \in SV} (\alpha_i)\, \boldsymbol{x}_i \boldsymbol{x}_j$$

Referencias:

Haykin, S. Neural Networks and Learning Machines 3ed. Prentice Hall. 2008

Vapnik V. The nature of Statistical Learning Theory. 2ed. Springer 1999.