Aprendizaje Automático y Minería de Datos

Support Vector Machines

Cristina Tîrnăucă

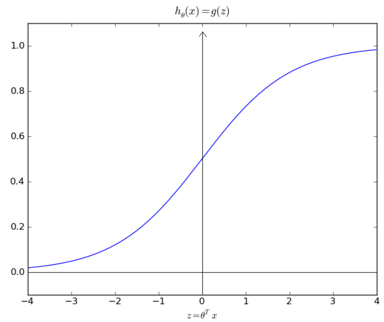
Dept. Matesco, Universidad de Cantabria

Fac. Ciencias – Grado en Ing. Informática

Regresión logística

Repaso

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



Predecimos "$y = 1$" si $h_\theta(x) \geq 0.5$
"$y = 0$" si $h_\theta(x) < 0.5$

Si $y = 1$, queremos $h_\theta(x) \approx 1$ (o equivalente, $\theta^T x \gg 0$)
$y = 0$, queremos $h_\theta(x) \approx 0$ (o equivalente, $\theta^T x \ll 0$)

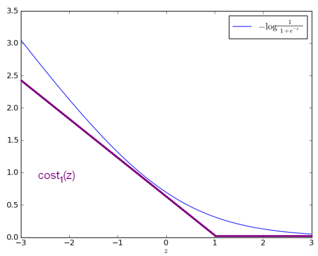
Regresión logística

Función de coste para un ejemplo (x,y)

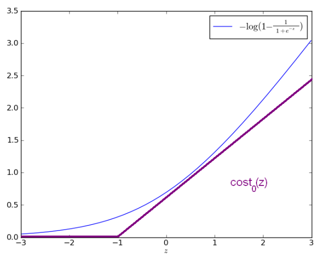
$$-y \log h_\theta(x) - (1-y)\log(1-h_\theta(x))$$

$$y(-\log \frac{1}{1+e^{-\theta^T x}}) + (1-y)(-\log(1-\frac{1}{1+e^{-\theta^T x}}))$$

Si $y = 1$ (queremos $\theta^T x \gg 0$)



Si $y = 0$ (queremos $\theta^T x \ll 0$)



Support Vector Machine

Objetivo: minimizar la función de coste

Regresión logística

$$\min_\theta \frac{1}{m}[\sum_{i=1}^{m} y^{(i)}(-\log h_\theta(x^{(i)}))+(1-y^{(i)})(-\log(1-h_\theta(x^{(i)})))]+\frac{\lambda}{2m}\sum_{j=1}^{p}\theta_j^2$$

Support Vector Machine

$$\min_\theta C[\sum_{i=1}^{m} y^{(i)} cost_1(\theta^T x^{(i)}) + (1-y^{(i)}) cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{j=1}^{p}\theta_j^2$$

Support Vector Machine

Función de coste

$$\min_{\theta} C[\sum_{i=1}^{m} y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{j=1}^{p} \theta_j^2$$

Hipótesis

$$h_\theta(x) = \begin{cases} 1 & \text{si } \theta^T x \geq 0 \\ 0 & \text{si } \theta^T x < 0 \end{cases}$$

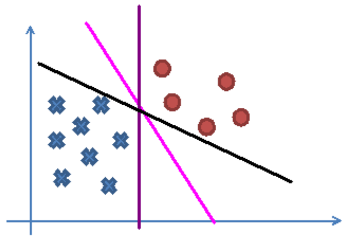
SVM es un Large Margin Classifier

Sí, cuando C es muy grande

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{p} \theta_j^2$$

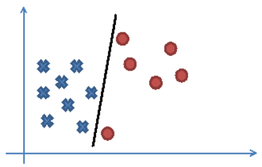
bajo las condiciones: $\theta^T x^{(i)} \geq 1$, si $y^{(i)} = 1$
$\theta^T x^{(i)} \leq -1$, si $y^{(i)} = 0$

El umbral de decisión: caso de puntos linealmente separables

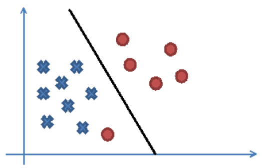


SVM en presencia de outliers

Cuando C es muy grande

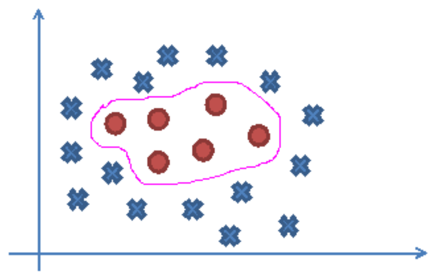


Cuando C no es muy grande



Umbral de decisión no lineal

Cuando los datos no son linealmente separables



Podríamos añadir más atributos ...

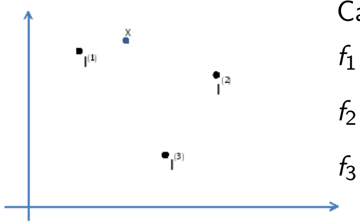
$$h_\theta(x) = \begin{cases} 1, & \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \geq 0 \\ 0, & \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots < 0 \end{cases}$$

Una alternativa mejor: utilizar una función Kernel (Núcleo)

proyectando la información a un espacio de características de mayor dimensión

Comenzaremos con un ejemplo:

Sean $l^{(1)}, l^{(2)}, l^{(3)}$ tres puntos de referencia y x un punto en el plano (todos en R^2)



Calculamos:

$$f_1 = K(x, l^{(1)}) = \exp(-\frac{\left\|x-l^{(1)}\right\|^2}{2\sigma^2})$$

$$f_2 = K(x, l^{(2)}) = \exp(-\frac{\left\|x-l^{(2)}\right\|^2}{2\sigma^2})$$

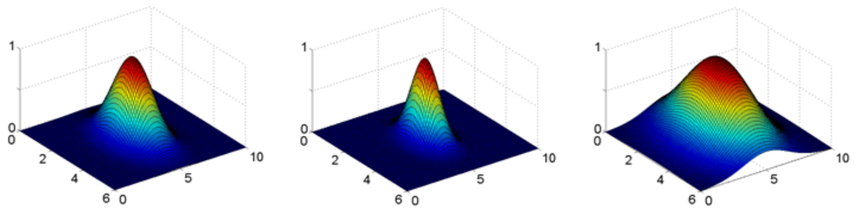
$$f_3 = K(x, l^{(3)}) = \exp(-\frac{\left\|x-l^{(3)}\right\|^2}{2\sigma^2})$$

Predecimos "$y = 1$" si $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$
(ejemplo con $\theta_0 = -0.5, \theta_1 = 1, \theta_2 = 1, \theta_3 = 0$)

El rol del parámetro σ

$$l^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

$$f^{(1)} = exp(-\frac{\left\|x-l^{(1)}\right\|^2}{2\sigma^2})$$



$\sigma^2 = 1$ $\sigma^2 = 0.5$ $\sigma^2 = 3$

SVM con Kernels

¿Cómo elegimos los puntos de referencia?

Dados $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$
Elegimos $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \ldots, l^{(m)} = x^{(m)}$
Para un ejemplo $x^{(i)}$ en R^{p+1} construimos $f^{(i)}$ en R^{m+1}:

$$f_0^{(i)} = 1$$
$$f_1^{(i)} = K(x^{(i)}, l^{(1)}) = K(x^{(i)}, x^{(1)})$$
.
.
.
$$f_i^{(i)} = K(x^{(i)}, l^{(i)}) = K(x^{(i)}, x^{(i)})$$
.
.
.
$$f_m^{(i)} = K(x^{(i)}, l^{(m)}) = K(x^{(i)}, x^{(m)})$$

Support Vector Machine con Kernels

Cambio de dimensión

Para cada uno de los m datos de entrenamiento $x^{(i)} \in R^p$, calcular $f^{(i)} \in R^m$.

Fase de entrenamiento

$$\min_{\theta} C[\sum_{i=1}^{m} y^{(i)} cost_1(\theta^T f^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T f^{(i)})] + \frac{1}{2} \sum_{j=1}^{m} \theta_j^2$$

Fase de predicción

Para cada nuevo ejemplo $x \in R^p$, calcular $f \in R^m$.
Predecir "$y = 1$" si $\theta^T f \geq 0$
Predecir "$y = 0$" si $\theta^T f < 0$

SVM en la práctica

Se recomienda utilizar los paquetes existentes (liblinear, libsvm,...) para obtener los parámetros θ.
Se deben elejir:

- el valor del parámetro C
- el kernel:
  - "No kernel" ("linear kernel"): predecir "$y = 1$" si $\theta^T x \geq 0$
  - Gaussian kernel: $K(x, l) = \exp\left(-\frac{\|x-l\|^2}{2\sigma^2}\right)$
    (hay que elegir σ y escalar los variables)
  - Polinomial kernel $K(x, l) = (x^T l + c)^d$
    (hay que elegir c y d)
  - otros: String kernel, chi-square kernel, histogram intersection kernel,...

Regresión logística versus SVM

Consejos prácticos

p= número de atributos, m = número de datos de entrada

Cuando p es mucho más grande que m

Se aconseja utilizar regresión logística o SVM sin kernel.

Cuando p es pequeño y m no muy grande

Se aconseja utilizar SVM con Gaussian kernel o similar.

Cuando p es pequeño y m es muy grande

Se aconseja aǹadir más atributos y utilizar regresión logística o SVM sin kernel.