L2 wtr

Tianrui Wang

April 2019

1 Supporting Vector Machine

Brief History
This is Theoretically Based and Global Minima

Optimization Goal
(1)Classification problem
Training set S=$\{(\mathbf{x_1}, y_1), (x_2, y_2) \ldots\}$
Our goal is to learn a function $g(x)$ *where* $f(x) = sgn(g(x))$ *fits the samples best*
linear function $g(x) = w^T x + b$
(2)**Destination**
$\mathbf{labels: \ y \in \{-1, 1\}}$
$\mathbf{geometricmargin: \ \gamma^(i) = y^{(i)}(w^T x + b)}$
in order to simplifiy the calculation ,we make $|\mathbf{w}| = \mathbf{1}$
$\gamma = \mathbf{min} \ \gamma^{(\mathbf{i})}$
We use γ **to judge how well the function proforms**
find $\mathbf{max}_{\gamma,\mathbf{w},\mathbf{b}}\gamma$
$\mathbf{s.t.y^{(i)}(w^T x^{(i)} + b) \geq \gamma}$
where $||\mathbf{w}|| = \mathbf{1}$
(3)**Transformation**
As Unit Circle is not a convex set , we transform the origin question to...
find $\mathbf{max}_{\gamma,\mathbf{w},\mathbf{b}} \ \gamma/||\mathbf{w}||$
$\mathbf{s.t.y^(i)(w^T x^{(i)} + b) \geq \gamma}$
We suppose $||\gamma|| = \mathbf{1}$ **to simplify the question**
So the problem is to find the $\mathbf{min_{w,b} \ 0.5||w||^2}$
(4)**Lagrangianform**
$\mathbf{L(w, b, \alpha) = \frac{||w||^2}{2} - \sum_{i=0}^{N} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1]}$
$\mathbf{s.t.} \ \alpha_\mathbf{i} \geq \mathbf{0}$

Dual Problem :
$\mathbf{max_\alpha min_{w,b} L(w, b, a) = L(w, b, \alpha) = \frac{||w||^2}{2} - \sum_{i=0}^{N} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1]}$
With KKT conditions the problem′s answer is the origin problem′s
We swap the min and the max to make it easier
figure the Partial guidance

$\mathbf{w}=\sum_{i=1}^{N}\alpha_i y^{(i)}x^{(i)}$

Improvement

(1)Allow error

We allow error ξ_i in classification;

it is based on the output of the discriminant function w^Tx+b.

So we can reach a balance between acuracy and generalization.

we replace the origin formula with $\min_{w,b,\xi}\frac{||w||}{2}+C\sum_i=1^N\xi_i$

s.t. $y(w^Tx^{(i)}+b)\geq 1-\xi_i \quad \xi_i\geq 0$

(2)Kernel function

In many real situations, dividing the samples linearly is not enough, so we use a map to increas

in order to get a linear dividing in a high dimensional space

define $K(x,z)=\phi(x)^T\phi(x)$

and replace all $(x^{(i)},x^{(j)})$ with $(\phi(x^{(i)}),\phi(x^{(j)}))$

However, it is difficult to find the proper kernel function...

Here are some commonly used function

$K(x,y)=(xy+1)^2$

$K(x,y)=\exp(\frac{||x-y||^2}{2\sigma^2})$

Implementation

(1)Coordinate descend method

change others with one fixed

Loop until convergence

{

For(i = 1 : m)

{

$a_i=\mathrm{argmax}_{a}iL(a_1,a_2,\ldots,a_{i-1},a_{i+1},\ldots,a_m)$

}

}

it is slow with low efficiency

(2)SMO algorithm

Repeat until convergence

{

First, select one pair a_i and a_j to update next

Second, reoptimize $L(a)$ with respect to a_i and a_j without changing other a_k

}

(3)Multi – class classification

Method 1 one – versus – rest

Advantages : low cost

Disadvantages : unbalanced data

Examples : circle and not circle, triangle and not triangle...

Method 1 one – versus – one

Advantages : make judgment by voting

Disadvantages : high cost (C_n^2 models)

Examples : circle and tirangle, circle and square, tirangle and sqaure...