Machine Learning

Lect 9

Dr. Pruane
Mukherjee

Solving the Optimization Problem

accuracy
misclassification

minimize $\frac{1}{2}\|w\|^2$

s.t. $y_i(w^T x_i + b) \geq 1$ (linear inequality constraints)

→ Optimization problem with convex quadratic objectives and linear constraints

→ Can be solved using QP.

→ Lagrange duality to get the optimization problem's dual form:

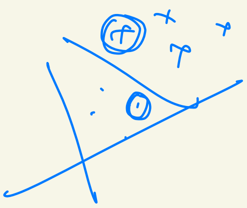
- Allow us to use kernels to get optimal margin classifiers to work efficiently in very high dimensional space
- Allow us to derive an efficient algorithm for solving the above optimization problem that will typically do much better than generic QP software.

Lagrangian Duality

The Primal Problem: $\min_w f(w)$ → Parameters: to find values of w to minimize $f(w)$, 'k'

s.t. $g_i(w) \leq 0, \; i = 1, \ldots, k$ (linear inequality constraints)

$h_i(w) = 0, \; i = 1, \ldots, l$ ('l' linear equality constraints)



The generalized Lagrangian:

$$\boxed{L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)}$$

the α_i's ($\alpha_i \geq 0$) and β's are called the Lagrange multipliers.

Lemma:

$$\max_{\alpha, \beta, \alpha_i \geq 0} L(w, \alpha, \beta) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise} \end{cases}$$

A re-written primal:

$$\min_w \max_{\alpha, \beta, \alpha_i \geq 0} L(w, \alpha, \beta)$$

The Primal Problem $p^* = \min_w \max_{\alpha, \beta, \alpha_i \geq 0} L(w, \alpha, \beta)$

The Dual Problem: $d^* = \max_{\alpha, \beta, \alpha_i \geq 0} \underline{\underline{\min_w L(w, \alpha, \beta)}}$

Theorem (weak duality):

$$d^* = \max_{\alpha, \beta, \alpha_i \geq 0} \min_w L(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta, \alpha_i \geq 0} L(w, \alpha, \beta) = p^*$$

Theorem (strong duality):

iff there exist a saddle point of $L(w, \alpha, \beta)$ we have $\boxed{d^* = p^*}$ → optimal value of both primal & dual forms

KKT conditions

If there exists some saddle point of L, then it satisfies the following Karush Kuhn Tucker (KKT) conditions:

$$\frac{\partial}{\partial w_i} L(w, \alpha, \beta) = 0, \quad i = 1, \dots, k$$

$$\frac{\partial}{\partial \beta_i} L(w, \alpha, \beta) = 0, \quad i = 1, \dots, l$$

$$\alpha_i g_i(w) = 0, \quad i = 1, \dots, m$$

$$g_i(w) \leq 0, \quad i = 1, \dots, m$$

$$\alpha_i \geq 0, \quad i = 1, \dots, m$$

Theorem: If w^*, α^*, β^* satisfy the KKT condition, then it is also a solution to the primal and the dual problems.

Support Vectors

In SVM, we just have g_i's (inequality constraint)
no h_i's (equality constraints)

- Only a few α_i's can be nonzero $\Rightarrow$ i's are S.Vs.
- Call the training data points whose α_i's are nonzero the support vectors.

$$\alpha_i g_i(w) = 0 \quad , i = 1, \dots, m$$

if $\alpha_i > 0$ then $g(w) = 0$

Solving the Optimization problem

Quadratic programming with linear constraints

$$\text{minimize } \boxed{\frac{1}{2} \|w\|^2} \quad f(x)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \Rightarrow$$

Lagrangian function $\updownarrow$ $\quad f(w) + \underline{\sum_{i=1}^{k} \alpha_i g_i(w)} +$

$$\text{minimize } L_p(\underline{w, b}, \alpha_i) = \frac{1}{2} \|w\|^2 - \underline{\sum_{i=1}^{m} \alpha_i \left(y_i (w^T x_i + b) - 1\right)} \to 1$$

$$\text{s.t. } \alpha_i > 0 \Rightarrow \text{S.V.}$$

minimize w.r.t w & b for fixed α

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \boxed{w = \sum_{i=1}^{n} \alpha_i y_i x_i} \to 2$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \to 3$$

Put 2 in 1

$$L_p(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j) - b \sum_{i=1}^{m} \alpha_i y_i \to 4$$

Put (3) in (4)

$$L_p(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

The Dual Problem :

Now we have the following dual opt problem :

$$\max_{\alpha} J(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

$$\text{s.t. } \alpha_i \geq 0 , \quad i = 1, \ldots, k$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

This is a quadratic programming problem :

→ A global maximum of α_i can always be found.

This dual opt. problem is much easier to solve than primal formulation hence constraints are simpler

SVM

Once we have the Lagrange multipliers $\{\alpha_j\}$ we can reconstruct the parameter vector w as a weighted combination of the training examples :

$$\boxed{w = \sum_{i=1}^{m} \alpha_i y_i x_i} \qquad \boxed{w = \sum_{i \in SV} \alpha_i y_i x_i}$$

$\alpha_i > 0$

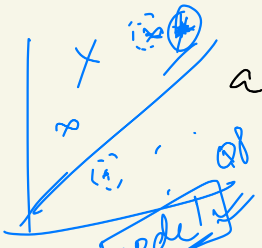
α_i's non zero for $i \in SV$

iris 3 labels

For testing with new data 'z'

– Compute

$$w^T z + b = \sum_{i \in SV} \alpha_i y_i (x_i^T z) + b \quad \text{— (1)}$$



and classify z as class 1 if sum is positive and class 2 otherwise

"2" model labels

Note: w need not be formed explicitly we can just use (1) as we take dot product of support vectors with z

1 → setosa
2 →
3 →

OvO classifier

OvA classifier

Solving the Opt. Problem

→ The discriminant function is:

$$g(x) = w^T x + b = \sum_{i \in SV} \alpha_i x_i^T x + b$$

→ It relies on a dot product b/w test point x and support vectors x ⇒ we get a scalar

→ Solving the opt. problem involved is computing dot products $x_i^T x_j$ b/w all pairs of training points

→ The optimal w is a linear combination of a small no of data points.

multilabel classifier :

CL 1 CL2 CL 3

OVO

CL1 vs CL 2 binary classifier

CL1 vs CL 3

OVA

CL1 vs CL2 CL3 KNN

~~posi~~ positive

negative