Week 2 Report

Li Zhengyuan
Tsinghua University
zhengyua17@mails.tsinghua.edu.cn

Abstract

Notes for SVM and kernel method.

1 SVM

1.1 Optimization Goal

- Given a training set S={(x1, y1),(x2, y2),. . . ,(xN, yN)}, and xi∈ X=Rm, i=1,2,. . . ,N
- To learn a function g(x), and make the decision function f(x)=sgn(g(x)) can classify new input x

1.2 Definitions

- Given x_i, y_i
- funtional margin :$\widehat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$
- geometric margin(can be seen as normalized) :$\widehat{\gamma}^{(i)} = y^i(\frac{w}{||w||}^T x + \frac{b}{||w||})$
- We define margin for a training set as:
  $\gamma = min\gamma^i$

1.3 Transformation

- Given x_i, y_i
- funtional margin :$\widehat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$
- geometric margin(can be seen as normalized) :$\widehat{\gamma}^{(i)} = y^{(i)}(\frac{w}{||w||}^T x + \frac{b}{||w||})$
- We define margin for a training set as:
  $\gamma = min\gamma^i$

1.4 Optimization Problem

Our problem is:

- $min_{\gamma,w,b} \quad \gamma$
- $s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma$

It can be transformed into:

- $min_{w,b} \quad \frac{1}{2}||w||^2$
- $s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.5 None Linearly Separable Case

We allow "error" ξ_i in classification Then the optimization goal become:

- $min_{w,b} \quad \frac{1}{2}||w||^2 + C\sum_{n=1}^{m} \xi_n$
- $s.t. \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$
- $\xi_i \geq 0$

we can form the Lagrangian:

$$L(w, b, \xi, \alpha, r) = 1/2w^T w + C\sum_{n=1}^{m} \alpha_i[y^i(x^T w + b) - 1 + \xi_i] - \sum_{n=1}^{m} r_i \xi_i \tag{1}$$

α_i 's andr_i 's are greater than 0.

By setting the derivatives to 0 to satisfy the KKT condition,we obtain the following dual form

- $max_\alpha \quad \sum_{i=1}^{m} \alpha_i - 1/2\sum_{i,j=1}^{m} y^i y^j \alpha_i \alpha_j < x^i, x^j >$
- $s.t. \quad \alpha_i \geq 0$
- $\sum_{i=1}^{m} \alpha_i y_i = 0$

We should notice that by introducing KKT condition,this form is almost equivalent to the prime problem.

BTW, by setting the derivatives with respect to w, we find that w is a linear combination of α_i.(The Representer Theorem)

1.6 How to solve SVM?

```
Repeat until convergence
{
   1. select some pair ai and aj to update next.
   2. reoptimize L(a) with respect to ai and aj, while holding all the other a.
}
```

2 Kernel Method

2.1 What is kernel?

If we have an feature map ϕ,then we define kernel corresponding to ϕ as:

$$k(x, x') =< \phi(x), \phi(x') > \tag{2}$$

2.2 Why kernel is right?

As we can see at the end of last section,the item relevant to x is an inner product.So,if we want to transform the x with some ϕ, to use the kernel function corresponding to ϕ is an equivalent .

In short, kernel method is an alternative choice besides feature map.

2.3 Why kernel?

- Overall,we can see that sometimes feature mapping is infeasible or it requires too much computing resource.That's why we use kernel.We should notice that a simple kernel function can be derived from many feature mapping functions.(e.g $k(x, x') = (1+ < x, x' >)^M$ corresponds to a feature map with all monomials up to degree M.)
- Kernel evaluation can be fast.

  Let's assume that $\phi(x) = allmonomialsuptodegreeM$,its dimension is $O(d^2)$.Now we try to calculate K(x,x').

- explicit computation:$O(d^2)$
- implicit computation:$O(d)$

- Kernel functions can allow access to infinite-dimensional feature spaces(e.g. RBF kernel).

2.4 notes

- Kernel method is off-line.
- To make a prediction, we need to touch all the training inputs,but we avoid the complexity of features.So that is a trade-off.
- There are Kernel Ridge Regression,Kernel Logistic Regression and others.