[Paper review 19]

Relevance Vector Machine Explained

(Tristan Fletcher, 2010)

[Contents]

- 1. Introduction
- 2. RVM for Regression
- 3. Analysis of Sparsity
- 4. RVM for Classification

1. Introduction

SVM

- Not a probabilistic prediction
- Only Binary decision
- have to tune hyperparameter C

RVM is more sparse, and can solve three problems above.

2. RVM for Regression

RVM = Linear Model + Modified prior for sparse solution

2.1 Model setup

1) conditional distribution : $p(t \mid x, w, \beta) = N(t \mid y(x), 1/\beta)$

2) prior:

- ullet (LM) $p\left(w_i
 ight)=N\left(0,1/lpha
 ight)$
- (RVM) $p(w_i) = N(0, 1/\alpha_i)$

3) posterior : $p(w \mid t, X, \alpha, \beta) = N(w \mid m, \Sigma)$, where

- $m = \beta \Sigma \Phi^T t$
- ullet $\Sigma = \left(A + eta \Phi^T \Phi
 ight)^{-1}$ (where $A = diag(a_i)$)

2.2 Maximize Marginal Likelihood

• find optimal α and β by maximizing marginal likelihood, $p(t \mid X, \alpha, \beta)$

$$egin{split} p(t\mid X,lpha,eta) &= \int p(t\mid X,w,eta)p(w\midlpha)dw \ &= \int N\left(t\mid w^{ op}\phi(x),rac{1}{eta}
ight)N\left(w\mid 0,A^{-1}
ight)dw \ &= N(oldsymbol{t}\mid 0,C) \end{split}$$

where $C = \beta^{-1}I + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T$

Woodbury identity

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$

In our case, $A=eta^{-1}I, B=\Phi, D=A, C=\Phi^T$

If we solve...

- $\begin{array}{l} \bullet \quad \frac{\partial}{\partial \alpha} p(t \mid X, \alpha, \beta) = 0 \\ \bullet \quad \frac{\partial}{\partial \beta} p(t \mid X, \alpha, \beta) = 0 \end{array}$

Solution:

$$lpha_i^{ ext{new}} = rac{r_i}{m_i^2} = \left(1 - lpha_i \Sigma_{ii}
ight)/m_i^2: ext{ implicit} \ \left(eta^{ ext{Nex}}
ight)^{-1} = \|t - oldsymbol{\Phi} m\|^2/\left(N - \sum_i r_i
ight)$$

- can not solve α_i^{new} directly...
- step 1) initialize α_0 and β_0

step 2) find posterior

step 3) update α and β

step 4) repeat step 2 & 3

Relevance Vector

- data(vector) with non-zero weight
- ullet $lpha_ipprox\infty$, $w_i=0$

RVM vs SVM

1) Sparsity: RVM > SVM

2) Generalization: RVM > SVM

3) Need to estimate hyperparameter: only SVM

4) Training Time: RVM>>SVM

3. Analysis of Sparsity

Alternative way to train RVM, due to long training time.

[Log Marginal Likelihood (L(lpha))]

$$L(\alpha) = L(\alpha_i) + \lambda(\alpha_i)$$

$$L(lpha) = \ln(p(m{t}\mid m{X}, m{lpha}, eta)) = \ln(N(m{t}\mid 0, C))$$
 where $C = eta^{-1}I + \sum_{i
eq i} lpha_i^{-1} \phi_i \phi_i^T + lpha_i^{-1} \phi_i \phi_i^T = C_{-i} + lpha_i^{-1} \phi_i \phi_i^T$

Solution:

$$\lambda\left(oldsymbol{lpha}_{i}
ight)=rac{1}{2}igg\{\lnig(ig|oldsymbol{1}+lpha_{i}^{-oldsymbol{1}}\phi_{i}^{T}oldsymbol{C}_{-i}^{-oldsymbol{1}}\phi_{i}ig|ig)-oldsymbol{t}^{T}\left(rac{C_{-i}^{-1}\phi_{i}\phi_{i}^{T}C_{-i}^{-1}}{lpha_{i}+\phi_{i}^{T}c_{-i}^{-1}\phi_{i}}
ight)oldsymbol{t}$$

$$s_i = \phi_i^T C_{-i}^{-1} \phi_i$$

- sparsity of ϕ_i
- ullet overlap between ϕ_i and ϕ_j

$$q_i = \phi_i^T C_{-i}^{-1} t$$

- quality of ϕ_i
- ullet $C_{-i}^{-1} t$: prediction error $o q_i:$ information about ϕ_i

$$s_i > q_i \rightarrow \phi_i = 0$$

$$s_i < q_i
ightarrow \phi_i
eq 0$$

$$rac{\partial}{\partial lpha} L(lpha) = rac{\partial}{\partial lpha} \lambda(lpha) = rac{lpha_i^{-1} s_i^2 + s_i - q_i^2}{\left(lpha_i + s_i
ight)^2} 0$$

$$ullet$$
 if $s_i \leq q_i^2
ightarrow lpha_i = rac{s_i^2}{s_i - q_i^2}$

$$ullet$$
 if $s_i>q_i^2
ightarrow lpha_i=\infty$

4. RVM for Classification

Binary case

$$ullet \ y(x) = \sigma\left(w^Tarphi(x)
ight)$$

Multi-class case

$$ullet \ y_K(x) = rac{\expig(w_K^T arphi(x)ig)}{\sum_j \expig(w_j^T arphi(x)ig)}$$

Hard to calculate marginal likelihood, so use Laplace Approximation