[Paper review 19]

Relevance Vector Machine Explained

(Tristan Fletcher, 2010)

[Contents]

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 :

- (LM) $p\left(w_i\right) = N\left(0, 1/\alpha\right)$
- (RVM) $p\left(w_i\right) = N\left(0, 1/\alpha_i\right)$

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

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

2.2 Maximize Marginal Likelihood

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

$$
\begin{aligned}
p(t \mid X, \alpha, \beta) &= \int p(t \mid X, w, \beta) p(w \mid \alpha) dw \\
&= \int N\left(t \mid w^{\top} \phi(x), \frac{1}{\beta}\right) N\left(w \mid 0, A^{-1}\right) dw \\
&= N(\boldsymbol{t} \mid 0, C)
\end{aligned}
$$

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

Woodbury identity

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

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

If we solve...

- $\frac{\partial}{\partial \alpha} p(t \mid X, \alpha, \beta) = 0$
- $\frac{\partial}{\partial \beta} p(t \mid X, \alpha, \beta) = 0$

Solution :

$\alpha_i^{\text{new}} = \frac{r_i}{m_i^2} = \left(1 - \alpha_i \Sigma_{ii}\right) / m_i^2 \text{ : implicit}$

$\left(\beta^{\text{Nex}}\right)^{-1} = \|t - \mathbf{\Phi} m\|^2 / \left(N - \sum_i r_i\right)$

- 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
- $\alpha_i \approx \infty$, $w_i = 0$

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 ($L(\alpha)$)]

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

$L(\alpha) = \ln(p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{\alpha}, \beta)) = \ln(N(\boldsymbol{t} \mid 0, C))$

where $C = \beta^{-1}I + \sum_{j\neq i}\alpha_j^{-1}\phi_j\phi_j^T + \alpha_i^{-1}\phi_i\phi_i^T = C_{-i} + \alpha_i^{-1}\phi_i\phi_i^T$

Solution :

$\lambda(\boldsymbol{\alpha}_i) = \frac{1}{2}\left\{\ln\left(\left|\mathbf{1} + \alpha_i^{-\mathbf{1}}\phi_i^T\boldsymbol{C}_{-i}^{-\mathbf{1}}\phi_i\right|\right) - \boldsymbol{t}^T\left(\frac{C_{-i}^{-1}\phi_i\phi_i^T C_{-i}^{-1}}{\alpha_i+\phi_i^T c_{-i}^{-1}\phi_i}\right)\boldsymbol{t}\right\}$

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

- sparsity of ϕ_i
- overlap between ϕ_i and ϕ_j

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

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

$s_i > q_i \rightarrow \phi_i = 0$

$s_i < q_i \rightarrow \phi_i \neq 0$

$\frac{\partial}{\partial\alpha}L(\alpha) = \frac{\partial}{\partial\alpha}\lambda(\alpha) = \frac{\alpha_i^{-1}s_i^2+s_i-q_i^2}{(\alpha_i+s_i)^2}0$

- if $s_i \leq q_i^2 \rightarrow \alpha_i = \frac{s_i^2}{s_i-q_i^2}$
- if $s_i > q_i^2 \rightarrow \alpha_i = \infty$

4. RVM for Classification

Binary case

- $y(x) = \sigma\left(w^T\varphi(x)\right)$

Multi-class case

- $y_K(x) = \frac{\exp\left(w_K^T\varphi(x)\right)}{\sum_j \exp\left(w_j^T\varphi(x)\right)}$

Hard to calculate marginal likelihood, so use Laplace Approximation