Relation to Logistic Regression

The objective function of SVM can be written in the form:

\[ \sum_{n=1}^{N} E_{sv}(y_n t_n) + \lambda \| W \|^2 \]

\( E_{sv}(. \) is the hinge error function

\[ E_{sv} (y_n t_n) = \begin{cases} 0 & \text{if } y_n t_n \geq 1, \\ 1 - y_n t_n & \text{otherwise}. \end{cases} \]

or equivalently,

\[ E_{sv} (y_n t_n) = \begin{cases} 0 & \text{if } y_n \geq 1, \\ 1 - y_n & \text{otherwise}. \end{cases} \]

The hinge loss is non-negative and increases linearly.

Consider the sigmoid function \( \sigma (y) = \frac{1}{1+e^{-y}} \) for logistic regression. For two-class classification, we have:

\( p (t=1 | y) = \sigma (y) = \frac{1}{1+e^{-y}} \)

and \( p (t=-1 | y) = 1 - \sigma (y) = \frac{e^y}{1+e^y} = \frac{1}{1+e^y} = 1 - \sigma (-y) \).

Therefore, we can write:

\[ p (t | y) = \sigma (y t) \]

The error function consists of the negative logarithm of the likelihood function with a quadratic regularizer:

\[ \sum_{n=1}^{N} E_{lr} (y_n t_n) + \lambda \| W \|^2 \]

where:

\[ E_{lr} (y_t) = \ln (1 + \exp (-y t)) \]

Given the iid data \( D = \{(t_n, y_n), \ldots, (t_N, y_N)\} \), the likelihood function is defined by:

\[ p(D) = \prod_{n=1}^{N} \sigma (y_n t_n) \]

\[ \ln p(D) = - \sum_{n=1}^{N} \ln \frac{1}{1+e^{-y_n t_n}} = \sum_{n=1}^{N} \ln (1 + e^{-y_n t_n}) \]

The approximation made by the hinge loss leads to sparse solutions.
SVM for Regression

In linear regression, we minimize a regularized error function given by

$$
\frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2 + \frac{\lambda}{2} \| \mathbf{w} \|^2.
$$

To obtain sparse solutions, we replace the quadratic error function by an $\epsilon$-insensitive error function

$$
E_\epsilon(y, t) = \begin{cases} 
0, & \text{if } |y - t| < \epsilon, \\
|y - t| - \epsilon, & \text{otherwise}.
\end{cases}
$$

We therefore minimize the following regularized error function

$$
C \sum_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{1}{2} \| \mathbf{w} \|^2,
$$

with $y(x_n) = \mathbf{w}^T \phi(x_n) + b$.

We can re-express the optimization problem by introducing slack variables.

$$
t_n = y(x_n) + b + \delta_n,
\delta_n \geq 0,
\lambda_n \geq 0
$$

if the predicted value $y_n$ lies inside the $\epsilon$-tube, then we do not penalize it.

The optimization problem we want to solve is then

$$
\begin{align*}
\text{minimize} & \quad C \sum_{n=1}^{N} (\lambda_n + \delta_n) + \frac{1}{2} \| \mathbf{w} \|^2 \\
\text{subject to} & \quad y_n + b + \epsilon - t_n + \delta_n \geq 0, \quad \lambda_n \geq 0, \\
& \quad y_n + b - \epsilon - t_n + \delta_n \geq 0, \quad \lambda_n \geq 0, \\
& \quad n = 1, \ldots, N.
\end{align*}
$$

Introducing Lagrange multipliers $\alpha_n, \lambda_n \geq 0, \mu_n \geq 0, \rho_n \geq 0$.

The Lagrangian function

$$
L = C \sum_{n=1}^{N} (\lambda_n + \delta_n) + \frac{1}{2} \| \mathbf{w} \|^2 - \sum_{n=1}^{N} (\mu_n \lambda_n + \rho_n \delta_n) - \sum_{n=1}^{N} \alpha_n (\epsilon + \delta_n - y_n - t_n)
$$

$$
\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} (\alpha_n - \lambda_n) \phi(x_n)
$$

$$
\frac{\partial L}{\partial \delta_n} = 0 \Rightarrow \sum_{n=1}^{N} (\alpha_n - \lambda_n) = 0
$$

$$
\frac{\partial L}{\partial \lambda_n} = 0 \Rightarrow \alpha_n + \mu_n = C \quad \text{if } \lambda_n > 0
$$

$$
\frac{\partial L}{\partial \rho_n} = 0 \Rightarrow \mu_n + \rho_n = C \quad \text{if } \rho_n > 0
$$

The prediction for new inputs can be made by

$$
y(x) = \sum_{n=1}^{N} (\alpha_n - \lambda_n) \phi(x, x_n) + b.
$$
KKT conditions

\[ a_n (e + \delta_n + y_n - t_n) = 0 \]
\[ \hat{a}_n (e + \delta_n - y_n + t_n) = 0 \]
\[ (C - a_n) \delta_n = 0 \]
\[ (C - \hat{a}_n) \delta_n = 0 \]

1. \( a_n \) can only be nonzero if \( e + \delta_n + y_n - t_n = 0 \), which implies that the data point either lies on the upper boundary of the \( \epsilon \)-tube (\( \delta_n = 0 \)) or lies above the upper boundary (\( \delta_n > 0 \), \( a_n = C \)).

2. Adding \( e + \delta_n + y_n - t_n = 0 \) and \( e + \delta_n - y_n + t_n = 0 \) but \( 2e + \delta_n + \delta_n = 0 \) is impossible, so \( \epsilon \), and \( \hat{a}_n \) cannot be both nonzero.

3. All points within the \( \epsilon \)-tube have \( a_n = \hat{a}_n = 0 \)  
   \( \Rightarrow \) sparse solution

   To decide \( b \): consider a data point for which \( 0 < a_i < C \)
   \[ b = t_n - e - \mathbf{w}^T \phi (x_n) \]
   \[ = t_n - e - \sum_{m=1}^{n} (c_m - \hat{a}_m) \phi (x_n, x_m) \]

RELEVANCE VECTOR MACHINES (RVMs)

Matrix Identities
\[ (C.7) \quad x^T (x + \epsilon) \]
\[ (C.14) \quad |\mathbf{I} + \mathbf{A}^T \mathbf{A}| = |\mathbf{I} + \mathbf{A}^T \mathbf{A}| \]
\[ (C.15) \quad |\mathbf{I} + \mathbf{a}_n^T \mathbf{a}_n| = 1 + \mathbf{a}_n \]

Gaussian Identities
\[ (2.113) \quad p(x) = \mathcal{N} (x | m, \Sigma) \]
\[ (2.114) \quad p(y|x) = \mathcal{N} (y | A\mathbf{x} + b, \Sigma) \]
\[ (2.115) \quad p(y) = \mathcal{N} (y | \mu, \Sigma) \]
\[ (2.116) \quad p(y|x) = \mathcal{N} (y | A\mathbf{x} + b, \Sigma) \]
\[ Z = (\hat{a} + A^T L A)^{-1} \]

Limitations of SVMs
- Posterior probabilities?
- Two-class \( \Rightarrow \) multi-class
- \( C, \nu, \epsilon \) parameter selection by cross-validation
- Positive definite kernels.

RVM for regression

\[ t \text{ : real-valued target value} \]
\[ x \text{ : input vector} \]

Gaussian Noise:
\[ p (t | x, w, \beta) = \mathcal{N} (t | y(x), \beta^{-1}) \]
\[ y (x) = \sum_{0}^{m} w_i \phi_i (x) = \mathbf{w}^T \phi (x) \]
for RVMs, we assume the following model

$$y(x) = \sum_{n=1}^{N} \omega_n \phi(x_n) + \beta$$

as $\phi(x)$ in a linear model no restriction to p.d kernels

Use the matrix notation and assume i.i.d.

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}$$

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \end{bmatrix}$$

Next introduce the prior on $\omega$

Key: each weight parameter $\omega_i$ has a separate hyperparameter $\alpha_i$

$$p(\omega | \alpha) = \prod_{i=1}^{M} N(\omega_i | 0, \alpha_i^{-1})$$

if $\alpha_i \rightarrow \infty \Rightarrow$ high precision, zero variance

$\Rightarrow$ $\omega_i$ centered at the mean $\mu_0$

$\Rightarrow$ sparse model

from the likelihood and the prior

we can write the posterior as

$$p(\omega | \mathbf{X}, \mathbf{t}, \beta) = N(\omega | \mu, \Sigma)$$

where

$$\mu = \Sigma \{ \mu_0^T \beta \}$$

$$\Sigma = (A + \beta \Sigma^T \Sigma)^{-1}$$

$$A = \text{diag}(\alpha)$$

(2.14)

this is obtained by applying (2.113) to (2.114)

as (2.113) yields

$$p(\omega | \mathbf{X}, \mathbf{t}, \beta) = \prod_{i=1}^{M} N(\omega_i | 0, \alpha_i)$$

as (2.114) yields

$$p(t | X, \omega, \beta) = \prod_{i=1}^{N} N(t_i | \phi(X_i), \beta^{-1})$$

(Note that $\Sigma = K$ for RVM)

The optimal weights $\mathbf{w}^*$ is given by the mean of the posterior

$$\mathbf{w}^* = \mu = \beta \Sigma \mathbf{X} \mathbf{t}$$

The next step is to decide the hyperparameters $\alpha, \beta$

We use “evidence approach”

EA: Compute the marginal likelihood

$$p(t | X, \omega, \beta) = \int p(t | X, \omega, \beta) p(\omega | \alpha) d\omega$$

$$= \int p(t | \mathbf{X}, \omega, \beta) d\omega$$
Again, we don’t need to do the integration explicitly for marginalization. Just by observing and applying (2.113) \rightarrow (2.114) \rightarrow (2.115) we get

\[ p(t | \beta, \alpha, \beta) = \int p(t | \beta, \alpha, \beta) \, p(\alpha | \beta) \, d\alpha \]

\[ = N(t | 0, C) \]

\[ C = \beta^{-1} I + \bar{x} A^{-1} \bar{\phi}^T \quad A = \text{diag}(\alpha_1) \]

EA \quad maximize the logarithm of the marginal likelihood w.r.t. \alpha and \beta

\[ \ln p(t | \beta, \alpha, \beta) = \ln N(t | 0, C) \]

\[ = -\frac{1}{2} \left\{ N \ln(2\pi) + \ln |C| + t^T C^{-1} t \right\} \]

It takes a couple of pages to write down the derivations. For now we write the results only, and go into the details later.

Take the derivatives and set them to zero, we get

\[ \alpha^{new} = \frac{\bar{y}_i}{m_i^2} \]

\[ (\beta^{new})^{-1} = \frac{[t_i \bar{y}_i - \bar{y}_i \bar{m}_i]^T}{N - \sum_i \bar{x}_i \bar{m}_i} \quad \beta_i = 1 - \alpha_i \sum_i \]

SLT09 Week 12 第 10 頁