Types of Support Vector Machine

March 24, 2010

1 Overview

1. SVMs belong to kernel method family. Types of kernels: linear, poly, Gaussian (RBF), sigmoid.
2. SVMs have various versions for different tasks: classification/ regression/ clustering/ semi-supervised learning.
3. SVM uses kernel as an input for data. All that we need is a kernel matrix or a kernel function.
4. SVMs have cost functions. The cost function includes 2 components: regularization and loss components. The regularization can be l_1 or l_2 norm.
5. SVC (SVM for classification) assumptions: data can be linearly separated in feature space.
6. SVC theory: The best separating hyperplane is the one with maximum margin.
7. SVC constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training datapoints of any class, since in general the larger the margin the lower the generalization error of the classifier.
8. SVC has primal form and dual form.
9. The dual form of SVC is a quadratic form.
10. SVC optimization problem is a quadratic optimization problem.
11. For special formulation of SVC quadratic optimization, many optimization techniques are applied to improve speed, efficiency of SVM optimization process.
12. SVC decision rules (decision function)

$$f(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b)$$

13. SVM model:

2 Hard Margin SVC

The assumption of Hard Margin SVC: data is linearly separate in feature space. It gives no solution if this assumption is wrong.

Primal form

In case the instances are linearly separable, the hyperlane $(\mathbf{w}^*, b^*)$ that solves the optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \langle \mathbf{w}.\mathbf{w} \rangle \\ \textit{subject to} \quad & y_i(\langle \mathbf{w}.\Phi(\mathbf{x}_i) \rangle + b) \geq 1, i = 1, ..., n \end{aligned} \tag{1}$$

Dual form

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \textit{subject to} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1, ..., n. \end{aligned} \tag{2}$$

For Soft Margin SVC, both L_1 and L_2 types, we do not have to assume that data is linearly separate in feature space. The error rate which can be tolerated is reflected in parameter $C, C \in [0, \infty])$

3 L_1 Soft Margin SVC

Primal form

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \langle \mathbf{w}.\mathbf{w} \rangle + C \sum_{i=1}^{n} \xi_i \\ \textit{subject to} \quad & y_i(\langle \mathbf{w}.\Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, i = 1, ..., n \\ & \xi_i \geq 0, i = 1, ..., n \end{aligned} \tag{3}$$

Dual form

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \textit{subject to} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0, C \geq \alpha_i \geq 0, i = 1, ..., n. \end{aligned} \tag{4}$$

4 L_2 Soft Margin SVC

L_2 Sorf Margin SVC can be consider as Hard Margin SVC, with a small change of kernel: adding $\frac{1}{C}$ to all diagonal elements of the kernel matrix. *Primal form*

$$\begin{aligned} \min_{\xi,\mathbf{w},b} \quad & \langle \mathbf{w}.\mathbf{w} \rangle + C \sum_{i=1}^{n} \xi_i^2 \\ \textit{subject to} \quad & y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, i = 1, ..., n \\ & \xi_i \geq 0, i = 1, ..., n \end{aligned} \tag{5}$$

Dual form

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C} \delta_{ij}) \\ \textit{subject to} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1, ..., n. \end{aligned} \tag{6}$$

5 ν-SVC

Primal form:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi} \in \mathcal{R}^n, \rho, b \in \mathcal{R}} \quad & \frac{1}{2} \|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{n} \sum_{i=1}^{n} \xi_i \\ s.t \quad & y_i(\langle \boldsymbol{x}, \boldsymbol{w} \rangle + b) \geq \rho - \xi_i, i = 1, ..., n \\ & \xi_i \geq 0, \rho \geq 0 \end{aligned} \tag{7}$$

Dual form:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathcal{R}^n} W(\boldsymbol{\alpha}) \quad & = -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ s.t \quad & 0 \leq \alpha_i \leq \frac{1}{n} \\ & \sum_{i=1}^{n} \alpha_i y_i = 0, \sum_{i=1}^{n} \alpha_i \geq \nu \end{aligned} \tag{8}$$

If ν-SVM classification leads to $\rho > 0$, then C-SVM (SVM using C parameter) classification, with C set a priori to $1/n\rho$, leads to the same decision function.

ν has range $[1, 0]$ while range of C is $[0, \infty]$

$\nu_{max} = 2min(n_+, n)/n$,

For any $\nu > \nu_{max}$, the dual ν-SVM is infeasible. That is, the set of feasible points is empty. For any $\nu \in (\nu_{min}, \nu_{max}]$, the optimal solution set of dual ν-SVM is the same as that of either one or some C-SVM where these C form an interval. In addition, the optimal objective value of ν-SVM is strictly positive. For any $0 \leq \nu \leq \nu_{min}$, dual ν-SVM is feasible with zero optimal objective value. If the kernel matrix is positive definite, then $\nu_{min} = 0$.

$$0 \leq \nu_{min} \leq \nu_{max} \leq 1$$

ν is a lower bound on the sum of α_i. The proportion of the training set that are margin errors is upper bounded by ν, while ν provides a lower bound on the total number of support vectors.

6 Advanced SVM

1. SVM for regression
2. SVM for semi-supervised learning: S3VM (or Transductive SVM)
3. SVM for unsupervised learning (clustering) (one-class)
4. SVM for multiclass classifcation
5. SVM modification
6. l_1 SVM (w are regularized by l_1 norm)
7. Structured SVM: for general structure output labels.
8. Multiple Kernel Learning
9. ...