Computational Intelligence: Methods and Applications

Lecture 24

SVM in the non-linear case

Włodzisław Duch

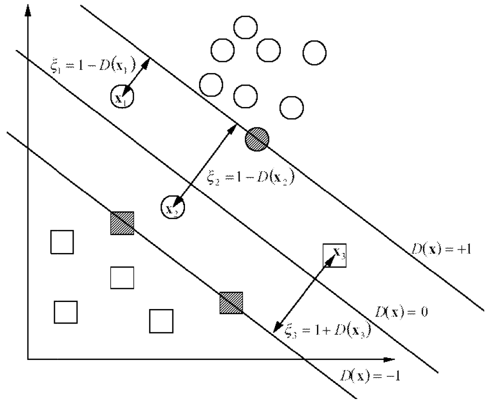
SCE, NTU, Singapore

Google: Duch

Non-separable picture

Unfortunately for non-separable data vectors not all conditions may be fulfilled, some data points are not outside of the two hyperplanes: new "slack" scalar variables are introduced in separability conditions.

Margin between the two distributions of points is defined as the distance between the two hyperplanes parallel to the decision border; it is still valid for most data, but there are now two points on the wrong side of the decision hyperplane, and one point inside the margin.



Non-separable case

The problem becomes slightly more difficult, since the quadratic optimization problem is not convex, saddle points appear.

Conditions:

$$g_{\mathbf{W}}\left(\mathbf{X}^{(i)}\right) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{(i)} + W_0 \geq +1 - \xi_i \quad \text{for } Y^{(i)} = +1$$

$$g_{\mathbf{W}}\left(\mathbf{X}^{(i)}\right) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{(i)} + W_0 \leq -1 + \xi_i \quad \text{for } Y^{(i)} = -1 \text{ and } \xi_i \geq 0$$

If $\xi_i > 1$ then the point is on the wrong side of the $g(\mathbf{X})>0$ plane, and is misclassified.

Dilemma: reduce the number of misclassified points, or keep large classification margin hoping for better generalization on future data, despite some errors made now. This is expressed by minimizing:

$$\frac{1}{2}\|\mathbf{W}\|^2 + C\sum_{i=1}^{n}\xi_i$$

adding a user-defined parameter *C* and leading to the same solution as before, with bound on α,

$0 \leq \alpha_i \leq C$ smaller *C* = larger margins (see Webb Chap. 4.2.5)

SVM: non-separable

Non-separable case conditions, using slack variables:

$$g_{\mathbf{W}}\left(\mathbf{X}\right) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{(i)} + W_0 \geq +1 - \xi_i \quad \text{for } Y^{(i)} = +1$$

$$g_{\mathbf{W}}\left(\mathbf{X}\right) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{(i)} + W_0 \leq -1 + \xi_i \quad \text{for } Y^{(i)} = -1 \text{ and } \xi_i \geq 0$$

Lagrangian with penalty for errors scaled by *C* coefficient:

$$L(\mathbf{W}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{W}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[Y^{(i)}\left(\mathbf{W}^{\mathrm{T}}\cdot\mathbf{X}^{(i)} + W_0\right) - 1\right], \alpha_i \geq 0$$

Min W, max $\boldsymbol{\alpha}$. Discriminant function with regularization conditions:

$$g(\mathbf{X}) = \mathbf{W}^{\mathrm{T}}\cdot\mathbf{X} + W_0 = \sum_{i=1}^{n}\alpha_i Y^{(i)}\mathbf{X}^{(i)\mathrm{T}}\cdot\mathbf{X} + W_0 \qquad 0 \leq \alpha_i \leq C$$

Coefficients α are obtained from the quadratic programming problem and $W_0 = Y^{(i)} - \mathbf{W}^{\mathrm{T}}\cdot\mathbf{X}^{(i)}$ from support vectors $Y^{(i)}g(\mathbf{X}^{(i)})=1$.

Support Vectors (SV)

Some α have to be non zero, otherwise classification conditions $Y^{(i)}g(\mathbf{X}^{(i)})-1>0$ will not be fulfilled and discriminating function will be reduce to W_0. The term known as the KKT sum (from Karush-Kuhn-Tacker, who used it in optimization theory) :

$$L_c\left(\mathbf{W},\boldsymbol{\alpha}\right)=-\sum_{i=1}^{n}\alpha_i\left[Y^{(i)}\mathbf{W}^{\mathrm{T}}\cdot\mathbf{X}^{(i)}-1\right],\ \alpha_i\geq 0$$

is large and positive for misclassified vectors, and therefore vectors near the border $g(\mathbf{X}^{(i)})=Y^{(i)}$ should have non zero α_i to influence W. This term is negative for correctly classified vectors, far from the H_i hyperplanes; selecting $\alpha_i=0$ will maximize the Lagrangian $L(\mathrm{W},\alpha)$.

The dual form with α is easier to use, it is maximized with one additional equality constraint:

$$L(\boldsymbol{\alpha})=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\alpha_i Y^{(i)}\sum_{j=1}^{n}\alpha_j Y^{(j)}\mathbf{X}^{(i)}\cdot\mathbf{X}^{(j)}$$

$$\sum_{i=1}^{n}\alpha_i Y^{(i)}=0;\quad 0\leq\alpha_i\leq C;\quad i=1..n$$

Mechanical analogy

Mechanical analogy: imagine the $g(\mathbf{X})=0$ hyperplane as a membrane, and SV $\mathbf{X}^{(i)}$ exerting force on it, in the $Y^{(i)}\mathbf{W}$ direction.
Stability conditions require forces to sum to zero leading to:

$$\mathbf{F}_i=\alpha_i Y^{(i)}\frac{\mathbf{W}}{\|\mathbf{W}\|}\quad i=1..n_{\mathrm{sv}}\quad \alpha_i\geq 0$$

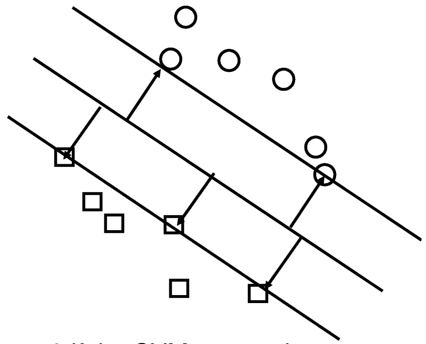
$$\sum_{i=1}^{n_{\mathrm{sv}}}\mathbf{F}_i=\frac{\mathbf{W}}{\|\mathbf{W}\|}\sum_{i=1}^{n_{\mathrm{sv}}}\alpha_i Y^{(i)}=0$$

Same as auxiliary SVM condition.

Also all torques should sum to 0

$$\sum_{i=1}^{n_{\mathrm{sv}}}\mathbf{X}^{(i)}\times\mathbf{F}_i=\sum_{i=1}^{n_{\mathrm{sv}}}\alpha_i Y^{(i)}\mathbf{X}^{(i)}\times\frac{\mathbf{W}}{\|\mathbf{W}\|}$$

$$=\mathbf{W}\times\frac{\mathbf{W}}{\|\mathbf{W}\|}=0$$



Sum=0 if the SVM expression for W is used.

Sequential Minimal Optimization

SMO: solve smallest possible optimization step (J. Platt, Microsoft).

Idea similar to the Jacobi rotation method with 2x2 rotations, but here applied to the quadratic optimization.

Valid solution for min $L(\boldsymbol{\alpha})$ is obtained when all conditions are fulfilled:

Complexity: problem size n^2,
solution complexity n_{sv}^2.

$$\alpha_i=0\Leftrightarrow Y^{(i)}g\left(\mathbf{X}^{(i)}\right)>1$$

$$0<\alpha_i<C\Leftrightarrow Y^{(i)}g\left(\mathbf{X}^{(i)}\right)=1$$

$$\alpha_i=C\Leftrightarrow Y^{(i)}g\left(\mathbf{X}^{(i)}\right)<1$$

ε- accuracy to which conditions should be fulfilled (typically 0.001)

SMO: find all examples $\mathbf{X}^{(i)}$ that violate these conditions;
select those that are neither 0 nor C (non-bound cases).

take a pair of α_i, α_j and find analytically the values that minimize their contribution to the $L(\boldsymbol{\alpha})$ Lagrangian.

Examples of linear SVM

SVM SMO, Sequential Multiple Optimization, is implemented in WEKA with linear and polynomial kernels.

The only user adjustable parameter for linear version is C; for non-linear version the polynomial degree may also be set.

In the GhostMiner 1.5 optimal value of C may automatically be found by crossvalidation training.

For non-linear version type of kernel function and parameters of kernels may be adjusted (GM).

Many free software packages/papers are at: www.kernel-machines.org

Example 1: Gaussians data clusters

Example 2: Cleveland Heart data

Examples of linear SVM

Examples of mixture data with overlapping classes; the Bayesian non-linear decision borders, and linear SVM with margins are shown.

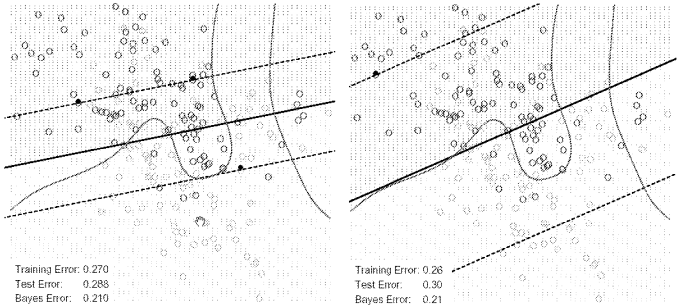


Fig. 12.2, from Hasti et. al 2001 data from mixture of Gaussians

With C=10000, with C=0.01, larger margin

Errors seem to be reversed here! Large C is better around decision plane, but not worse overall (the model is too simple), so it should have lower training error but higher test; for small C margin is large, training error slightly larger but test lower.

Letter recognition

Categorization of text samples. Set different rejection rate and calculate Recall = $P_{+|+} = P_{++} / P_{+}$ and Precision= $P_{++}/(P_{++} + P_{-+}) = TP/(TP+FP)$