Support Vector Machines (SVMs)
Part 4: Non-separable data

Yingming Li
yingming@zju.edu.cn

Data Science & Engineering Research Center, ZJU

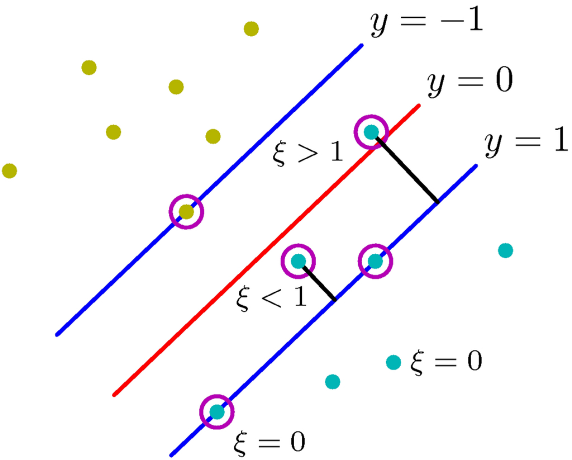
16th April 2018

Adapted from slides provided by Prof. Michael Mandel.

What if the classes overlap?

- Allow mis-classifications, but penalize them
  - in proportion to distance on the wrong side of the margin
  - Add to existing cost, minimize sum of the two
- Introduce "slack variables" $\xi_p \geqslant 0$
  - one per training point
  - $\xi_p = \max(1 - d_p y(\mathbf{x}_p), 0)$
- Interpretation
  - $\xi_p = 0$ for points on the correct side of the margin
  - $0 < \xi_p < 1$ for correctly classified points within margin
  - $\xi_p > 1$ for mis-classified points

Meaning of ξ_p



Incorporate slack variables in optimization

- New problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\operatorname{argmin}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_p \xi_p \\ \text{s.t.} \quad & d_p y(\mathbf{x}_p) \geqslant 1 - \xi_p \\ & \xi_p \geqslant 0 \end{aligned}$$

- So constraint $d_p y(\mathbf{x}_p) \geqslant 1$ has been relaxed
- But now minimize the sum of the ξ_p too
- C controls trade-off between margin and slack
    - As $C \to \infty$, return to SVM for separable data

New primal Lagrangian adds two new terms

- Primal Lagrangian (still QP with linear constraints):

$$\begin{aligned} L(\mathbf{w}, b, \mathbf{a}, \mu) &= \frac{1}{2}||\mathbf{w}||^2 + C\sum_p \xi_p - \sum_p \mu_p \xi_p \\ &- \sum_p a_p(d_p(\mathbf{w}^t\mathbf{x}_p + b) - 1 + \xi_p) \end{aligned}$$

- KKT conditions:

$$\begin{array}{rr} a_p \geqslant 0 & \xi_p \geqslant 0 \\ d_p y(\mathbf{x}_p) - 1 + \xi_p \geqslant 0 & \mu_p \geqslant 0 \\ a_p(d_p y(\mathbf{x}_p) - 1 + \xi_p) = 0 & \mu_p \xi_p = 0 \end{array}$$

Derive dual Lagrangian by solving for $\mathbf{w}$, b, ξ

- The matrix

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_p a_p d_p \mathbf{x}_p \quad \text{Unchanged}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow b = \sum_p a_p d_p = 0 \quad \text{Unchanged}$$

$$\frac{\partial L}{\partial \xi_p} = 0 \Rightarrow a_p = C - \mu_p \quad \text{New}$$

- So $\boldsymbol{\mu}$ can be replaceed by $\mathbf{a}$

New dual Lagrangian changed very little

- Dual Lagrangian

$$\tilde{L}(\mathbf{a}) = \sum_p a_p - \frac{1}{2} \sum_p \sum_q a_p a_q d_p d_q \mathbf{x}_p^T \mathbf{x}_q$$

- With constraints

$$0 \leqslant a_p \leqslant C \quad \sum_p a_p d_p = 0$$

- Only difference is upper bound on a_p from $\mu_p \geqslant 0$
- Still a quadratic program with linear constraints
- Predictions still made identically

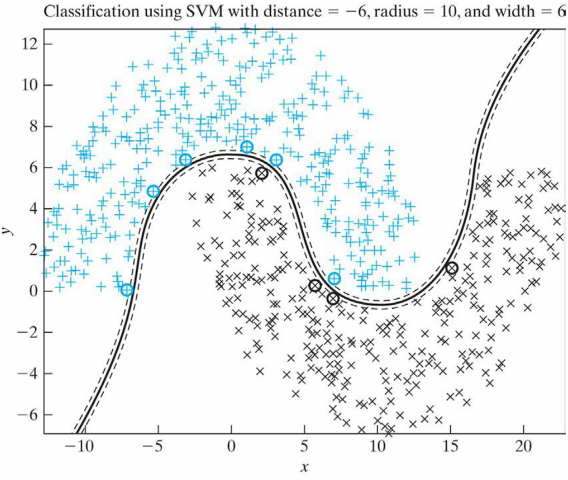
Now many types of points

- Points with $a_p = 0$ are still non-support vectors
  - Do not contribute to classification
- Points with $a_p > 0$
  - Must satisfy KKT condition $d_p y(x_p) = 1 - \xi_p$
  - Points with $0 < a_p < C$ have margin 1
    - KKT condition that $\xi_p = 0$
  - Points with $a_p = C$ can lie inside the margin
    - Correctly classfied if $\xi_p \leqslant 1$
    - Incorrectly classified if $\xi_p > 1$

Remarks on points with $a_p = C$

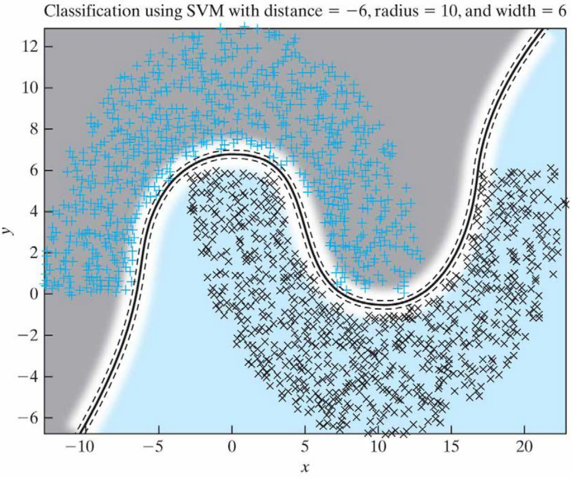
- It is undesirable that these points are support vectors
- All misclassified training points must be SVs
- Makes decisions sensitive to outliers in training
- Need to evaluate kernel on them at test time

SVM double-moon training set, $d = -6$



(a) Training result

SVM double-moon test set, $d = -6$



(b) Testing result

Thank you!