CSE 5526: Introduction to Neural Networks

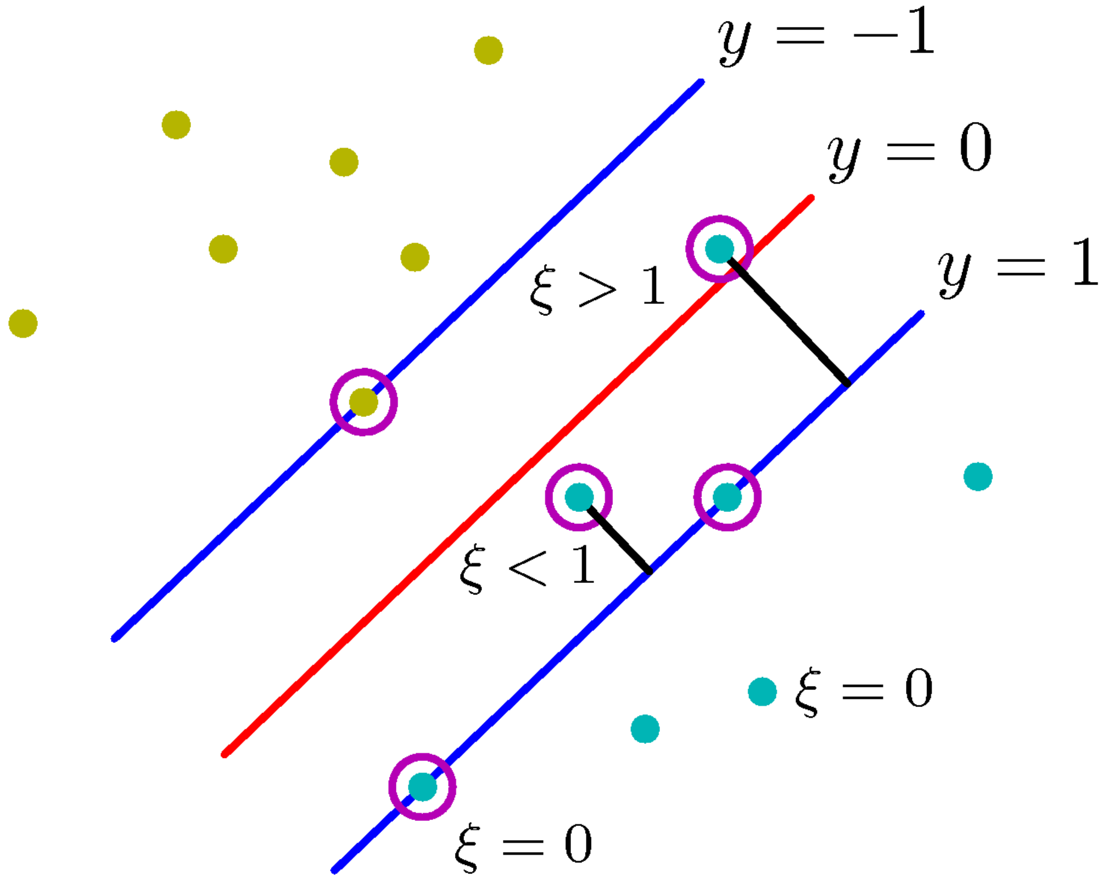
Support Vector Machines (SVMs)

Part 4: Non-separable data

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” $\xi_p \geq 0$
  - one per training point
  - $\xi_p = \max\left(1 - d_p y\left(\boldsymbol{x}_p\right), 0\right)$
- 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:

$$\text{argmin}_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_p \xi_p$$

$$\text{Subject to } \; d_p y(\boldsymbol{x}_p) \geq 1 - \xi_p$$

- So constraint $d_p y(\boldsymbol{x}_p) \geq 1$ has been relaxed
- But now minimize the sum of the ξ_ps 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):

$$L(\boldsymbol{w}, b, \boldsymbol{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_p \xi_p - \sum_p \mu_p \xi_p + \sum_p a_p\left(d_p\left(\boldsymbol{w}^T\boldsymbol{x}_p + b\right) - 1 + \xi_p\right)$$

- KKT conditions:

$$a_p \geq 0 \qquad \xi_p \geq 0$$

$$d_p y\left(\boldsymbol{x}_p\right) - 1 + \xi_p \geq 0 \qquad \mu_p \geq 0$$

$$a_p\left(d_p y\left(\boldsymbol{x}_p\right) - 1 + \xi_p\right) = 0 \qquad \mu_p \xi_p = 0$$

Derive dual Lagrangian by solving for $\boldsymbol{w}$, b, $\boldsymbol{\xi}$

- Set gradient of Lagrangian to 0

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

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

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

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

New dual Lagrangian changed very little

- Dual Lagrangian

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

- With constraints

$$0 \le a_p \le C \qquad \sum_p a_p d_p = 0$$

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

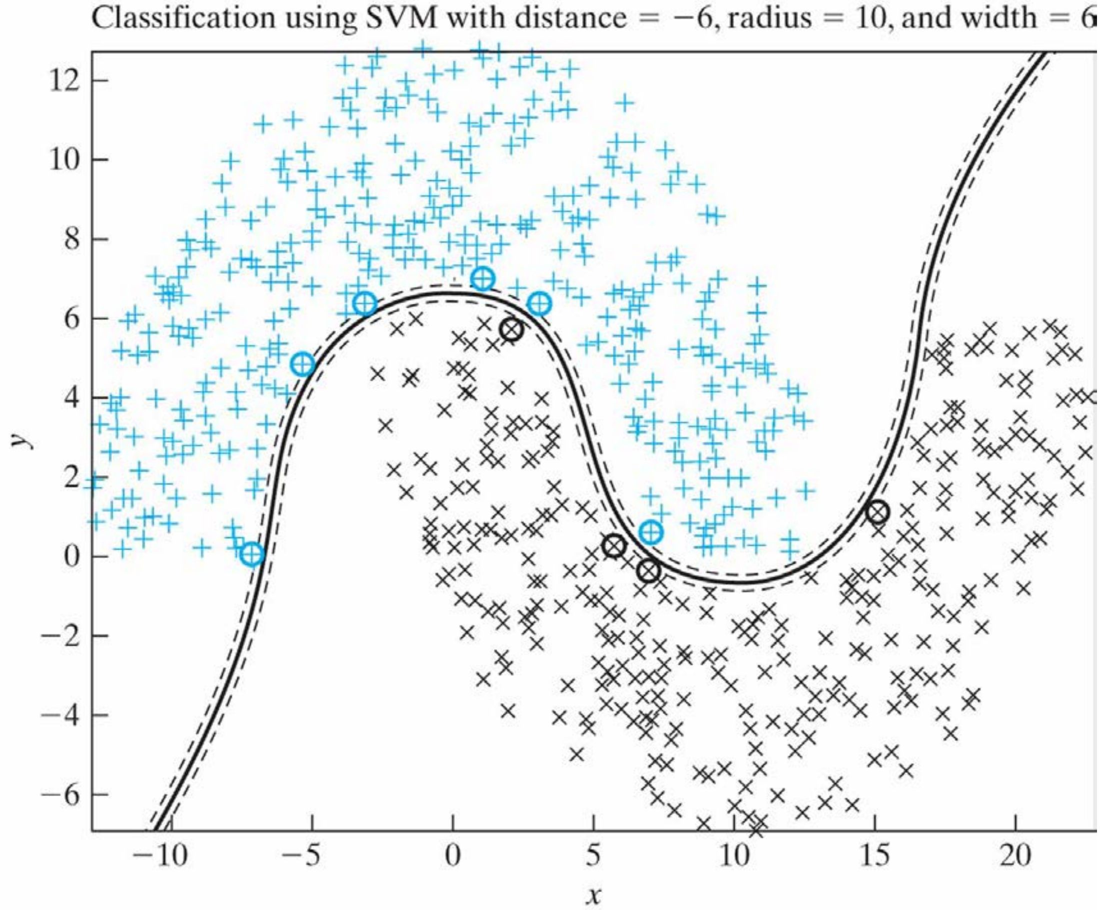
Now many types of points

- 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 classified if $\xi_p \leq 1$
    - Incorrectly classified if $\xi_p > 1$

Remarks on points with $a_p = C$

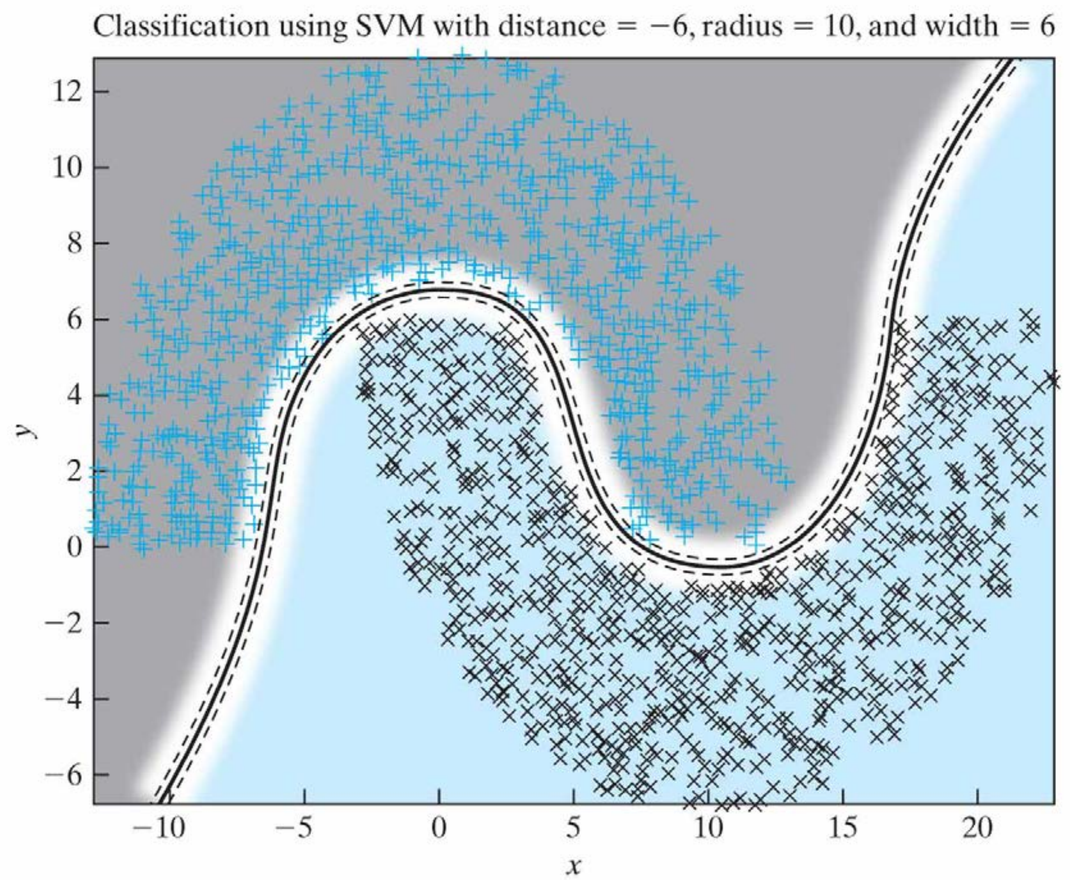
- It is undesirable that these points are support vectors
- All misclassified training points must be SVs
- Makes decisions sensitive to outliers in training
- Need to evaluate kernel on them at test time

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



(a) Training result

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



(b) Testing result