Support Vector Machine

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Dual SVM with Kernel Trick

- $max_{\alpha \geq 0} \sum_j \alpha_j - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \varphi(x_i) \varphi(x_j)$
- $max_{\alpha \geq 0} \sum_j \alpha_j - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
  - $\alpha_i \left((wx_j + b) y_j - 1 \right) = 0, C > \alpha_i > 0$
    - $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i)$
    - $b = y_j - \sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i) \, \varphi(x_j)$
  - $\sum_{i=1}^{N} \alpha_i y_i = 0$
  - $C \geq \alpha_i \geq 0, \forall i$
- Dual formulation lets SVM utilize
  - Kernel trick
    - Reduced parameters to estimate
  - Only store alpha values instead of w
    - How many alpha values are needed?
    - Consider meaningful alphas

Dual Problem of Linearly Separable SVM

$$max_{\alpha \geq 0} \sum_j \alpha_j - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$$

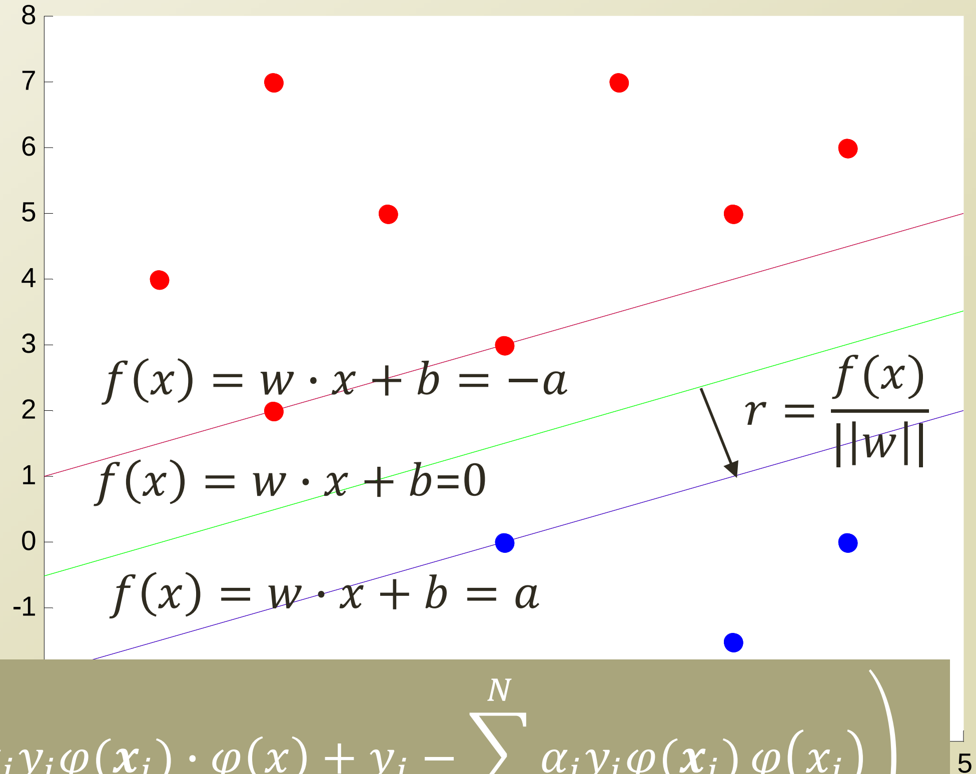
$$C \geq \alpha_i \geq 0, \forall i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \left((wx_j + b) y_j - 1 \right) = 0, C > \alpha_i > 0$$

Classification with SVM Kernel Trick

- Linear case
  - $sign(w \cdot x + b)$
  - $min_{w,b} ||w||$
    - $(wx_j + b)y_j \geq 1, \forall j$
- Transformed case
  - $sign(w \cdot \varphi(x) + b)$
  - $min_{w,b,\xi_j} ||w|| + C \sum_j \xi_j$
    - $(w\varphi(x_j) + b)y_j \geq 1 - \xi_j, \forall j$
    - $\xi_j \geq 0, \forall j$
- Kernel trick case
  - $sign(w \cdot \varphi(x) + b)$
  - $max_{\alpha \geq 0} \sum_j \alpha_j - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
    - $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i)$
    - $b = y_j - w\varphi(x_j)\ when 0 < \alpha_j < C$
    - $\sum_{i=1}^{N} \alpha_i y_i = 0$
    - $0 \leq \alpha_i \leq C, \forall i$



$$sign(w \cdot \varphi(x) + b) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i) \cdot \varphi(x) + y_j - \sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i)\, \varphi(x_j)\right)$$

$$= sign\left(\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, x) + y_j - \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, x_j)\right)$$

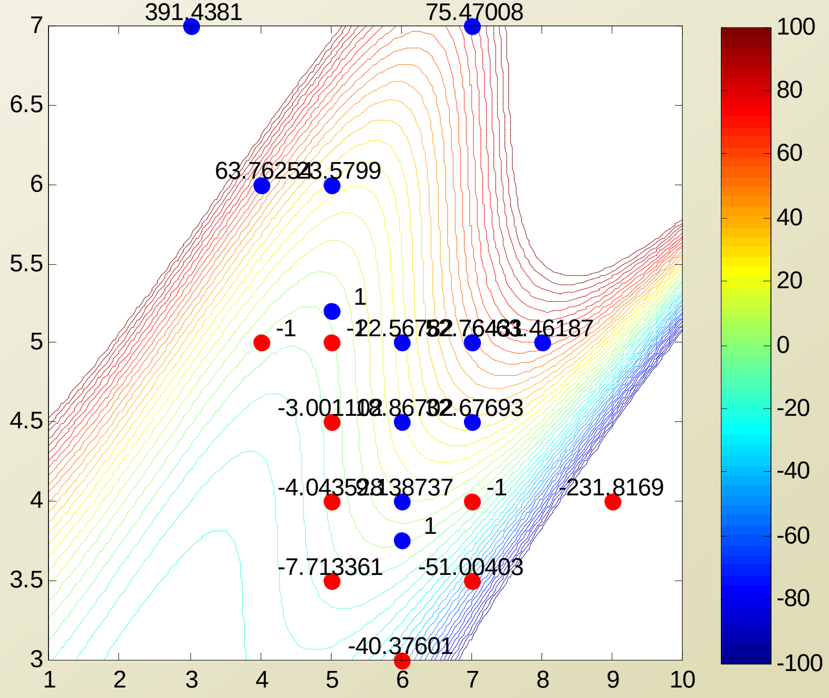
$0 < \alpha_j < C$

KAIST

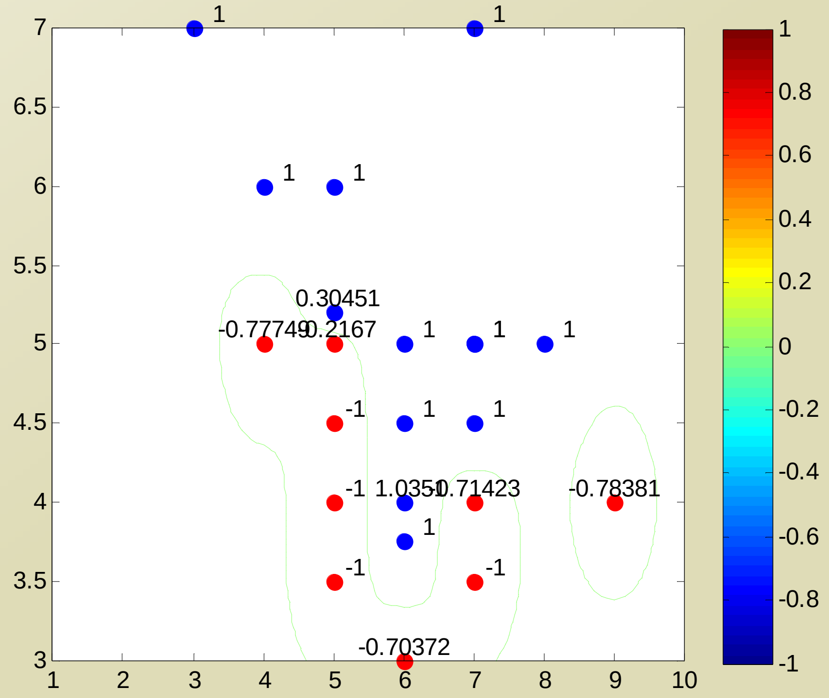
SVM with Various Kernels

- SVM is very adaptable to the non-linearly separable cases with the kernel trick
  - Easy expand to the high dimension features (for free!)

Polynomial Kernel with Degree 4



RBF Kernel with γ=4



Logistic Regression with Kernel

- Logistic regression
  - $P(Y|X) = \frac{1}{1+e^{-\dot{\theta}^T x}}$
  - Finding the MLE of θ
- Can we kernelize the logistic regression?
  - $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \varphi(\boldsymbol{x}_i)$
  - $P(Y|X) = \frac{1}{1+e^{-\dot{\theta}^T x}} = \frac{1}{1+e^{\sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)\varphi(x)+b}} = \frac{1}{1+e^{\sum_{i=1}^{N} \alpha_i y_i K(x_i,x)+b}}$
- Problem changes
  - From finding θ to finding α_i
  - How to solve this problem?
    - In other words…
    - Is this a constrained optimization?
    - If not, what does it imply?

Acknowledgement

- This slideset is greatly influenced
  - By Prof. Carlos Guestrin at CMU
  - By Prof. Eric Xing at CMU

Further Readings

- Bishop Chapter 7 → 6