CS578 Statistical Machine Learning Lecture 4

Jean Honorio
Purdue University

(based on slides by Tommi Jaakkola, MIT CSAIL)

Today's topics

- Perceptron solution and kernels
- Support vector machine with kernels
  - dual solution, with offset, slack

- Today midnight: Homework 1, due Jan 30, 11.59pm EST

- Office hours: Doodle poll closes Sunday

- If Homework 0 grade less than or equal to 6.5
  - In the past, students got C, D or F
  - **Advise**: prepare on linear algebra, statistics, take course next semester

The perceptron solution

- Suppose the training set is linearly separable through origin given a particular feature mapping, i.e.,

$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) > 0, \; i = 1, \ldots, n$$

  for some $\underline{\theta}$

- The perceptron algorithm, applied repeatedly over the training set, will find a solution of the form

$$\underline{\theta} = \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i), \quad \alpha_i \in \{0, 1, \ldots\}$$

the number of mistakes made on
the ith training example until convergence

- We can recast the algorithm entirely in terms of these "mistake counts" α_i

Kernel perceptron

- We don't need the parameters nor the feature vectors explicitly
- All we need for predictions as well as updates is the value of the discriminant function

$$\underline{\theta} \cdot \underline{\phi}(\underline{x}) = \sum_{i=1}^{n} \alpha_i y_i [\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x})] = \sum_{i=1}^{n} \alpha_i y_i \underbrace{K(\underline{x}_i, \underline{x})}_{\text{kernel}}$$

Initialize: $\alpha_i = 0,\ i = 1, \ldots, n$

Repeat until convergence:

$$\text{for } t = 1, \ldots, n$$

$$\text{if } y_t \left(\sum_{i=1}^{n} \alpha_i y_i K(\underline{x}_i, \underline{x}_t) \right) \leq 0 \ \text{(mistake)}$$

$$\alpha_t \leftarrow \alpha_t + 1$$

value of the discriminant function prior to the update

Kernels

- By writing the algorithm in a "kernel" form, we can try to work with the kernel (inner product) directly rather than explicating the high dimensional feature vectors

$$\begin{aligned} K(\underline{x}, \underline{x}') &= \underline{\phi}(\underline{x}) \cdot \underline{\phi}(\underline{x}') \\ &= \begin{bmatrix} \\ ? \\ \\ \end{bmatrix} \cdot \begin{bmatrix} \\ ? \\ \\ \end{bmatrix} \\ &= \exp(-\|\underline{x} - \underline{x}'\|^2) \quad \text{(e.g.)} \end{aligned}$$

- All we need to ensure is that the kernel is "valid", i.e., there exists some underlying feature representation

Valid kernels

- A kernel function is valid (is a kernel) if there exists some feature mapping such that

$$K(\underline{x}, \underline{x}') = \phi(\underline{x}) \cdot \phi(\underline{x}')$$

- Equivalently, a kernel is valid if it is symmetric and for all training sets, the Gram matrix

$$\begin{bmatrix} K(\underline{x}_1, \underline{x}_1) & \cdots & K(\underline{x}_1, \underline{x}_n) \\ \cdots & \cdots & \cdots \\ K(\underline{x}_n, \underline{x}_1) & \cdots & K(\underline{x}_n, \underline{x}_n) \end{bmatrix}$$

is positive semi-definite

Primal SVM

- Consider a simple max-margin classifier through origin

$$\text{minimize } \frac{1}{2}\|\underline{\theta}\|^2 \text{ subject to}$$
$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \quad i = 1, \ldots, n$$

- We claim that the solution has the same form as in the perceptron case

$$\underline{\theta}(\alpha) = \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i), \quad \alpha_i \geq 0$$

non-negative Lagrange multipliers used to enforce the classification constraints

- As before, we focus on estimating α_i which are now non-negative real numbers

Primal SVM

- Consider a simple max-margin classifier through origin

$$\text{minimize} \quad \frac{1}{2}\,||\underline{\theta}||^2 \quad \text{subject to}$$

$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \quad i = 1, \ldots, n$$

- To solve this, we can introduce Lagrange multipliers for the classification constraints and minimize the resulting Lagrangian with respect to the parameters $\underline{\theta}$

$$L(\underline{\theta}, \alpha) = \frac{1}{2}\,||\underline{\theta}||^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

Primal SVM

- Consider a simple max-margin classifier through origin

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{subject to}$$

$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \quad i = 1, \ldots, n$$

- To solve this, we can introduce Lagrange multipliers for the classification constraints and minimize the resulting Lagrangian with respect to the parameters $\underline{\theta}$

should become non-negative

$$L(\underline{\theta}, \alpha) = \frac{1}{2}\|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

positive values enforce classification constraints

Understanding the Lagrangian

$$L(\underline{\theta}, \alpha) = \frac{1}{2}\|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right]$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

- To **maximize** $L(\underline{\theta}, \alpha)$ with respect to α :
  - Assume $y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1$, for instance:

$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) = 3, \qquad \begin{array}{ll} \alpha_i = 10, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = -20 \\ \alpha_i = 5, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = -10 \\ \alpha_i = 0, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = 0 \end{array}$$

  - Assume $y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) < 1$, for instance:

$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) = -1, \qquad \begin{array}{ll} \alpha_i = 10, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = 20 \\ \alpha_i = 20, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = 40 \\ \alpha_i = \infty, & -\alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right] = \infty \end{array}$$

Understanding the Lagrangian

$$L(\underline{\theta}, \alpha) = \frac{1}{2} \|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

- We can recover the primal problem by maximizing the Lagrangian with respect to the Lagrange multipliers

to maximize:
since $y_i(\underline{\theta} \bullet \underline{\phi}(\underline{x}_i)) \geq 1$, for all i
make $\alpha_i = 0$

$$\max_{\alpha \geq 0} L(\underline{\theta}, \alpha) = \begin{cases} \frac{1}{2}\|\underline{\theta}\|^2, & \text{if } y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \ i = 1, \ldots, n \\ \infty, & \text{otherwise} \end{cases}$$

to maximize:
since $y_i(\underline{\theta} \bullet \underline{\phi}(\underline{x}_i)) < 1$, for some i
make $\alpha_i = \infty$

Note: to **minimize** { $\max_{\alpha \geq 0} L(\theta, \alpha)$ } with respect to θ, we should fulfill constraints, to avoid ∞

Primal - Dual

minimize $\frac{1}{2}||\underline{\theta}||^2$ subject to
$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \quad i = 1, \ldots, n$

$$\min_{\underline{\theta}} \overbrace{\left[\max_{\alpha \geq 0} L(\underline{\theta}, \alpha) \right]}^{\text{primal}(\underline{\theta})}$$

- expressed in terms of $\underline{\theta}$
- explicit feature vectors $\underline{\phi}(\underline{x})$

?

Primal - Dual

(See Slater conditions in [1] and [2] if interested)

$$\text{minimize } \frac{1}{2}||\underline{\theta}||^2 \text{ subject to}$$
$$y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) \geq 1, \quad i = 1, \ldots, n$$

?

$$\min_{\underline{\theta}} \overbrace{\left[\max_{\alpha \geq 0} L(\underline{\theta}, \alpha)\right]}^{\text{primal}(\underline{\theta})} = \underbrace{\max_{\alpha \geq 0}}_{\text{step 2}} \underbrace{\overbrace{\left[\min_{\underline{\theta}} L(\underline{\theta}, \alpha)\right]}^{\text{dual}(\alpha)}}_{\text{step 1}}$$

- expressed in terms of $\underline{\theta}$
- explicit feature vectors $\underline{\phi}(\underline{x})$

- expressed in terms of α
- kernels $K(\underline{x}, \underline{x}')$

Lagrangian Dual (step 1)

$$L(\underline{\theta}, \alpha) = \frac{1}{2} \|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}, \alpha) \qquad = 0$$

Lagrangian Dual (step 1)

$$L(\underline{\theta}, \alpha) = \frac{1}{2}\left\|\underline{\theta}\right\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1\right]$$

$$\frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}, \alpha) = \underline{\theta} - \qquad\qquad\qquad\qquad = 0$$

Lagrangian Dual (step 1)

$$L(\underline{\theta}, \alpha) = \frac{1}{2} \|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i (\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}, \alpha) = \underline{\theta} - \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i) = 0$$

Lagrangian Dual (step 1)

$$L(\underline{\theta}, \alpha) = \frac{1}{2} \|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}, \alpha) = \underline{\theta} - \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i) = 0$$

$$\Rightarrow \underline{\theta} = \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i) = \underline{\theta}(\alpha)$$

Lagrangian Dual (step 1)

$$L(\underline{\theta}, \alpha) = \frac{1}{2}\|\underline{\theta}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\underline{\theta} \cdot \underline{\phi}(\underline{x}_i)) - 1 \right]$$

$$\frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}, \alpha) = \underline{\theta} - \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i) = 0$$

$$\Rightarrow \underline{\theta} = \sum_{i=1}^{n} \alpha_i \, y_i \underline{\phi}(\underline{x}_i) = \underline{\theta}(\alpha)$$

- The dual problem is obtained by substituting this solution back into the Lagrangian and recalling that the Lagrange multipliers are non-negative

$$\underline{\text{maximize}} \quad \text{dual}(\alpha) = \min_{\underline{\theta}} L(\underline{\theta}, \alpha) = L(\underline{\theta}(\alpha), \alpha)$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

Dual SVM (step 2)

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

- This is again a quadratic programming problem but with simpler "box" constraints

Dual SVM (step 2)

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

- This is again a quadratic programming problem but with simpler "box" constraints

$$\underline{\theta}(\alpha^*) = \sum_{i=1}^{n} \alpha_i^* y_i \underline{\phi}(\underline{x}_i)$$

Dual SVM (step 2)

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

- This is again a quadratic programming problem but with simpler "box" constraints

$$\underline{\theta}(\alpha^*) = \sum_{i=1}^{n} \alpha_i^* y_i \underline{\phi}(\underline{x}_i)$$

$$\text{if } \alpha_i^* > 0 \quad \Rightarrow \quad y_i(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i)) = 1$$

(support vector)

Dual SVM (step 2)

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

- This is again a quadratic programming problem but with simpler "box" constraints

$$\underline{\theta}(\alpha^*) = \sum_{i=1}^{n} \alpha_i^* y_i \underline{\phi}(\underline{x}_i)$$

$$\text{if } \alpha_i^* > 0 \quad \Rightarrow \quad y_i(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i)) = 1 \qquad \text{(support vector)}$$

$$\text{if } \alpha_i^* = 0 \quad \Rightarrow \quad y_i(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i)) > 1$$

Dual SVM

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n$$

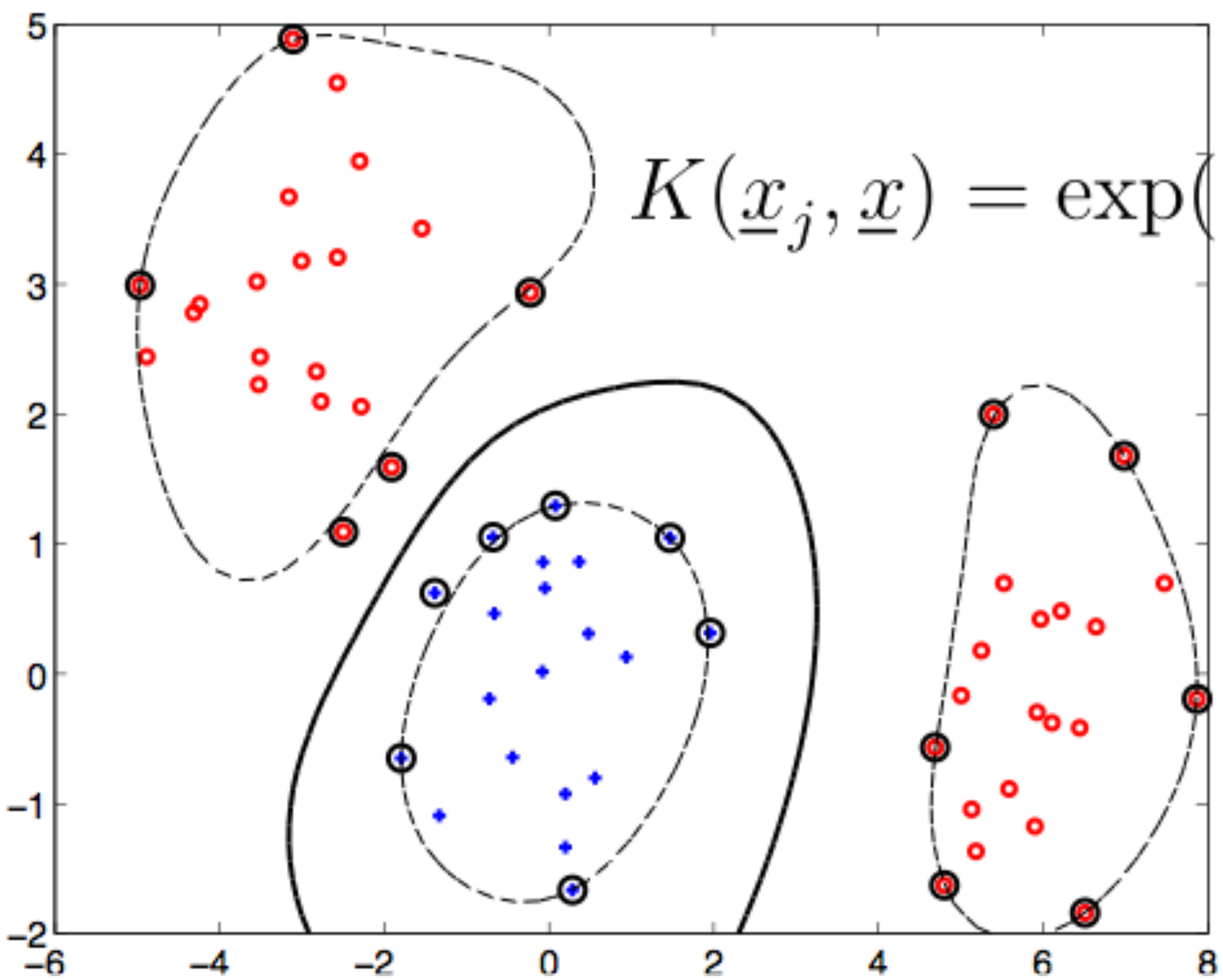
- This is again a quadratic programming problem but with simpler "box" constraints
- Once we solve for α_i^* , we predict labels according to

$$\begin{aligned} f(\underline{x}; \alpha^*) &= \operatorname{sign}\big(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x})\big) \\ &= \operatorname{sign}\Big(\sum_{i=1}^{n} \alpha_i^* y_i [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x})}_{\text{kernel}}]\Big) \end{aligned}$$

Kernel SVM

- Solving the SVM dual implicitly finds the max-margin linear separator in the feature space

$$f(\underline{x};\alpha) = \mathrm{sign}\Big(\sum_{i=1}^{n} \alpha_i y_i K(\underline{x}_i, \underline{x}) \Big)$$



$$K(\underline{x}_j, \underline{x}) = \exp\Big(-\frac{1}{4}\|\underline{x}_j - \underline{x}\|^2\Big)$$

Dual SVM with offset

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{[\phi(\underline{x}_i) \cdot \phi(\underline{x}_j)]}_{\text{kernel}}$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n, \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

- Where's the offset parameter? How do we solve for it?

Dual SVM with offset

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{[\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)]}_{\text{kernel}}$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n, \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

- Where's the offset parameter? How do we solve for it?
- We know that the classification constraints are tight for support vectors. If the ith point is a support vector, then $y_i(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i) + \theta_0^*) = 1$

Dual SVM with offset

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j [\underbrace{\underline{\phi}(\underline{x}_i) \cdot \underline{\phi}(\underline{x}_j)}_{\text{kernel}}]$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n, \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

- Where's the offset parameter? How do we solve for it?
- We know that the classification constraints are tight for support vectors. If the ith point is a support vector, then $$y_i(\underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i) + \theta_0^*) = 1$$

$$\Rightarrow \theta_0^* = y_i - \underline{\theta}(\alpha^*) \cdot \underline{\phi}(\underline{x}_i) = y_i - \sum_{j=1}^{n} \alpha_j^* y_j [\underbrace{\underline{\phi}(\underline{x}_j) \cdot \underline{\phi}(\underline{x}_i)}_{\text{kernel}}]$$

Note: you can pick any SV

Dual SVM with offset and slack

$$\underline{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{[\phi(\underline{x}_i) \cdot \phi(\underline{x}_j)]}_{\text{kernel}}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \; i = 1, \ldots, n, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- C is the same slack penalty as in the primal formulation