

# Compsci 571 HW4

Yilin Gao (yg95)

February 20, 2018

## 1 Constructing Kernels

## 2 Reproducing Kernel Hilbert Spaces

## 3 Convexity and KKT Conditions

(a) The Lagrangian function for the primal form is:

$$\begin{aligned} \min L(\mathbf{w}, \eta, \eta^*, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) = & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\eta_i + \eta_i^*) + \sum_{i=1}^n a_i [y_i - \mathbf{w}^T \mathbf{x}_i - \epsilon - \eta_i] \\ & + \sum_{i=1}^n b_i [\mathbf{w}^T \mathbf{x}_i - y_i - \epsilon - \eta_i^*] - \sum_{i=1}^n c_i \eta_i - \sum_{i=1}^n d_i \eta_i^* \end{aligned}$$

It's KKT conditions are:

- Primal feasibility:

$$y_i - \mathbf{w}^T \mathbf{x}_i - \epsilon - \eta_i \leq 0, i = 1, \dots, n$$

$$\mathbf{w}^T \mathbf{x}_i - y_i - \epsilon - \eta_i^* \leq 0, i = 1, \dots, n$$

$$\eta_i \geq 0, i = 1, \dots, n$$

$$\eta_i^* \geq 0, i = 1, \dots, n$$

- Dual feasibility:

$$a_i \geq 0, i = 1, \dots, n$$

$$b_i \geq 0, i = 1, \dots, n$$

$$c_i \geq 0, i = 1, \dots, n$$

$$d_i \geq 0, i = 1, \dots, n$$

- Complementary slackness:

$$a_i [y_i - \mathbf{w}^T \mathbf{x}_i - \epsilon - \eta_i] = 0, i = 1, \dots, n$$

$$b_i [\mathbf{w}^T \mathbf{x}_i - y_i - \epsilon - \eta_i^*] = 0, i = 1, \dots, n$$

$$c_i \eta_i = 0, i = 1, \dots, n$$

$$d_i \eta_i^* = 0, i = 1, \dots, n$$

- Lagrangian stationary:

$$\nabla_{\mathbf{w}L} = \mathbf{w} - \sum_{i=1}^n (a_i - b_i) \mathbf{x}_i = 0$$

$$\nabla_{\eta L} = C - \mathbf{a} - \mathbf{c} = 0$$

$$\nabla_{\eta^* L} = C - \mathbf{b} - \mathbf{d} = 0$$

With these conditions, we can transform Lagrangian function into dual form:

$$\max L(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n (a_i - b_i) y_i - \epsilon \sum_{i=1}^n (a_i + b_i) - \frac{1}{2} \sum_{i,j=1}^n (a_i - b_i)(a_j - b_j) \mathbf{x}_i^T \mathbf{x}_j$$

subject to

$$0 \leq a_i, b_i \leq C, i = 1, \dots, n$$

- (b) Support vectors are the points  $i$  such that  $|y_i - \mathbf{w}^T \mathbf{x}_i| \leq \epsilon$ .
- (c) Increasing  $\epsilon$  makes the model less likely to overfit. Because the model penalizes the points that have training error larger than  $\epsilon$ . If  $\epsilon$  increases, the allowed/unpenalized training error increases, and the model tends to overfit less.
- (d) Increasing  $C$  makes the model more likely to overfit.  $C$  is the penalty for each point that has training error larger than  $\epsilon$ . If the penalty increases, the model will try to make points have smaller training error, and thus overfits.
- (e) Assume we've computed the optimal dual variables as  $\mathbf{a}^*$  and  $\mathbf{b}^*$ .

From one of the KKT conditions, we can get the optimal primal variable is  $\mathbf{w}^* = \sum_{i=1}^n (a_i^* - b_i^*) \mathbf{x}_i$ .

So for a new point  $\mathbf{x}^{new}$ , its evaluation is  $f(\mathbf{x}^{new}) = \sum_{j=1}^p w_j^* x_j^{new} = \sum_{j=1}^p \sum_{i=1}^n (a_i^* - b_i^*) x_{ij} x_j^{new} = \sum_{i=1}^n (a_i^* - b_i^*) \mathbf{x}_i \cdot \mathbf{x}^{new}$ .

## 4 SVM Implementation

- (a) See Jupyter notebook
- (b)
- (c)