Machine Learning

CSCI 567 Spring 2021

Discussion: Kernel Method, SVM

Date: Feb 26th, 2021

1. In the lecture, we discuss both the primal and dual formulations for SVM that learn a linear model $\boldsymbol{w}^\top \phi(\boldsymbol{x}) + b$ with an explicit bias term b. In this exercise, follow the steps below to derive the primal and dual formulations for SVM that learn a linear model *without a bias term* (that is, set the bias term b to be zero).
(a) Write down the primal formulation where the variables are w and $\xi_1, \ldots, \xi_N$? 0 (note again that b is gone). Ans:

$$\begin{aligned} \min_{\boldsymbol{w}, \xi_n} \quad & C \sum_n \xi_n + \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1 - y_n \boldsymbol{w}^\top \phi(\boldsymbol{x}_n) \leq \xi_n, \forall n \\ & \xi_n \geq 0 \end{aligned}$$

(b) Write down the Lagrangian of the primal formulation, with $\alpha_1, \ldots, \alpha_N \geq 0$ and $\lambda_1, \ldots, \lambda_N$ being the Lagrangian multipliers. Ans:

$$L(\boldsymbol{w}, \{\xi_n\}, \{\alpha_n\}, \{\lambda_n\}) = C \sum_n \xi_n + \frac{1}{2}\|w\|_2^2 - \sum_n \lambda_n \xi_n + \sum_n \alpha_n (1 - y_n \boldsymbol{w}^\top \phi(\boldsymbol{x}_n) - \xi_n)$$

(c) Apply the stationarity condition from KKT to obtain equations linking the primal variables and the dual variables. Ans:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_n y_n \alpha_n \phi(\boldsymbol{x}_n) = 0 \implies \boldsymbol{w} = \sum_n y_n \alpha_n \phi(\boldsymbol{x}_n)$$

$$\frac{\partial L}{\partial \xi_n} = C - \lambda_n - \alpha_n = 0, \forall n$$

2. What is the corresponding $\phi(x)$ for the polynomial kernel of 2-dimensional vectors:

$$k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z} + 1)^2$$

Ans:

$$(1+\boldsymbol{x}^\top \boldsymbol{z})^2 = (1+x_1z_1+x_2z_2)^2 = 1+(x_1z_1)^2+(x_2z_2)^2+2x_1z_1x_2z_2+2x_1z_1+2x_2z_2$$

The feature vector is $\phi(\boldsymbol{x}) = (1, x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)$.

3. How many parameters do we need to learn for the following network structure? An $8\times 8\times 3$image input, followed by a convolution layer with 2 filters of size 2×2 (stride 1, no zero-padding), then another convolution layer with 4 filters of size 3×3 (stride 2, no zero-padding), and finally a max-pooling layer with a 2×2 filter (stride 1, no zero-padding). (Note: the depth of all filters are not explicitly spelled out, and we assume no bias/intercept terms are used.)

(A) 96
(B) 44
(C) 100
(D) 48

Ans: A. $2\times(2\times 2\times 3)+4\times(3\times 3\times 2) = 96$

4. Which is *not* a valid kernel function, for samples x and y and kernels $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$?

(A) $k(x,y) = 5$
(B) $k(x,y) = x+y$
(C) $k(x,y) = e^{x+y}$
(D) $k(x,y) = \langle x,y\rangle^3 + (\langle x,y\rangle + 1)^2$

Ans: B

5. In this problem, we prove that if we are using the Newton's method to solve the least square optimization problem, then it only takes one step to converge. Recall that the Newton's method update the parameters as follow:

$$w^{t+1} = w^t - H^{-1}\nabla L(w^t)$$

where $H = \nabla^2 L(w^t)$ is the Hessian matrix of the loss function, i.e., $H_{ij} = \frac{\partial}{\partial w_i \partial w_j} L(w^t)$.

(1) Find the Hessian of least square loss function: $L(w) = \frac{1}{2}\sum_{n=1}^N (w^T x_n - y_n)^2$.

Ans:

$$\frac{\partial}{\partial w_j} L(w) = \sum_{n=1}^{N} (w^T x_n - y_n) x_{nj}$$
$$\frac{\partial^2}{\partial w_i \partial w_j} L(w) = \sum_{n=1}^{N} x_{ni} x_{nj} = (X^T X)_{ij}$$

Therefore, $H = \nabla^2 L(w) = X^T X$.

(2) Show that given any w^0, after first iteration of Newton's method, we obtain the optimal $w^* = (X^T X)^{-1} X^T y$.

Ans:

$$\begin{aligned} w^1 &= w^0 - H^{-1} \nabla L(w^0) \\ &= w^0 - H^{-1} X^T (X w^0 - y) \\ &= w^0 - w^0 + (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T y \end{aligned}$$