Data Application Lab

Tuesday Machine Learning Basis Quiz Answer

Part I: Deeper understanding on SVM

Problem 1: Non-seperable SVM

(i) The original constraint optimization problem:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i$$
$$\text{s.t. } y^i(w^T x^i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0.$$

The generalized Lagrangian:

$$L(w, b, \alpha, r, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i + \sum_i^m \alpha_i[1 - \xi_i - y^i(w^T x^i + b)] - \sum_i^m r_i \xi_i \quad ①$$

$\underline{\min_{w} \max_{\alpha, r : \alpha_i \geq 0, r_i \geq 0} L(w, b, \alpha, r, \xi)}$ has the same solution as the original problem.

$\Rightarrow$ Dual problem $\max_{\alpha, r : \alpha_i \geq 0, r_i \geq 0} \min_w L(w, b, \alpha, r, \xi)$

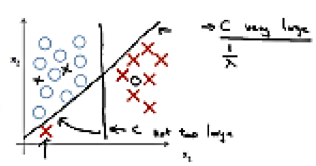
$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_i^m \alpha_i y^i x^i = 0 \Rightarrow w = \sum_i^m \alpha_i y^i x^i & ② \\ \frac{\partial L}{\partial b} = -\sum_i^m \alpha_i y^i = 0 \Rightarrow \sum_i^m \alpha_i y^i = 0. & ③ \\ \frac{\partial L}{\partial \xi_i} = C - \alpha_i - r_i = 0 & ④ \end{cases}$$

②③④ $\Rightarrow$ ①:

$$L(w^*, b^*, \xi^*) = \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i + \sum_i^m \alpha_i[1 - \xi_i - y^i(w^T x^i + b)] - \sum_i^m r_i \xi_i$$
$$= \frac{1}{2}\|w\|^2 + \cancel{C\sum_i^m \xi_i} + \sum_i^m \alpha_i[1 - y^i(w^T x^i)] - \cancel{\sum_i^m \alpha_i y^i b} - \cancel{\sum_i^m \alpha_i \xi_i} - \cancel{\sum_i^m r_i}$$
$$= \frac{1}{2}\|w\|^2 + \sum_i^m \alpha_i - w^T \underbrace{\left[\sum_i^m \alpha_i y^i x^i\right]}_{w}$$
$$= \sum_i^m \alpha_i - \frac{1}{2}\|w\|^2 = \sum_i^m \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle$$

$\therefore$ the final result:

$$\max_{\alpha, r} \sum_i^m \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle$$
$$\text{s.t. } \alpha_i \geq 0,\ \alpha_i \leq C,\ \sum_i \alpha_i y^i = 0.$$
$$r_i \geq 0.$$



(ii) $C \to \infty$. the obnormal points will affect the classifier more, it will have a smaller margin hyperplane when it can seperate the points correctly

$C \to 0$ the classifier will have a larger margin hyperplane even it will misclassify points.

Problem 2: Kernels
Please try to understand the kernel tricks basing on this question.

$$\text{kernel} \quad k(x,z) = (x^T z + c)^2$$
$$= (x_1 z_1 + x_2 z_2 + c)^2$$
$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 + 2cx_1z_1 + 2cx_2z_2 + c^2$$

$$\phi(x) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2c}\,x_1, \sqrt{2c}\,x_2, c]^T$$
$$\phi(z) = [z_1^2, \sqrt{2}z_1z_2, z_2^2, \sqrt{2c}\,z_1, \sqrt{2c}\,z_2, c]^T$$

2

$$\langle \phi(x), \phi(z) \rangle = x_1^2 z_1^2 + 2x_1x_2z_1z_2 + x_2^2z_2^2 + 2cx_1z_1 + 2cx_2z_2 + c^2$$

$$\therefore\ k(x,z) = \langle \phi(x), \phi(z) \rangle.$$

Part II: The basic knowledge on ANN

Problem 3: Backpropagation

$$\frac{\partial J(w,x,y)}{\partial w_{ji}^{(1)}} = \frac{\partial J}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial h_j} \cdot \frac{\partial h_j}{\partial w_{ji}^{(1)}}$$

sigmoid function:
$$\sigma'(z) = \sigma(z)[1-\sigma(z)]$$

$$\frac{\partial J}{\partial \hat{y}_k} = (\hat{y}_k - y_k).$$

$$\frac{\partial \hat{y}_k}{\partial h_j} = \frac{\partial\sigma(\sum_j h_j w_{kj}^{(2)})}{\partial h_j} = \sigma'(\textstyle\sum_j h_j w_{kj}^{(2)}) \cdot w_{kj}^{(2)} = \hat{y}_k(1-\hat{y}_k)\, w_{kj}^{(2)}.$$

$$\frac{\partial h_j}{\partial w_{ji}^{(1)}} = \frac{\partial\sigma(\sum_i x_i w_{ji}^{(1)})}{\partial w_{ji}^{(1)}} = h_j(1-h_j)\, x_i$$

$$\therefore\ \frac{\partial J(w,x,y)}{\partial w_{ji}^{(1)}} = \sum_k (\hat{y}_k - y_k)\, \hat{y}_k(1-\hat{y}_k)\, w_{kj}^{(2)}\, h_j(1-h_j)\, x_i.$$

Problem 4: What are the advantages of ReLU over sigmoid function?

Sigmoid function has the problem of vanish gradient because the gradient of sigmoid becomes increasingly small as the absolute value of x increases. But ReLU can **reduce the likelihood of the gradient to vanish** and the constant gradient of ReLU when x>0 will **result in faster learning**.

Another advantage of ReLU is **sparsity**, which arises when x <=0. The more such units that exist in a layer the more sparse the resulting representation.