# Machine Learning Exercise 5

何舜成

2012011515

**1. Show KKT condition is necessary condition and sufficient in some cases**:

Proof:

A standard form of optimal problem can be described as follow

$$\begin{aligned} min \quad & f_0(x) \\ s.t. \quad & f_i(x) \le 0, i = 1, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p \end{aligned}$$

where $x \in \mathcal{R}^n$ and the domian $\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom} f_i \cap \bigcap_{i=1}^{p} \mathbf{dom} h_i$ not empty. And the optimal solution is $p^*$.

Define Lagrangian function

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

where $\lambda$ and $\mu$ are called Lagrangian multiplier.

Define Lagrangian dual function

$$g(\lambda, \mu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \mu) = \inf_{x \in \mathcal{D}} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x))$$

Suppose $\tilde{x}$ is feasible to primal problem, then for all $\lambda \ge 0$ and $\mu$

$$\sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x) \le 0$$

$$L(\tilde{x},\lambda,\mu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x) \le f_0(\tilde{x})$$

therefore

$$g(\lambda,\mu) = \inf_{x\in\mathcal{D}} L(x,\lambda,\mu) \le L(\tilde{x},\lambda,\mu) \le f_0(\tilde{x})$$

Consider the Lagrangian dual problem

$$\begin{array}{ll} max & g(\lambda,\mu) \\ s.t. & \lambda_i \ge 0, i = 1,\cdots,m \end{array}$$

the dual problem is convex optimal problem. If $(\lambda^*,\mu^*)$ is the optimal solution and the target function reaches $d^*$ at $(\lambda^*,\mu^*)$, we easily get this inequality

$$d^* \le p^*$$

and we call this weak duality. Likewise, if

$$d^* = p^*$$

holds true, we call this strong dualtiy.

In condition of strong duality and the existence of optimal solution $x^*$ of primal problem, we have

$$\begin{aligned} f_0(x^*) &= g(\lambda^*,\mu^*) && (1) \\ &= \inf_x (f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \mu_i^* h_i(x)) && (2) \\ &\le f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \mu_i^* h_i(x^*) && (3) \\ &\le f_0(x^*) && (4) \end{aligned}$$

Therefore the inequalities are forced to be equal. We can infer

$$\left.\frac{\partial L(x, \lambda^*, \mu^*)}{\partial x}\right|_{x=x^*} = 0$$

from the third equation (supposing $f_i$ and $h_i$ are differentiable), and infer

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

or $\lambda_i^* f_i(x^*) = 0, i = 1, \cdots, m$ from the fourth equation.

To conclude, if $x^*$ and $(\lambda^*, \mu^*)$ are optimal solutions of the primal and dual problem respectively, and strong duality is sufficed, KKT condition

(1) $x^*$ is primal feasible

$$f_i(x) \leq 0, i = 1, \cdots, m$$

$$h_i(x) = 0, i = 1, \cdots, p$$

(2) $(\lambda^*, \mu^*)$ are dual feasible

$$\lambda^* \geq 0, i = 1, \cdots, m$$

(3) complementary slackness

$$\lambda^* f_i(x^*) = 0, i = 1, \cdots, m$$

(4) stationary

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla h_i(x^*) = 0$$

will hold true simultaneously. KKT condition is **necessary**.

KKT condition is also sufficient if

$$\left.\frac{\partial L(x, \lambda^*, \mu^*)}{\partial x}\right|_{x=x^*} = 0 \Rightarrow L(x^*, \lambda^*, \mu^*) = \inf_{x \in \mathcal{D}} (x, \lambda^*, \mu^*)$$

Proof:

Combining the condition above with the (4) equation in KKT condition, we get

$$g(\lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*)$$

According to other 3 conditions, we know

$$g(\lambda^*, \mu^*) = f_0(x^*)$$

Therefore we can infer $x^*$ and $(\lambda^*, \mu^*)$ are optimal solutions of primal and dual problem respectively from weak duality. KKT condition is **sufficient**.

**2. Give the dual problem of SVM when linear inseparable**

Slackness variables are introduces when data are linear inseparable. The primal problem

$$\begin{aligned} min \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i \in [n] \\ & \xi_i \geq 0, \forall i \in [n] \end{aligned}$$

The Lagrangian function is

$$L(w, b, \xi, \alpha, r) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} r_i\xi_i$$

for all $i \in [n], \alpha_i \geq 0, r_i \geq 0$.

Set the partial derivatives to zero, and we get

$$\begin{aligned} \frac{\partial L}{\partial w} = 0 \quad & \Rightarrow \quad w = \sum_{i=1}^{n}\alpha_i y_i x_i \\ \frac{\partial L}{\partial b} = 0 \quad & \Rightarrow \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \quad & \Rightarrow \quad C - \alpha_i - r_i = 0, \forall i \in [n] \end{aligned}$$

According to the previous exercise,

$$L(w,b,\xi,\alpha,r)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

Therefore the dual can be described as follow

$$\begin{aligned}
max \quad & \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
s.t. \quad & 0\le\alpha_i\le C,\forall i\in[n] \\
& \sum_{i=1}^{n}\alpha_i y_i=0
\end{aligned}$$

**3. Design an algorithm to calculate gradient of the loss function of neural network**

Mathematical model of artificial neural network:

$$x^l=f(u^l),u^l=(W^{l-1})^T x^{l-1}+b^l$$

where $l$ denotes the current layer with the output layer designated to be layer $L$ and the input layer desiganted to ba layer 1. Function $f(\cdot)$ is a nonlinear function (i.e. sigmoid or hyperbolic tangent).Define the loss function as

$$E(x^L,t)$$

where $x^L$ is the network output and $t$ is the target output. Usually we choose

$$E(x^L,t)=\frac{1}{2}||t-x^L||^2$$

Since

$$E(x^L,t)=E(f((W^{L-1})^T x^{L-1}),t)$$

we can write the derivatives w.r.t. $W^{L-1}$

$$\frac{\partial E}{\partial W^{L-1}}=x^{L-1}(f'(u^L)\star\frac{\partial E}{\partial x^L})^T$$

where $\star$ denotes elementwise multipying, and if we define

$$\delta^L = f'(u^L) \star \frac{\partial E}{\partial x^L}$$

we get

$$\frac{\partial E}{\partial W^{L-1}} = x^{L-1}(\delta^L)^T$$

If we calculate the $\delta$ term recursively

$$\delta^l = f'(u^l) \star ((W^l)^T \delta^{l+1}), l = L-1, \cdots, 2$$

it is easy to write

$$\frac{\partial E}{\partial W^l} = x^l(\delta^{l+1})^T, l = L-2, \cdots, 1$$