

Problem 1

(a) the difference between ASVM primal problem is the slack variables which allow the i th point to violate the margin. and the role is to maximise the margin between the updated classifier and new training data.

$$(b) L(w, \xi, \alpha, r) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i (y_i (f_0(x_i) + w^T x_i) - 1 + \xi_i) - \sum_i r_i \xi_i$$

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w^* = \sum_i \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - r_i = 0 \Rightarrow r_i = C - \alpha_i$$

$$(c) L = \frac{1}{2} \|\sum_i \alpha_i y_i x_i\|^2 + C \sum_i \xi_i - \sum_i \alpha_i (y_i [f_0(x_i) + (\sum_j \alpha_j y_j x_j)^T x_i] - 1 + \xi_i) - \sum_i r_i \xi_i$$

$$= \frac{1}{2} (\sum_i \alpha_i y_i x_i)^T (\sum_j \alpha_j y_j x_j) - \sum_i \alpha_i y_i \sum_j \alpha_j y_j x_j^T x_i - \sum_i \alpha_i y_i f_0(x_i) + \sum_i \alpha_i + \sum_i (C - \alpha_i - r_i) \xi_i$$

Hence, the dual function is

$$L(\alpha) = \sum_i \alpha_i (1 - y_i f_0(x_i)) - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

and we have $\begin{cases} \alpha_i \geq 0 \\ r_i \geq 0 \\ r_i = C - \alpha_i \end{cases} \Rightarrow 0 \leq \alpha_i \leq C$

Hence, the ASVM dual problem is

$$\max_{\alpha} \sum_i \alpha_i (1 - y_i f_0(x_i)) - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t. } 0 \leq \alpha_i \leq C$$



(d) we have $g(x) = y_i(f_0(x_i) + w^T x_i) - 1 + \xi$

so the KKT conditions are : $\begin{cases} \alpha_i = 0, g(x) > 0 & \text{inactive} \\ \alpha_i > 0, g(x) = 0 & \text{active} \end{cases}$

① $r_i > 0, \alpha_i = 0, \xi_i = 0, y_i(f_0(x_i) + w^T x_i) > 1 - \xi_i > 1$. beyond the margin

② $r_i > 0, 0 < \alpha_i < C, \xi_i = 0, y_i(f_0(x_i) + w^T x_i) = 1 - \xi_i = 1$ on the margin

③ $r_i = 0, \alpha_i = C, \xi_i > 0, y_i(f_0(x_i) + w^T x_i) < 1 - \xi_i < 1$ violating the margin

(e) compared to the original soft-SVM dual, ASVM put an condition into soft-SVM's goal function

$f_0(x)$ simplify the computation because only wrong classified data will be considered.

Problem 2

(a) in the stock market, the x is return percentage. the y is the number of corresponding quantity
this is a normal distribution and the mean equal to 0, the deviation is volatility that we care.

(b) in this question, $x_i: d \times 1$ $y_i: 1 \times 1$ $w: d \times 1$
 $X: d \times n$ $y: n \times 1$

the MAP objective function is $l = \log p(y|X, w) + \log p(w)$

$$= \sum_{i=1}^n \log p(y_i|x_i, w) + \log p(w) = \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi e^{-w^T x_i}) - \frac{1}{2} e^{w^T x_i} y_i^2 \right] - \frac{1}{2} \log[(2\pi)^d |\Sigma|] - \frac{1}{2} w^T \Sigma^{-1} w$$

dropping terms that are constant, the MAP solution is equivalent to

$$w^* = \underset{w}{\operatorname{argmax}} l = \underset{w}{\operatorname{argmin}} \underbrace{\sum_{i=1}^n (e^{w^T x_i} y_i^2 - w^T x_i) + w^T \Sigma^{-1} w}_{}$$

$$(C) \text{ we use } E(w) = \sum_{i=1}^n (e^{w^T x_i} y_i^2 - w^T x_i) + w^T \Sigma^{-1} w$$

$$\text{the gradient of } E(w) \text{ is } \nabla E(w) = \sum_{i=1}^n (y_i^2 e^{w^T x_i} x_i - x_i) + 2 \Sigma^{-1} w \quad \dots (1)$$

$$= \sum_{i=1}^n (y_i^2 e^{w^T x_i} - 1) x_i + 2 \Sigma^{-1} w = X(Ry^2 - I) + 2 \Sigma^{-1} w \quad \text{where } R = \text{diag}(e^{w^T x_1}, \dots, e^{w^T x_n})$$

$$\text{the Hessian of } E(w) \text{ is } \nabla^2 E(w) = \frac{\partial}{\partial w} \frac{\partial}{\partial w} E(w) = \frac{\partial}{\partial w} [(Ry^2 - I)^T X^T + 2 w^T \Sigma^{-1}]$$

$$= \frac{\partial}{\partial w} [y^T R X^T - X^T + 2 w^T \Sigma^{-1}]$$

$$\text{because } \frac{\partial}{\partial w} R = \frac{\partial}{\partial w} \begin{bmatrix} e^{w^T x_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{w^T x_n} \end{bmatrix} = [e^{w^T x_1} x_1, \dots, e^{w^T x_n} x_n] = X R$$

$$\text{using (1), we can also get } \nabla^2 E(w) = \sum_{i=1}^n y_i^2 e^{w^T x_i} x_i x_i^T + 2 \Sigma^{-1}$$

$$\text{so } \nabla^2 E(w) = X S X^T + 2 \Sigma^{-1} \quad \text{where } S = \text{diag}(e^{w^T x_1} y_1^2, \dots, e^{w^T x_n} y_n^2) \\ = \text{diag}(Ry^2) = R \text{diag}(y^2)$$

$$\text{the Newton-Raphson iteration is } w_{n+1} = w_n - [\nabla^2 E(w)]^{-1} \nabla E(w) \\ = w_n - (X S X^T + 2 \Sigma^{-1})^{-1} [X(Ry^2 - I) + 2 \Sigma^{-1} w_n] \\ = (X S X^T + 2 \Sigma^{-1})^{-1} X S [X^T w_n - S^{-1}(Ry^2 - I)]$$

(d) if $\Sigma = \lambda I$, $\Sigma^{-1} = \frac{1}{\lambda} I$

$$\text{come back to see } w_{n+1} = (X S X^T + 2 \Sigma^{-1})^{-1} X S z = (X S X^T + \frac{2}{\lambda} I)^{-1} X S z$$

Σ is added to the term $y^T R X X^T$ which is then inverted. Hence, Σ helps to ensure that we avoid inverting a singular matrix.

$$P = \frac{1}{2} \sum \quad B^T = X \quad R^T = S$$

(e) Define $d_* = w^T x_* = x_*^T w = x_*^T (X^T S X + (\frac{1}{2}\Sigma)^{-1})^{-1} X^T S z$
 $= \frac{1}{2} x_*^T \Sigma X (\frac{1}{2} X^T \Sigma X + S^{-1})^{-1} z = \underline{\underline{k_*^T (K + S^{-1}) z}}$

where $K_* = \frac{1}{2} x_*^T \Sigma X, \quad K = \frac{1}{2} X^T \Sigma X$

so we can use $k_* = [k(x_*, x_1), \dots, k(x_*, x_n)]^T$ as the test kernel, and
 $K = [k(x_i, x_j)]_{ij}$ is the training kernel matrix. Hence, the regression model
is $\sigma_*^2 = e^{-d_*}$

(f) we have kernelized w in (e),

so we can see $\begin{cases} w_{n+1} = \frac{1}{2} \sum X (\frac{1}{2} X^T \Sigma X + S^{-1})^{-1} z_n \\ z_n = X^T w_n - S^{-1} (Ry^2 - I_m) \end{cases}$

or can be written as $\begin{cases} \alpha^{(new)} = K (K + S^{-1}) z^{(old)} \\ z^{(new)} = \alpha^{(old)} - S^{-1} (Ry^2 - I_m) \end{cases}$

(g) Σ change the kernel regression model on scale, $\frac{1}{2} x_i^T \Sigma x_j$ for an example, we use $d=2$, and $\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix}$. so $\frac{1}{2} x_i^T \Sigma x_j = \frac{1}{2} (a x_{i1} x_{j1} + b x_{i2} x_{j2})$, a and b determine the weights of each x components.

(h) the kernel function method is used to solve the problem which can not be linearly separable. so the advantage of kernelized algorithms is it can solve higher dimensional problems; and the disadvantage is when it faces low dimensional problems, the computing is complicated.

The work in these answer sheets are my own work. I have not discussed this quiz with anyone else. I have only used the allowed materials

YANG Zhiyuan 55834243 1b/12/2019 杨致远