Solution of Assignment 3

November 6, 2014

1. Shown that in regularized linear regression, $\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{r} = \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_N)^{-1} \boldsymbol{r}$, a linear combination of the examples.
(Hint: use the matrix inversion in block form: `http://www.cs.nthu.edu.tw/~jang/book/addenda/matinv/matinv/`)

   Answer :

   $(X^\top X + \lambda I_d)^{-1} X^\top$

   Since the matrix inversion lemma, let $A = \lambda I_d,\ B = X^\top,\ C = I_N,\ D = X$
   $= \frac{1}{\lambda} X^\top - \frac{1}{\lambda^2} X^\top (I_N + \frac{1}{\lambda} X X^\top)^{-1} X X^T$
   $= \frac{1}{\lambda} (X^\top - X^\top (\lambda)^{-1} (I_N + \frac{1}{\lambda} X X^\top)^{-1} X X^T)$
   $= \frac{1}{\lambda} (X^\top - X^\top (\lambda I_N + X X^\top)^{-1} X X^T)$
   $= \frac{1}{\lambda} (X^\top (\lambda I_N + X X^\top)^{-1} (\lambda I_N + X X^\top) - X^\top (\lambda I_N + X X^\top)^{-1} X X^T)$
   $= \frac{1}{\lambda} ((X^\top (\lambda I_N + X X^\top)^{-1})(\lambda I_N + X X^\top - X X^T)$
   $= X^\top (X X^\top + \lambda I_N)^{-1}$

2. Show that in an RKHS, the inner product $\langle f, g \rangle := \sum_{i=1}^n \sum_{j=1}^m \alpha^{(i)} \beta^{(j)} k(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ for any $f = \sum_{i=1}^n \alpha^{(i)} k(\boldsymbol{x}^{(i)}, \cdot)$ and $g = \sum_{j=1}^m \beta^{(j)} k(\boldsymbol{y}^{(j)}, \cdot)$ is well-defined; i.e., it satisfies

   (a) *symmetry*: $\langle f, g \rangle = \langle g, f \rangle$;

   (b) *linearity*: $\langle af + bg, h \rangle = a \langle f, h \rangle + b \langle g, h \rangle$ for any $a, b \in \mathbb{R}$; and

   (c) *positive definiteness*: $\langle f, f \rangle \geq 0$ with equality holds iff $f(\cdot) = 0(\cdot)$.

   Answer :

   (a) $\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha^{(i)} \beta^{(j)} k(x^{(i)}, y^{(j)})$
   $= \sum_{j=1}^m \sum_{i=1}^n \beta^{(j)} \alpha^{(i)} k(y^{(j)}, x^{(i)}) = \langle g, f \rangle$

   (b) $\langle af + bg, h \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha^{(i)} \beta^{(j)} k(ax^{(i)} + by^{(i)}, z^{(j)}) = \sum_{i=1}^n \sum_{j=1}^m \alpha^{(i)} \beta^{(j)} (k(ax^{(i)}, z^{(j)}) + k(by^{(i)}, z^{(j)}))$
   $= \sum_{i=1}^n \sum_{j=1}^m \alpha^{(i)} \beta^{(j)} (a \times k(x^{(i)}, z^{(j)}) + b \times k(y^{(i)}, z^{(j)})) = a \langle f, h \rangle + b \langle g, h \rangle$.

(c) $\langle f, f\rangle = \left\langle \sum_{i=1}^n \alpha^{(i)} k(x^{(i)}, \cdot), \sum_{j=1}^n \alpha^{(i)} k(x^{(i)}, \cdot) \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha^{(i)} \alpha^{(j)} k(x^{(i)}, x^{(j)}) \geq 0$ (by definition of kernel).
Next, we show that $\langle f, f\rangle = 0$ iff $\forall x, \sum_{i=1}^n \alpha^{(i)} k(x^{(i)}, \cdot) = 0(\cdot)$
$\Rightarrow 0 = \sum_{i=1}^n \sum_{j=1}^n \alpha^{(i)} \alpha^{(j)} k(x^{(i)}, x^{(j)}) = \sum_{j=1}^n \alpha^{(j)} (\sum_{i=1}^n \alpha^{(i)} k(x^{(i)}, x^{(j)})) \Rightarrow f = 0(\cdot)$.
$\Leftarrow$ *Trivial.*

3. Show that in a large-margin linear classifier, the margin between the hyperplanes $\{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} - b = 1\}$ and $\{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} - b = -1\}$ is $2/\|\boldsymbol{w}\|$.

Answer :

Let p be a point on h_1, i.e., $\boldsymbol{w}^\top p - b = 1$, and q be the point on h_2, $w^\top q - b = -1$, which has minimum distance to p. Since $\frac{w}{||w||}$ is the unit normal vector of h_1, we have $q = p + t\frac{w}{||w||}$. We can solve t by $\boldsymbol{w}^\top q - b = -1 \Rightarrow \boldsymbol{w}^\top (q = p + t\frac{w}{||w||}) - b = -1 \Rightarrow t = \frac{(-1-(w^\top p - b))}{||w||} = \frac{-2}{||w||}$. So, the distance between p and q is $||p - q|| = ||p - (p + t\frac{w}{||w||})|| = \frac{-2}{||w||}$.

4. Prove the Semiparametric Representer theorem below:

Theorem 1. *Let $\widetilde{g} := g + b\psi$, where $g \in \mathcal{H}$, $b \in \mathbb{R}$, and $\psi : \mathcal{I} \to \mathbb{R}$. Then each minimizer $\widetilde{h}$ of the regularized risk functional:*

$$\arg\min_{\widetilde{g}} L((\boldsymbol{x}^{(1)}, r^{(1)}, \widetilde{g}(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, \widetilde{g}(\boldsymbol{x}^{(N)}))) + \Omega(\|g\|_{\mathcal{RKHS}})$$
$$\textit{subject to } C_k((\boldsymbol{x}^{(1)}, r^{(1)}, \widetilde{g}(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, \widetilde{g}(\boldsymbol{x}^{(N)}))) \leq 0, \forall k$$

admits the form $\widetilde{h}(\boldsymbol{x}) = \sum_{t=1}^N c_t k(\boldsymbol{x}^{(t)}, \boldsymbol{x}) + b\psi(\boldsymbol{x})$.

Answer :

(a) Consider $\tilde{\Omega}(||g||^2_{RKHS})$ for convenience.

(b) By the fundamental theorem of linear algebra, decompose any $g \in H$ into two vectors parallel and orthogonal to $V = span(k(\cdot, x_i))_{i=1,\cdots,m}$ respectively. We have $g(x) = f_p(x) + f_o(x) \Rightarrow \tilde{g}(x) = f_p(x) + f_o(x) + b\psi(x)$, where $f_p \in V$, and $f_o \in V^\perp$.

(c) Since $f_o \in V^\perp \Rightarrow f_o(x) = 0$, $\tilde{g}(x) = f_p(x) + b\psi(x)$.

(d) Suppose the minimizer $\tilde{h}$ has the form $\tilde{h}(x) = f_p(x) + f_o(x) + b\psi(x)$; however, $\tilde{h} - f_o$ is always a better solution.

i. First, $\tilde{h} - f_o$ satisfies all constrains $C'_k s$ as $(\tilde{h} - f_o)(x^{(i)}) = \tilde{h}(x^{(i)})$ for all $1 \leq i \leq N$.

ii. Due to the same reason, $\tilde{h} - f_o$ has the same loss score from L as h.

iii. $\tilde{\Omega}(||h - f_o||^2) = \tilde{\Omega}(||f_p||^2) \leq \tilde{\Omega}(||f_p||^2 + ||f_o||^2)$
$= \tilde{\Omega}(||f_p + f_o||^2) = \tilde{\Omega}(||h||^2)$

(e) By contradiction, $\tilde{h}(x) = f_p(x) + b\psi(x) = \sum_{t=1}^N c_t k(x^{(t)}, x) + b\psi(x)$.

5. Consider the necessary optimality condition for the SVM dual (that is, if the following holds, then $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$):

$$\exists \rho \in \mathbb{R} \text{ such that } \max_{i \in I_{up}} r^{(i)} g_i \leq \rho \leq \min_{j \in I_{down}} r^{(j)} g_j \tag{1}$$

where $I_{up} := \{i | r^{(i)} \alpha_i < B^{(i)}\}$ and $I_{down} := \{j | A^{(j)} < r^{(j)} \alpha_j\}$.

I_{up} :
正Instance : $r^{(i)}\alpha_i<C$ · $\alpha_i<C$ · $r^{(i)}$在線上或線外
負Instance : $r^{(i)}\alpha_i<0$ · $\alpha_i>0$ · $r^{(i)}$在線上或線內

I_{down} :
正Instance : $r^{(i)}\alpha_i>0$ · $\alpha_i>0$ · $r^{(i)}$在線上或線內
負Instance : $r^{(i)}\alpha_i>-C$ · $\alpha_i<C$ · $r^{(i)}$在線上或線外

正Instance : rg>ρ · g>ρ · rα=C · α=C · SV
負Instance : rg>ρ · g<ρ · rα=0 · α=0 · Non SV
正Instance又在I_{down}才是SV

正Instance : rg<ρ · g<ρ · rα=0 · α=0 · Non SV
負Instance : rg<ρ · g>ρ · rα=-C · α=C · SV
負Instance又在I_{up}才是SV

(a) Show that the above condition can be rewritten as

$$\exists \rho \in \mathbb{R} \text{ such that } \forall t, \begin{cases} r^{(t)}\alpha_t^* = B^{(t)}, & \text{if } r^{(t)}g_t^* > \rho \\ r^{(t)}\alpha_t^* = A^{(t)}, & \text{if } r^{(t)}g_t^* < \rho \end{cases}, \quad (2)$$

I_{down}的正Ins或I_{up}的非SV
I_{up}的SV或I_{down}的非SV

which is equivalent to

結論 · g>rρ才是SV · 此時α=C

$$\exists \rho \in \mathbb{R} \text{ such that } \forall t, \begin{cases} \alpha_t^* = C, & \text{if } g_t^* > r^{(t)}\rho \\ \alpha_t^* = 0, & \text{if } g_t^* < r^{(t)}\rho \end{cases} \quad (3)$$

是SV ·

(b) Show that Condition (1) is also sufficient by letting $\boldsymbol{w}^* = \sum_{t=1}^N \alpha_t^* r^{(t)} \Phi(\boldsymbol{x}^{(t)})$, $b^* = \rho$, and $\xi_t^* = \max\{0, g_t^* - r^{(t)}\rho\}$.
(Hint: show that there is no duality gap, i.e., $primal(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) - dual(\boldsymbol{\alpha}^*) = 0$, by Condition (3))

Answer :

不等於B只會<B

(a) If $r^{(t)}g_t^* > \rho$ and $r^{(t)}\alpha_t^* < B^{(t)}$, then the instance t must be in I_{up}, meaning $\max_{i \in I_{up}} r^{(i)} g_i > \rho$ in Condition 1, a contradiction. Similarly, if $r^{(t)}g_t^* < \rho$ and $r^{(t)}\alpha_t^* > A^{(t)}$, then the instance t must be in I_{down}, leading to another contradiction $\rho > \min_{j \in I_{down}} r^{(j)} g_j$ to Condition 1. Condition (3) is just a simple rewrite of Condition (2).

(b) Recall that

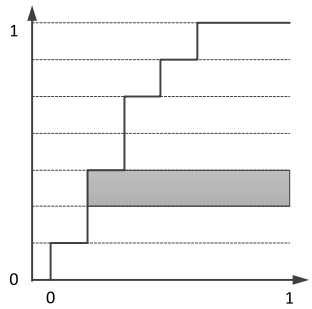
$$primal(w, b, \xi) = \frac{1}{2}||w||^2 + C\sum_{t=1}^N \xi_t,$$

$$dual(\alpha) = \alpha 1 - \frac{1}{2}\alpha^\top \tilde{K}\alpha = \sum_{t=1}^N \alpha_t - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N r^{(i)} r^{(j)} \alpha_i \alpha_j K_{i,j}.$$

We show that there is no duality gap, i.e., $primal(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*) - dual(\boldsymbol{\alpha}^*) = 0$, by letting $\boldsymbol{w}^* = \sum_{t=1}^N \alpha_t^* r^{(t)} \Phi(\boldsymbol{x}^{(t)})$, $b^* = \rho$, and $\xi_t^* = \max\{0, g_t^* - r^{(t)}\rho\}$.
$primal(w^*, b^*, \xi^*) - dual(\alpha^*)$
$= (\frac{1}{2}||w^*||^2 + C\sum_{t=1}^N \xi_t^*) - (\sum_{t=1}^N \alpha_t^* - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N r^{(i)} r^{(j)} \alpha_i^* \alpha_j^* K_{i,j})$
$= C\sum_{t=1}^N \xi_t^* - \sum_{t=1}^N \alpha_t^* + \sum_{i=1}^N \sum_{j=1}^N r^{(i)} r^{(j)} \alpha_i^* \alpha_j^* K_{i,j}$ (due to $\frac{1}{2}||w^*||^2 = \frac{1}{2}(\alpha^*)^\top \tilde{K}\alpha^*$)
$= C\sum_{t=1}^N \xi_t^* - \sum_{i=1}^N \alpha_t^* g_t^*$ (due to $g_t^* = 1 - r^{(t)}\sum_{i=1}^N r^{(i)}\alpha_i^* K_{i,t}$)
$= \sum_{t=1}^N (C\xi_t^* - \alpha_t^* g_t^*)$
$= \sum_{t=1}^N (-\rho r^{(t)} \alpha_t^*)$ (due to Condition (3))
$= 0$ (due to $r^\top \alpha = 0$).
The optimality holds.

6. Show that the *Area Under the ROC Curve* (AUC) is equal to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one.
(Hint: by partitioning the AUC horizontally)

Answer :



(a) Consider the horizontal partition of AUC based on each positive instance. The height of each partition (shown as the shaded area in the above figure) is the probability that a positive instance is chosen. On the other hand, the width of the partition is the conditional probability that given a positive instance is chosen, a randomly chosen negative instance is ranked after that positive instance (note that each negative instance ranked before that positive instance contributes to the portion of the horizontal bar before the shaded area). Therefore, by summing up all the partitions, we have the joint probability that a randomly chosen positive instance is ranked before a randomly chosen negative one.