سوال یک

بخش اول

پاسخ سوال ۳ *Properties of Kernels*

ابتدا $g(x+y,x+y)$ را ساده می‌کنیم.

$$g(x+y,x+y)$$

$$= g(x,x+y) + g(y,x+y)$$

$$= g(x+y,x) + g(x+y,y)$$

$$= g(x,x) + g(y,x) + g(x,y) + g(y,y)$$

$$= g(x,x) + 2g(x,y) + g(y,y)$$

مشابه روند بالا $g(x-y,x-y)$ را هم ساده می‌کنیم.

$$g(x-y,x-y)$$

$$= g(x,x-y) + g(-y,x-y)$$

$$= g(x-y,x) + g(x-y,-y)$$

$$= g(x,x) + g(-y,x) + g(x,-y) + g(-y,-y)$$

$$= g(x,x) - 2g(x,y) + g(y,y)$$

در نهایت نتایج فوق را در عبارت اصلی جایگزین می‌کنیم.

$$h(x,y) = \frac{1}{4}(g(x,x) + 2g(x,y) + g(y,y) - g(x,x) + 2g(x,y) - g(y,y)) = \frac{1}{4}(4g(x,y)) = g(x,y)$$

از معتبر بودن هسته $g(x,y)$ نتیجه می‌گیریم $h(x,y)$ هم یک هسته معتبر است.

سوال یک

بخش دوم

پاسخ سوال ۲ *Kernel*

۱.

$$k_1(x_1, x_2) = \phi_1(x_1)^T \phi_1(x_2)$$

$$k_2(x_1, x_2) = \phi_2(x_1)^T \phi_2(x_2)$$

$$k_3(x_1, x_2) = k_1(x_1, x_2) + k_2(x_1, x_2)$$

$$= \phi_1(x_1)^T \phi_1(x_2) + \phi_2(x_1)^T \phi_2(x_2)$$

$$= (\phi_1(x_1), \phi_2(x_1))^T (\phi_1(x_2), \phi_2(x_2))$$

حال تعریف می‌کنیم: $\forall x: \phi_3(x) = (\phi_1(x), \phi_2(x))$. در واقع ϕ_3 را از *concat* کردن ϕ_1 و ϕ_2 بدست می‌آوریم.

بنابراین داریم: $\phi_3(x_1)^T \phi_3(x_2) = k_3(x_1, x_2)$ و طبق تعریف k_3 یک هسته معتبر است.

۲. مشابه قسمت قبل

$$k_4(x_1, x_2) = k_1(x_1, x_2) k_2(x_1, x_2)$$

$$k_1(x_1, x_2) = \phi_1(x_1)^T \phi_1(x_2) = \sum_i \phi_{1i}(x_1) \phi_{1i}(x_2)$$

$$k_2(x_1, x_2) = \phi_2(x_1)^T \phi_2(x_2) = \sum_j \phi_{2j}(x_1) \phi_{2j}(x_2)$$

سوال اول بخش سوم:

پرسش اول: ۳. چون K کرنل معتبر است لذا:

$$K(x, x') = \phi(x)^T \phi(x')$$

$$K'(A, B) = \sum_{x \in A,\, x' \in B} K(x, x') = \sum_{x \in A,\, x' \in B} \phi(x)^T \phi(x') = \Big(\sum_{x \in A} \phi(x)\Big)^T \Big(\sum_{x' \in B} \phi(x')\Big)$$

$$\rightarrow \phi_1(A) = \sum_{x \in A} \phi(x) \quad \rightarrow K'(A, B) = \langle \phi_1(A), \phi_1(B) \rangle$$

سوال دوم

پرسش دوم: ۱. همانطور که خواندید در SVM برای نزدیکترین داده به hyperplane داریم:

$$y_\cdot(w^Tx_\cdot+b)=1 \rightarrow |w^Tx_\cdot+b|=1$$

از طرفی برای هر نقطه، فاصله آن از hyperplane برابر است با:

$$\frac{|w^Tx+b|}{\|w\|}$$

لذا نزدیکترین نقطه فاصله‌اش $\frac{1}{\|w\|}$ است یعنی مقدار مارجین $\frac{1}{\|w\|}$ است.

$$L(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jy^{(i)}y^{(j)}K(x^{(i)},x^{(j)}) \qquad ①$$

همچنین: $$L(a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}a_i\{y^{(i)}(w^Tx^{(i)}+b)-1\} \qquad ②$$

فرض کنید a_i ها جواب بهینه باشند، چون شرط KKT برقرار خواهد بود، خواهیم داشت:

$$\forall i:\ a_i\{y^{(i)}(w^Tx^{(i)}+b)-1\}=0$$

پس در ② ، عبارت دوم حذف خواهد شد و لذا داریم:

$$L(a)=\frac{1}{2}\|w\|^2 \qquad (*)$$

از طرفی چون ثابت شد $w=\sum a_iy^{(i)}x^{(i)}$ پس در ① داریم:

$$L(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\|\omega\|^2$$

$$\xrightarrow{(*)} \frac{1}{2}\|\omega\|^2 = \sum_{i=1}^{n} a_i - \frac{1}{2}\|\omega\|^2 \rightarrow \|\omega\|^2 = \sum_{i=1}^{n} a_i$$

$$\text{مارجین} = \frac{1}{\|\omega\|} \rightarrow \boxed{\text{مارجین} = \frac{1}{\sqrt{\sum_{i=1}^{n} a_i}}} \quad \blacksquare$$

۲. همانطور که به یاد دارید در SVM، $\|\omega\|^2$ (که عکس مارجین بود) را ماکسیمم می‌کردیم. وقتی یک ویژگی جدید اضافه شود که $\|\omega^*\|^2$ بهینه، بیشتر نشود، ω مربوط به آن مؤلفه را می‌توان 0 در نظر گرفت چون $\|\omega\|^2 = \sum \omega_i^2$ یعنی به آن ویژگی ضریب صفر می‌دهد و آن ویژگی را در نظر نمی‌گیرد. پس نسبت به ویژگی‌های بی‌ربط، robust است.

۳.

$$\text{مسئله SVM}: \min_{\omega, b} \frac{1}{2}\|\omega\|^2$$

$$\text{s.t.} \quad \left.\begin{aligned} \omega^T x_1 + b &= +1 \\ \omega^T x_2 + b &= -1 \end{aligned}\right\} \#$$

$$\xrightarrow{\text{لاگرانژ}} \min_{\omega, b} \Big\{ \underbrace{\frac{1}{2}\|\omega\|^2 + \lambda(\omega^T x_1 + b - 1) + \eta(\omega^T x_2 + b + 1)}_{L(\omega, b, \lambda, \eta)} \Big\}$$

$$\rightarrow \begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \rightarrow \omega + \lambda x_1 + \eta x_2 = 0 \\ \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \lambda + \eta = 0 \end{cases} \Rightarrow \omega = \eta(x_1 - x_2) \quad (1)$$

$$\#_{\text{جمع}} \rightarrow \omega^T(x_1 + x_2) = -2b \xrightarrow{(1)} b = \frac{\eta}{2}(x_2 - x_1)^T(x_1 + x_2) \quad (2)$$

فاصله ابرصفحه طبق آنچه از مبدأ $\frac{b}{\|w\|}$ است، بنابراین طبق ① و ② داریم:

$$\frac{b}{\|w\|} = \frac{\eta/2\left[\|x_2\|^2 - \|x_1\|^2\right]}{\eta\,\|x_1 - x_2\|} = \frac{1}{2}\,\boxed{\frac{\|x_2\|^2 - \|x_1\|^2}{\|x_1 - x_2\|}} \quad \blacksquare$$

سوال سوم

1. In the above cost function, ϵ defines the region inside which errors are ignored. The loss function defined above is non-differentiable due to the absolute value in the loss function. We can introduce slack variables ξ and ξ^* to account for errors in points that lie outside the ϵ tube as follows. (These are similar to the slack variables used in classification.)

$$y_i - \langle w, x_i \rangle - \epsilon \leq \xi_i \tag{1}$$

$$\langle w, x_i \rangle - y_i - \epsilon \leq \xi_i^* \tag{2}$$

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, n \tag{3}$$

   Thus, we can rewrite the primal form as,

$$\min_{w \in \mathbb{R}^m, \xi \in \mathbb{R}^n, \xi^* \in \mathbb{R}^n} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

   s.t. equations (1)-(3) are satisfied.

   Rubric: 2 points for constraints. 1 point for the objective. Partial grade if there is a mistake using one of the slack variables etc.

2. Having the above constraints and objective, the Lagrangian function can be written as follows.

$$\begin{aligned} L = L(w, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) := \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{n}(\beta_i \xi_i + \beta_i^* \xi_i^*) \\ & - \sum_{i=1}^{n} \alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i \rangle) \\ & - \sum_{i=1}^{n} \alpha_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i \rangle), \end{aligned} \tag{4}$$

   where the Lagrange multipliers have to satisfy the positivity constraints,

$$\alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0, \quad (i = 1, \ldots, n).$$

   Rubric: 2 points for the Lagrangian. Subtract 1 if the constraints are missing. Partial grade if minor typo in equation.

3. We need to solve the following min-max problem:

$$(w,\xi,\xi^*,\alpha,\alpha^*,\beta,\beta^*) = \min_{w,\xi,\xi^*}\max_{\alpha,\alpha^*,\beta,\beta^*} L(w,\xi,\xi^*,\alpha,\alpha^*,\beta,\beta^*) \tag{5}$$

$$= \max_{\alpha,\alpha^*,\beta,\beta^*}\min_{w,\xi,\xi^*} L(w,\xi,\xi^*,\alpha,\alpha^*,\beta,\beta^*) \tag{6}$$

[Similarly as we discussed in class for classification, the max and min can be switched because the so-called strong duality holds for quadratic problems.]

Taking the derivative of L w.r.t the primal variables (w, ξ_i and ξ_i^*), we get

$$\partial_w L = w - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i = 0$$

$$\partial_{\xi_i} L = C - \alpha_i - \beta_i = 0$$

$$\partial_{\xi_i^*} L = C - \alpha_i^* - \beta_i^* = 0$$

From the last two equations we have that

$$0 \le \beta_i = C - \alpha_i$$

$$0 \le \beta_i^* = C - \alpha_i^*$$

Substituting the results back into the Lagrangian (4), we get

$$\begin{aligned}
L = \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{i=1}^{n}(\beta_i\xi_i + \beta_i^*\xi_i^*) - \sum_{i=1}^{n}\alpha_i(\epsilon+\xi_i-y_i+\langle w,x_i\rangle) - \sum_{i=1}^{n}\alpha_i^*(\epsilon+\xi_i^*+y_i-\langle w,x_i\rangle) \\
= & \frac{1}{2}\|\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i\|^2 + \sum_{i=1}^{n}\xi_i\underbrace{(C-\beta_i-\alpha_i)}_{0} + \sum_{i=1}^{n}\xi_i^*\underbrace{(C-\beta_i^*-\alpha_i^*)}_{0} - \epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \\
& + \sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\quad\underbrace{\langle w,x_i\rangle}_{\langle\sum_{j=1}^{n}(\alpha_j-\alpha_j^*)x_j,x_i\rangle} \\
= & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle - \epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)
\end{aligned}$$

Therefore, the dual problem is

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle - \epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \tag{7}$$

$$s.t. \quad \alpha_i,\alpha_i^* \in [0,C] \tag{8}$$

[Note that if you use $\langle w,x\rangle + b$ instead of $\langle w,x\rangle$, then you have an extra constraint: $\sum_{i=1}^{n}(\alpha_i-\alpha_i^*) = 0$.]

<u>Rubric</u>: 2 points for the derivatives of L, 1 point for (7), 1 point for (8), 1 point for explaining the details well. Partial grade for minor mistakes in derivation.

4. The problem has a quadratic objective with linear constraints, therefore it can be solved by a Quadratic Programming solver.

   Rubric: 1 point for correct answer. No partial credit.

5. The KKT complementary slackness conditions are as follows. In the optimal solutions of (5) and (6) we have that

$$\alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i \rangle) = 0 \tag{9}$$
$$\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i \rangle) = 0 \tag{10}$$
$$\beta_i \xi_i = 0 \tag{11}$$
$$\beta_i^* \xi_i^* = 0 \tag{12}$$

   for all $i = 1, \ldots, n$.

   Equation (9) implies that if $\alpha_i > 0$, then $(\epsilon + \xi_i - y_i + \langle w, x_i \rangle) = 0$.

   Now, if $\xi_i = 0$, then it implies that x_i is on the border of the ϵ-tube, therefore x_i is a margin support vector. If $\xi_i > 0$, then it means we are outside of the ϵ-tube. These x_i vectors are the non-margin support vectors. Similar reasoning holds for ξ_i^* and α_i^*.

   Rubric: 1 point for margin support vectors. 1 point for non-margin support vectors.

6. Since for prediction we use $f(x) = \langle w, x \rangle$, and $w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i$, therefore

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle$$

   Rubric: 1 point for the correct prediction, 1 point for reasoning.

7. Yes, we can write the above equation in the kernel form.

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x)$$

   Rubric: 1 point for the correct answer.

8. ϵ plays the opposite role of C. The smaller the value of ϵ, the harder SVM tries to fit smaller errors around the learnt SVM function, and leads to a more complex model. Smaller ϵ also leads to a less sparse solution (more support vectors).
   Small ϵ - More complex model. Low Bias, High variance.
   Large ϵ - Less complex model. High Bias, Low Variance.

   Rubric: 1 point for reasoning. 1 point for mentioning the relationship with Bias/Variance

9. C plays a similar role as it did during classification. It is a measure of how strongly we penalize errors. It should be tuned for bias vs variance with model selection. The higher the value of C, the larger the tendency of SVM to penalize errors and overfit the data. The lower the value of C, the larger its tendency to ignore errors and underfit the data.
   Large C - More complex model. Low Bias, High variance.
   Small C - Less complex model. High Bias, Low Variance.

سوال چهارم

1.1

In the linearly separable case, if one of the training samples is removed:(1) If the point is not a support vector, then the margin remains unchanged. (2) If the point is a support vector, then the margin length can become larger and move towards the point which is removed if the point was the only support vector or remain unchanged otherwise. Logistic regression focuses on maximizing the probability of the data. The farther the data lies from the separating hyperplane, the happier LR is as opposed to SVM which tries to explicitly find the maximum margin. If a point is not a support vector, it doesn't really matter. Since LR is a density estimation technique, each point will carry some weight and have some effect on the decision boundary

1.2

The hinge loss is the upper bound on the number of misclassified instances. Here, we choose the hinge loss function as $h(z) = \max(0, 1 - z)$. The primal optimization of SVM is given by

$$\underset{w,\xi_i}{\text{minimize}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

Now the slack variable appears in 1 constraint and we try to minimize that, i.e., we satisfy one of the two constraints given, i.e.,

$$y_i \left(w^T x_i\right) \geq 1 - \xi_i$$

or

$$\xi_i \geq 0$$

The slack variable is the tighter/larger one of the two numbers. So it can either be zero or $1 - y_i \left(w^T x_i\right)$ Thus,

$$\xi_i = \max\left(0, 1 - y_i \left(w^T x_i\right)\right)$$

which is the hinge loss function. Hence,

$$\xi_i = \max\left(0, 1 - y_i \left(w^T x_i\right)\right) = h\left(y_i \left(w^T x_i\right)\right)$$

And we know that the hinge loss is the upper bound on the number of misclassified instances. Thus, the upper bound is given by

$$\Longrightarrow \sum_{i=1}^{n} h\left(y_i \left(w^T x_i\right)\right)$$

or simply

$$\sum_{i=1}^{n} (\xi_i > 1)$$

1.3

C is the trade-off parameter that tells us whether we would rather have a small norm of w, meaning a large margin, or rather have no violations of the margin constraints, meaning the small sum of hinge loss. (1) When $C \to 0$, more emphasis is given to finding the largest margin irrespective of several noises, i.e. no misclassification will be penalized. (2) When $C \to \infty$, we put a higher and higher weight on violations of margin constraints, so we find a hyperplane where the required slack is minimized even at the expense of the margin.

1.4

When two classes are linearly separable, both Hard SVM and Logistic Regression can always find a solution. The major difference is that Logistic Regression finds a decision boundary that maximizes its likelihood function while Hard SVM finds a decision boundary with a maximal margin.

1.5

When the two classes are not linearly separable, Logistic Regression will still find a decision boundary that maximizes its likelihood function. For Soft SVM, it will find a decision boundary that best balances the margin and errors. (Note: the key difference between SVM and Logistic Regression is that SVM is more of a geometric-motivated model while Logistic Regression is more of a probability-motivated model.)

ادامه‌ی پاسخ سوال یک، بخش دوم

حال دو عبارت فوق را در هم ضرب می‌کنیم:

$$\sum_i \sum_j \phi_{1i}(x_1)\phi_{1i}(x_2)\phi_{2j}(x_1)\phi_{2j}(x_2)$$

حال تعریف می‌کنیم:

$$\phi'_{ij}(x) = \phi_{1i}(x)\phi_{2j}(x)$$

در نتیجه داریم:

$$k_4(x_1, x_2) = \sum_{i,j} \phi'_{ij}(x_1)\phi'_{ij}(x_2) = \phi'(x_1)^T\phi'(x_2)$$

بنابراین k_4 طبق تعریف هسته‌ای معتبر است.

۳. اگر k یک هسته معتبر باشد به ازای یک اسکالر مثبت c ، ck نیز یک هسته معتبر است.

$$k'(x_1, x_2) = ck(x_1.x_2) = c\phi(x_1)^T\phi(x_2) = \sqrt{c}\phi(x_1)^T\sqrt{c}\phi(x_2) = \phi'(x_1)^T\phi'(x_2)$$

همچنین با استقرا بر روی بخش ب، اگر k هسته معتبر باشد k^n نیز هسته‌ای معتبر است.

می دانیم بسط تیلور e^x به صورت زیر است :

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

بنابراین می‌توان برای هسته k_1 اینطور نوشت:

$$e^{k_1(x_1,x_2)} = \sum_{n=0}^{\infty} \frac{k_1(x_1, x_2)^n}{n!}.$$

همان‌طور که گفتیم، $k_1(x_1, x_2)^n$ یک هسته معتبر است. ضرب یک عدد مثبت در هسته نیز خود یک هسته معتبر است. بنابراین $\frac{k_1(x_1,x_2)^n}{n!}$ هسته‌ای معتبر است. از طرفی حاصل جمع هسته‌های معتبر خود یک هسته معتبر است (بخش ۱). در نتیجه عبارت نهایی، هسته‌ای معتبر است.

۴. بسط تیلور $\frac{1}{1-\alpha}$ به صورت زیر است:

$$\frac{1}{1-\alpha} = 1 + \alpha + \alpha^2 + \alpha^3 + ... = \sum_{n=0}^{\infty} \alpha^n$$

همچنین $k(x_1, x_2) = x_1^T x_2$ یک هسته معتبر است. زیرا در صورتی که $\phi(x) = x$ آن‌گاه داریم:

$$k(x_1, x_2) = x_1^T x_2 = \phi(x_1)^T\phi(x_2) \Rightarrow valid\ kernel$$

بنابراین عبارت $\frac{1}{1-k(x_1,x_2)}$ مجموع عباراتی به فرم k^n است که طبق اثبات بخش قبل هسته‌ای معتبر است.