M146 – Homework set #3

Name: Yangyang Mao   UID: 504945234

Problem 1

(a) $\because K(x,z) = (x^T z)^d$ is a valid function

$\therefore \forall$ vector v, $v^T K(x,z) v \geq 0$, $v^T (x^T z)^d v \geq 0$

$K_{new}(x,z) = \left(\sum_{i=1}^{n} \sqrt{x_i}\sqrt{z_i}\right)^d$

$v^T K_{new}(x,z) v = v^T (A^T B)^d v \geq 0$

where $A = \begin{bmatrix} \sqrt{x_1} \\ \vdots \\ \sqrt{x_n} \end{bmatrix}$ $B = \begin{bmatrix} \sqrt{z_1} \\ \vdots \\ \sqrt{z_n} \end{bmatrix}$

Thus, $K_{new}(x,z) = \left(\sum_{i=1}^{n} \sqrt{x_i}\sqrt{z_i}\right)^d$ is a valid kernel.

(b) $\because K_1$ and K_2 are both valid kernels

$\therefore \forall$ vector v, $v^T K_1 v \geq 0$ $v^T K_2 v \geq 0$

Since $\alpha, \beta \in R^+$

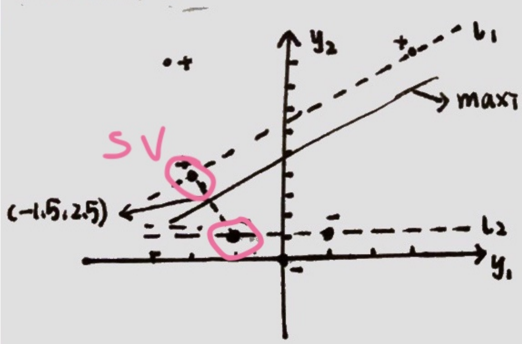
$\therefore \forall$ vector v, $v^T \alpha K_1 v \geq 0$, $v^T \beta K_2 v \geq 0$

$\therefore$ $\cancel{v^T (K_{new}) v}$ $v^T \alpha K_1 v + v^T \beta K_2 v \geq 0$

$\therefore v^T (\alpha K_1 + \beta K_2) v \geq 0$

$\therefore K_{new}(x,z) = \alpha K_1(x,z) + \beta K_2(x,z)$ is a valid kernel

Problem 2



maximum margin separating hyperplane in R^2

$Y_2 - 2.5 = \frac{1}{3}(y_1 + 1.5)$

separating hyperplane must between $l_1 \sim l_2$

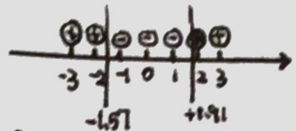
We take a midpoint of $(1,1)$ and $(-2,4)$. And draw a line which is perpendicular to the line through $(1,1)$ and $(-2,4)$. Slop $= \frac{1}{3}$.

Thus, maximum margin separating hyper plane in R^2 is $3y_2 - y_1 - 9 = 0$

margin $(w,b) = \min_n \frac{y_n[w^T \phi(x_n) + b]}{\|w\|_2} = \frac{\sqrt{10}}{2}$

$w = \begin{bmatrix} -9 \\ -1 \\ 3 \end{bmatrix} \begin{array}{l} \leftarrow w_0 \\ \leftarrow w_1 \\ \leftarrow w_2 \end{array}$

(c)



Decision boundary of the separating hyperplane in original R^1 feature space

(d) $\because$ support vectors : u_1, u_2

$\therefore$ only a_1 and a_2 are non-zero. suppose $a_1 = a_2 = a$

$$L(a) = \sum_{n=1}^{N} a_n + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m K(x_n, x_m)$$

$$= a_1 + a_2 - \frac{1}{2}\left(a_1^2 (u_1 \cdot u_1) - 2a_1 a_2 (u_1 \cdot u_2) + a_2^2 (u_2 \cdot u_2)\right)$$

$$= 2a - 5a^2$$

$$L'(a) = 0 \Rightarrow a = \frac{1}{5}$$

$$\sum_{n=1}^{|SV|} \alpha_n y_n K(x_i, u_n) + b = 1$$

$\therefore b = -9/5$

(e) The hyperplane will not change, since the traning point is classified correctly.

Problem 3

(a) $X = \begin{bmatrix} a \\ e \end{bmatrix}$. suppose $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

constrain: $y_n w^T x_n \geq 1$. equals to $-1 \cdot [w_1, w_2] \begin{bmatrix} a \\ e \end{bmatrix} \geq 1$

~~$\therefore w_1 a + w_2 e \leq -1$~~

~~Let $w_1 a + w_2 e = -1 \Rightarrow w_2 = \frac{w_1 a + 1}{-e}$~~

~~$w^* = \begin{bmatrix} w_1 \\ \frac{-w_1 a - 1}{e} \end{bmatrix}$~~

~~$\frac{1}{2}\|w\|^2 = \frac{1}{2}\left(|w_1| + \left|\frac{-w_1 a - 1}{e}\right|\right)^2$~~

We move the constraint into objective function and introduce a Lagrange multiplier:

$$\max_{\lambda} \min_{w} \frac{\|w\|^2}{2} + \lambda(1 + w^T x), \quad \lambda \geq 0$$

$y = -1$. And take derivative above wrt w:

$$w + \lambda x = 0$$

There is only a single point x, Hence x must be support vector.

$\therefore 1 = y(w^T x + b)$ for any support vectories on margin. And here hard margin:

$$1 + wx = 0$$

$$\therefore \begin{cases} w_1 + \lambda a = 0 \\ w_2 + \lambda e = 0 \\ 1 + w_1 a + w_2 e = 0 \end{cases}$$

$$\therefore w^* = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= -\frac{1}{a^2 + e^2}\begin{bmatrix} a \\ e \end{bmatrix}$$

(b) $x_1 = (1,1)^T$, $y_1 = 1$. $x_2 = (1,0)^T$, $y_2 = -1$

These two points lie on hard margin, so:

$1 = y(w^T x + b)$

$$\begin{cases} 1-(w_1+w_2)=0 \\ 1+w_1=0 \end{cases} \Rightarrow \begin{cases} w_1 = -1 \\ w_2 = 2 \end{cases}$$

$\therefore w = [-1, 2]^T$

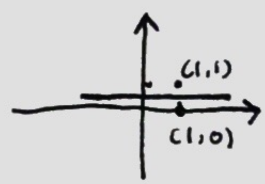
(c) In this case, $b \neq 0$. $1 = y(w^T x + b)$ Since data only vary in vertical axis, the boundary will be horizontal $\Rightarrow w_1 = 0$

$$\begin{cases} 1-(w_1+w_2)-b=0 \\ 1+w_1+b=0 \end{cases} \Rightarrow \begin{cases} w_1 = 0 \\ w_2 = 2 \\ b = -1 \end{cases} \Rightarrow w^* = \begin{bmatrix} 0 \\ 2 \end{bmatrix},\ b^* = -1$$

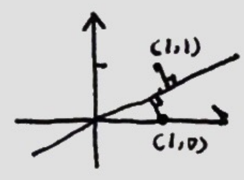
~~$\text{margin}(w,b) = \frac{1}{\|w\|_2}$~~

~~with offset: $\text{margin} = \frac{1}{\sqrt{4}} = \frac{1}{2}$~~

~~without offset: $\text{margin} = \frac{1}{\sqrt{1+4}} = \frac{1}{\sqrt{5}}$~~



with offset b.



without offset b.

Thus the margin is larger if the classifier has offset parameter b.

with offset: $\text{margin} = \frac{1}{2}$. without offset: $\text{margin} = \frac{1}{\sqrt{5}}$

Problem 4

0.2

(b) Stratified splits are important, since the fundamental assumption of most ML algorithms is that the training set should represent the test set. the training and test data are drawn from the same underlying distributions. If the ratio of positive to negative examples (the class balance) differs significantly between the training and test sets (across folds), the assumption will not hold.

(d)

C	accuracy	F1-score	AUROC	Precision	Sensitivity	Specificity
0.001	0.708941953964	0.829682822742	0.5	0.708941953964	1.0	0.0
0.01	0.710743755766	0.830562800464	0.503125	0.710235975258	1.0	0.00625
0.1	0.806032676165	0.875472682956	0.71878715957	0.835683713447	0.929430379747	0.508143939394
1	0.814627111309	0.87486483275	0.753111334868	0.856161851838	0.90167721519	0.604545454545
10	0.818182737099	0.876562152887	0.759171940928	0.859521253319	0.90167721519	0.616666666667
100	0.818182737099	0.876562152887	0.759171940928	0.859521253319	0.90167721519	0.616666666667
Best C	10;100	10;100	10;100	10;100	0.001;0.01	10;100

As we can see from the table shown above, accuracy, F1-score, AUROC, precision and Specificity increase with C increasing in range [0.001,100], they all achieve maximum value when C =10 and 100. In contrast, Sensitivity is decreasing with C increasing in range [0.001,100], and it achieves maximum value when C=0.001 and 0.01.

0.3

(a)The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The behavior of the model is very sensitive to the gamma parameter. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. When gamma is very small, the model is too constrained and cannot capture the complexity or “shape” of the data.

(b)gamma ranges in 10^{-3},10^{-2},10^{-1},1,10,10^{2},10^{3}, and C ranges in 10^{-3},10^{-2},10^{-1},1,10,10^{2},10^{3},10^{4}. Since gamma and C must be larger than 0, and I use a relative large grid to search the best C and gamma, and this range is usually sufficient. If the best parameters lie on the boundaries of the grid, I extended in that direction in a subsequent search.

(c)

Metric	score	C	γ
accuracy	0.816460518673	100	0.01
F1-score	0.876285736173	100	0.01
AUROC	0.756141637898	1000	0.01
precision	0.858265970904	100	0.01
sensitivity	1.0	0.001	1000
specificity	0.610606060606	1000	0.01

From the table shown above, we can see that accuracy, F1-score, precision achieve maximum value when C=100 and gamma=0.01. AUROC and specificity achieves maximum value when C=1000 and gamma=0.01. There exists many sets of C and gamma that can make sensitivity achieve maximum value, which is 1.0, the one I extracted is C=0.001 and gamma=1000.

0.4

(a)I choose C=100 as parameter for linear-kernel SVM and C=100, gamma=0.01 for RBF-kernel SVM. Since in 0.2 we see that performance measures achieves maximum value when C=10 or 100 in linear-kernel SVM. In 0.3 we see that performance measures achieves maximum or relative high value when C=100, gamma=0.01.

(c)

Performance metric	Linear kernel (C=100;10)	RBF kernel (C=100; γ=0.01)
accuracy	0.742857142857	0.757142857143
F1-score	0.4375	0.451612903226
AUROC	0.625850340136	0.636054421769
precision	0.636363636364	0.7

sensitivity	0.333333333333	0.333333333333
specificity	0.918367346939	0.938775510204

From the table shown above, we can see that in the listed 6 performance metrics, the results achieved by RBF-kernel SVM are slightly better than results achieved by linear-kernel SVM.