005230642 Rui Xu

1.(a). $k_\beta(\vec{x},\vec{z}) = (1+\beta\vec{x}^T\vec{z})^3 = (1+\beta\vec{x}^T\vec{z})(1+2\beta\vec{x}^T\vec{z}+\beta^2(\vec{x}^T\vec{z})^2)$

$= 1+2\beta\vec{x}^T\vec{z}+\beta^2(\vec{x}^T\vec{z})^2+\beta\vec{x}^T\vec{z}+2\beta^2(\vec{x}^T\vec{z})^2+\beta^3(\vec{x}^T\vec{z})^3$

$= 1+3\beta\vec{x}^T\vec{z}+3\beta^2(\vec{x}^T\vec{z})^2+\beta^3(\vec{x}^T\vec{z})^3$

$\vec{x} = [x_1\ x_2]^T \quad \vec{z} = [z_1\ z_2]^T$

$\Rightarrow k_\beta(\vec{x},\vec{z}) = 1+3\beta(x_1z_1+x_2z_2)+3\beta^2(x_1z_1+x_2z_2)^2+\beta^3(x_1z_1+x_2z_2)^3$

$= 1+3\beta x_1z_1+3\beta x_2z_2+3\beta^2(x_1^2z_1^2+x_2^2z_2^2+2x_1x_2z_1z_2)$

$= 1+3\beta x_1z_1+3\beta x_2z_2+3\beta^2x_1^2z_1^2+3\beta^2x_2^2z_2^2+6\beta^2x_1x_2z_1z_2+\beta^3x_1^3z_1^3+\beta^3x_2^3z_2^3+$

$3\beta^3x_1^2x_2z_1^2z_2+3\beta^3x_1x_2^2z_1z_2^2$

(b). $\phi_\beta(\vec{x}) = [1\ \sqrt{3\beta}x_1\ \sqrt{3\beta}x_2\ \sqrt{3}\beta x_1^2\ \sqrt{6}\beta x_1x_2\ \sqrt{3}\beta x_2^2\ \sqrt{\beta^3}x_1^3\ \sqrt{3\beta^3}x_1^2x_2\ \sqrt{3\beta^3}x_1x_2^2\ \sqrt{\beta^3}x_2^3]^T$

(c). $K(\vec{x},\vec{z})$ and $k_\beta(\vec{x},\vec{z})$ have similar feature map.

The difference is that $k_\beta(\vec{x},\vec{z})$ has a parameter β, which scales each term in the $\phi_\beta(\vec{x})$.

The term with ith order is scaled by $\beta^{\frac{i}{2}}$ $(i>0)$

$\Rightarrow$ The weight of different order terms is intluenced by the value of β

① If $\beta=1$, $k_\beta(\vec{x},\vec{z}) = (1+\beta\vec{x}^T\vec{z})^3 = (1+\vec{x}^T\vec{z})^3$ is the same with $K(\vec{x},\vec{z}) = (1+\vec{x}^T\vec{z})^3$

$\phi_\beta(\vec{x}) = [1\ \sqrt{3}x_1\ \sqrt{3}x_2\ \sqrt{3}x_1^2\ \sqrt{6}x_1x_2\ \sqrt{3}x_2^2\ x_1^3\ \sqrt{3}x_1^2x_2\ \sqrt{3}x_1x_2^2\ x_2^3]^T$

② If $0<\beta<1$, $0<\beta^{\frac{i}{2}}<1$, the larger i is, the smaller $\beta^{\frac{i}{2}}$ is $\Rightarrow \beta^{\frac{1}{2}}>\beta>\beta^{\frac{3}{2}}$

$\Rightarrow$ Higher order terms in $\phi_\beta(\vec{x})$ have less weight.

Lower order terms in $\phi_\beta(\vec{x})$ have more weight.

③ If $\beta>1$, $\beta^{\frac{i}{2}}>1$, the larger i is, the larger $\beta^{\frac{i}{2}}$ is $\Rightarrow \beta^{\frac{1}{2}}<\beta<\beta^{\frac{3}{2}}$

$\Rightarrow$ Higher order terms in $\phi_\beta(\vec{x})$ have more weight.

Lower order terms in $\phi_\beta(\vec{x})$ have less weight.

④ If $\beta \to \infty$, $\beta^{\frac{i}{2}}$ with the highest i will be much larger than the other $\beta^{\frac{i}{2}}$

$\Rightarrow \beta^{\frac{3}{2}} >> \beta >> \beta^{\frac{1}{2}}$

$\Rightarrow$ only the constant and cubic order terms in $k_\beta(\vec{x}, \vec{z})$ remain.

$\Rightarrow \phi_\beta(\vec{x}) = [1 \ \sqrt{\beta^3}x_1^3 \ \sqrt{3\beta^3}x_1^2x_2 \ \sqrt{3\beta^3}x_1x_2^2 \ \sqrt{\beta^3}x_2^3]^T$

⑤ If $\beta \to 0$, $\beta^{\frac{i}{2}}$ with the lowest i will be much larger than the other $\beta^{\frac{i}{2}}$

$\Rightarrow \beta^{\frac{1}{2}} >> \beta >> \beta^{\frac{3}{2}}$

$\Rightarrow$ only the constant and first order terms in $k_\beta(\vec{x}, \vec{z})$ remain.

$\Rightarrow \phi_\beta(\vec{x}) = [1 \ \sqrt{3\beta}x_1 \ \sqrt{3\beta}x_2]^T$

2.(a). $\vec{w}=[w_1\ w_2]^T$

$y_n\vec{w}^T\vec{x}_n \geq 1$

$\Rightarrow \begin{cases} 1\,(w_1+w_2)\geq 1 \\ (-1)(w_1+0)\geq 1 \end{cases} \Rightarrow \begin{cases} w_1+w_2\geq 1 \Rightarrow w_2\geq 1-w_1>0 \\ w_1\leq -1 \Rightarrow w_1<0 \end{cases}$

minimize: $J(\vec{w})=\frac{1}{2}\|\vec{w}\|^2=\frac{1}{2}(w_1^2+w_2^2)\geq\frac{1}{2}(w_1^2+(1-w_1)^2)=w_1^2-w_1+\frac{1}{2}$

$\geq w_1^2-w_1+\frac{1}{2}\Big|_{w_1=-1}=1-(-1)+\frac{1}{2}=\frac{5}{2}$

$\Rightarrow w_2\geq 1-w_1=2$

$\Rightarrow w_1, w_2$ with minimum absolute value will minimize $J(\vec{w})$

$\Rightarrow \begin{cases} w_1=-1 \\ w_2=2 \end{cases}$

$\Rightarrow \vec{w}^*=[-1\ 2]^T$

(b). $\vec{w}=[\vec{w}_1\ \vec{w}_2]^T$ b

$y_n\vec{w}^T\vec{x}_n \geq 1$

$\Rightarrow \begin{cases} 1\,(w_1+w_2+b)\geq 1 \\ (-1)(w_1+0+b)\geq 1 \end{cases} \Rightarrow \begin{cases} w_1+w_2+b\geq 1 \Rightarrow w_2\geq 1-w_1-b \\ w_1+b\leq -1 \Rightarrow w_1\leq -1-b \end{cases}$

minimize: $J(\vec{w})=\frac{1}{2}\|w\|^2=\frac{1}{2}(w_1^2+w_2^2)\geq\frac{1}{2}(w_1^2+(1-w_1-b))^2=w_1^2+(b-1)w_1+\frac{1}{2}(b-1)^2$

If $-\frac{b-1}{2}>-1-b \Rightarrow b>-3$

$\Rightarrow J(\vec{w})\geq w_1^2+(b-1)w_1+\frac{1}{2}(b-1)^2\Big|_{w_1=(-1-b)}=\frac{1}{2}b^2+b+\frac{5}{2}\geq\frac{1}{2}b^2+b+\frac{5}{2}\Big|_{b=-1}=2$

$\Rightarrow w_1=-1-b\big|_{b=-1}=0$

$w_2\geq 1-w_1-b=2$

If $-\frac{b-1}{2}\leq -1-b \Rightarrow b\leq -3$

$\Rightarrow J(\vec{w})\geq w_1^2+(b-1)w_1+\frac{1}{2}(b-1)^2\Big|_{w_1=-\frac{b-1}{2}}=\frac{b^2}{4}-\frac{b}{2}+\frac{1}{4}\geq\frac{b^2}{4}-\frac{b}{2}+\frac{1}{4}\Big|_{b=-3}=4$

$\Rightarrow w_1=-\frac{b-1}{2}\Big|_{b=-3}=2$

$w_2\geq 1-w_1-b=2$

$J(\vec{w}) = 2 < J(\vec{w}) = 4$

$\Rightarrow \vec{w}^* = [0 \ 2]^T \quad b = -1$

margin without offset from (a) $= \frac{1}{\frac{5}{2}} = \frac{2}{5}$

margin with offset $= \frac{1}{2}$

$\Rightarrow$ The classifier changes from $y = \text{sign}(-x_1 + 2x_2)$ to $y = \text{sign}(2x_2 - 1)$

The margin becomes larger from $\frac{2}{5}$ to $\frac{1}{2}$.

(from without offset to with offset)

3.

3.1(d). training data set feature dimensionality : (560, 1811)

test data set feature dimensionality : (70, 1811)

3.2(b). It might be beneficial to maintain class portions across folds because most machine learning algorithms assume that the training data set is a representative sample of the test data set. For example, the training data set and test data set are drawn from the same distribution. If the portion of positive examples from training data set differs significantly from the portion of positive examples from test data, the training data set cannot represent the test data set. The model built using this kind of training data will have large change to be overfitting and give large test error.

(d).

C	accuracy	F1-score	AUROC
10^{-3}	0.7089	0.8297	0.5000
10^{-2}	0.7107	0.8306	0.5031
10^{-1}	0.8060	0.8755	0.7188
10^{0}	0.8146	0.8749	0.7531
10^{1}	0.8182	0.8766	0.7592
10^{2}	0.8182	0.8766	0.7592
best C	10,100	10,100	10,100

The 5-fold CV performance increases as C increases.

The 5-fold CV performance based on F1-score is the best and the 5-fold CV performance based on AUROC is the worst corresponding to the same C.

The 5-fold CV performance based on F1-score has the smallest variance while the 5-fold CV performance based on AUROC has the largest variance.

3.3.(c). C=10 or 100

performance metric	performance on the test data
accuracy	0.7429
F1-score	0.4375
AUROC	0.6259