CSM146 PS3

1 a) Yes, instead of counting the number of unique words for each document and taking the cross product of the entire vocabulary, taking the size of the intersection of the sets of words in both documents is a significantly faster way to compute the same thing.

b) $k(x, z) = x \cdot z$

$k_1(x, z) = \frac{1}{\|x\| \, \|z\|} k(x, z)$ → scaling

$k_2(x, z) = k_1(x, z) + 1$ → sum

$k_3(x, z) = k_2(x, z) \, k_2(x, z) \, k_2(x, z)$ → product

$$\left(1 + \left(\frac{x}{\|x\|}\right) \cdot \left(\frac{z}{\|z\|}\right)\right)^3$$

c) $k_\beta(x, z) = (1 + \beta x \cdot z)^3$

$= (\beta x \cdot z)^3 + 3(\beta x \cdot z)^2 + 3(\beta x \cdot z) + 1$

$= \beta^3 \sum_{i,j,k=1}^{D} x_i z_i x_j z_j x_k z_k + 3\beta^2 \sum_{i,j=1}^{D} x_i z_i x_j z_j$

$+ 3\beta \sum_{i=1}^{D} x_i z_i + 1$

Typo

$$= \langle 1, \sqrt{3\beta}\, x_i \Big|_{i=1}^{D}, \sqrt{3}\,\beta\, x_i x_j \Big|_{i,j=1}^{D}, \beta^{3/2} x_i x_j x_k \Big|_{i,j,k=1}^{D} \rangle$$

$$\cdot$$

$$\langle 1, \sqrt{3\beta}\, z_i \Big|_{i=1}^{D}, \sqrt{3}\,\beta\, z_i z_j \Big|_{i,j=1}^{D},$$

$$\beta^{3/2} z_i z_j z_k \Big|_{i,j,k=1}^{D} \rangle$$

$$\therefore \phi_\beta(x) = \langle 1, \sqrt{3\beta}\, x_i \Big|_{i=1}^{D}, \sqrt{3}\,\beta\, x_i x_j \Big|_{i,j=1}^{D},$$

$$\beta^{3/2} x_i x_j x_k \Big|_{i,j,k=1}^{D} \rangle$$

In $K(x,z) = (1 + x \cdot z)^3$, $\beta = 1$

$$\therefore \phi(x) = \langle 1, \sqrt{3}\, x_i \Big|_{i=1}^{D}, \sqrt{3}\, x_i x_j \Big|_{i,j=1}^{D},$$

$$x_i x_j x_k \Big|_{i,j,k=1}^{D} \rangle$$

The parameter scales the features and makes the higher order terms have more weight.

For example if $\beta = 4$, $\sqrt{3\beta} < \beta^{3/2}$.

2 a) goal: minimize $\frac{1}{2}\|\theta\|^2$ s.t. $y_n\theta^T x_n \geq 1$

$$\frac{1}{2}\|\theta\|^2 = \frac{1}{2}(\theta_1^2+\theta_2^2).$$

constraint:

$$y\,\theta^T x = (-1)(\theta^T)(a,e)^T$$
$$= (-1)(\theta_1\ \theta_2)\begin{pmatrix} a \\ e \end{pmatrix}$$
$$= (-1)(a\theta_1 + e\theta_2) \geq 1$$
$$a\theta_1 + e\theta_2 \leq -1 \qquad \underline{1 + a\theta_1 + e\theta_2 \leq 0}$$

$$L(\theta,\alpha) = \frac{1}{2}(\theta_1^2+\theta_2^2) + \alpha(1 + a\theta_1 + e\theta_2)$$

$$g(\alpha) = \min_{\theta} L(\theta,\alpha)$$

$$\frac{\partial L(\theta,\alpha)}{\theta_1} = \frac{1}{2}(2)(\theta_1) + \alpha a = 0$$

$$\frac{\partial L(\theta,\alpha)}{\theta_2} = \frac{1}{2}(2)(\theta_2) + \alpha e = 0$$

$$\theta_1 = -\alpha a$$
$$\theta_2 = -\alpha e$$

$$g(\alpha) = \frac{1}{2}(\alpha^2 a^2 + \alpha^2 e^2) + (\alpha^2 a^2 + \alpha^2 e^2 + \alpha)$$
$$= \frac{3}{2}\alpha^2 a^2 + \frac{3}{2}\alpha^2 e^2 + \alpha$$

solution = max $g(\alpha)$

$$g'(\alpha) = 3\alpha a^2 + 3\alpha e^2 + 1 = 0$$

$$\alpha(3a^2+3e^2) = -1$$

$$\alpha = \frac{-1}{3a^2+3e^2}$$

$$\Theta_1 = \frac{a}{3a^2+3e^2}$$

$$\Theta_2 = \frac{e}{3a^2+3e^2}$$

$$\Theta^* = \begin{pmatrix} \frac{a}{3a^2+3e^2} \\ \frac{e}{3a^2+3e^2} \end{pmatrix}$$

b)

$$y_1\theta^T x_1 = \theta_1 + \theta_2 \geq 1 \qquad 1-(\theta_1+\theta_2) \leq 0$$

$$y_2\theta^T x_2 = -\theta_1 \geq 1 \qquad 1+\theta_1 \leq 0$$

$$L(\theta, \alpha)$$

$$= \frac{1}{2}(\theta_1^2+\theta_2^2) + \alpha_1(1-\theta_1-\theta_2) + \alpha_2(1+\theta_1)$$

$$g(\alpha) = \min L(\theta, \alpha)$$

$$\frac{\partial L(\theta,\alpha)}{\partial \theta_1} = \theta_1 + (-\alpha_1) + \alpha_2 = 0$$

$$\theta_1 = \alpha_1 - \alpha_2$$

$$\frac{\partial L(\theta,\alpha)}{\partial \theta_2} = \theta_2 + (-\alpha_1) = 0$$

$$\theta_2 = \alpha_1$$

$$g(\alpha) = \frac{1}{2}((\alpha_1-\alpha_2)^2 + \alpha_1^2) + \alpha_1(1-\alpha_1+\alpha_2-\alpha_1)$$
$$+ \alpha_2(1+\alpha_1-\alpha_2)$$
$$= \frac{1}{2}(\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 + \alpha_1^2)$$
$$+ \alpha_1(1-2\alpha_1+\alpha_2) + \alpha_2(1+\alpha_1-\alpha_2)$$
$$= \frac{1}{2}(2\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2) + \alpha_1 - 2\alpha_1^2 + \alpha_1\alpha_2$$
$$+ \alpha_2 + \alpha_1\alpha_2 - \alpha_2^2$$
$$= \cancel{\alpha_1^2} - \cancel{\alpha_1\alpha_2} + \frac{1}{2}\cancel{\alpha_2^2} + \alpha_1 - \cancel{2\alpha_1^2} + \cancel{\alpha_1\alpha_2} + \alpha_2 + \alpha_1\alpha_2 - \cancel{\alpha_2^2}$$

$$= -\alpha_1^2 - \frac{1}{2}\alpha_2^2 + \alpha_1 + \alpha_2 + \alpha_1\alpha_2$$

$$\frac{\partial(g(\alpha))}{\partial \alpha_1} = -2\alpha_1 + 1 + \alpha_2 = 0$$

$$\frac{\partial(g(\alpha))}{\partial \alpha_2} = -\alpha_2 + 1 + \alpha_1 = 0$$

$$2\alpha_1 - \alpha_2 = 1$$

$$\alpha_2 = 1 + \alpha_1$$

$$2\alpha_1 - 1 - \alpha_1 = 1$$

$$\alpha_1 = 2$$

$$\alpha_2 = 3$$

$$\therefore \theta_1 = -1$$

$$\theta_2 = 2$$

$$\theta^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\text{margin} = \frac{1}{\|\theta^*\|}$$

$$= \frac{1}{\sqrt{5}}$$

Typo

c)

with offset, the conditions will change

$\theta_1 + \theta_2 + b \geq 1$

$-(\theta_1 + b) \geq 1$

$1 - \theta_1 - \theta_2 - b \leq 0$

$1 + \theta_1 + b \leq 0$

$L(\theta, \alpha, b)$

$= \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha_1(1 - \theta_1 - \theta_2 - b) + \alpha_2(1 + \theta_1 + b)$

$\frac{\partial L}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0$

$\frac{\partial L}{\partial \theta_2} = \theta_2 - \alpha_1 = 0$

$g(\alpha) = \frac{1}{2}((\alpha_1 - \alpha_2)^2 + \alpha_1^2) + \alpha_1(1 - \alpha_1 + \alpha_2 - \alpha_1 - b)$

$+ \alpha_2(1 + \alpha_1 - \alpha_2 + b)$

$\frac{\partial g(\alpha)}{\alpha_1} = \alpha_1 + \alpha_1 + 1 - 2\alpha_1 - 2\alpha_1 - b$

$+ \alpha_2 = 0$

$2\alpha_1 = 1 - b + \alpha_2 \qquad \alpha_1 = \frac{1 - b + \alpha_2}{2}$

$\frac{\partial g(\alpha)}{\alpha_2} = -\alpha_2 + \alpha_1 + 1 + b = 0$

$\alpha_2 = \alpha_1 + b + 1 \qquad \alpha_1 = 2$

$\alpha_2 = \frac{3}{2} + \frac{b}{2} + \frac{\alpha_2}{2} + b + 1 \qquad \alpha_1 = \frac{1 - b + \alpha_1 + b + 1}{2} \qquad \cancel{\alpha_2}$

$\alpha_2 = 3b + 5$

$3b+5 = 3+b$

$2b = -2$

$\underline{b = -1}$

$\underline{d_2 = 2 - 1 + 1 = 2}$

$\theta_1 = 0$

$\theta_2 = 2$

$\theta^* = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \quad b = -1$

$\text{margin} = \dfrac{1}{2}$

With offset, the margin is larger -

It is $\frac{1}{2}$ as compared to $\frac{1}{\sqrt{3}}$

3.2b) It might be beneficial to maintain class proportions across folds as it prevents extreme cases where the fold is almost all positive (since the proportion is not equal in the training data) which may result in skewed models as a hyperplane separating the positive and negative data cannot be correctly formed.

3.2d)

C	accuracy	F1-score	AUROC	precision	specificity	sensitivity
10^-3	0.7089	0.8297	0.5	0.7089	0.0	1.0
10^-2	0.7107	0.8306	0.5031	0.7102	0.0063	1.0
10^-1	0.8060	0.8755	0.7188	0.8357	0.5081	0.9294
10^0	0.8146	0.8749	0.7531	0.8562	0.6045	0.9017
10^1	0.8182	0.8766	0.7592	0.8595	0.6167	0.9017
10^2	0.8182	0.8766	0.7592	0.8595	0.6167	0.9017
Best C	10^1/10^2	10^1/10^2	10^1/10^2	10^1/10^2	10^1/10^2	10^-3/10^-2

For all the metrics except for sensitivity, performance increases with C. For sensitivity performance decreases with C.

3.3 a) The hyper-parameter defines how far the influence of a single training example reaches, with low values meaning far and high close. Lower gamma means lower generalization error.

b) For gamma values larger than 1, the score for each C value was the same. So I used a grid with C values from 10^-3 to 10^2 as used in the linear kernel measurement and gamma values from 1 to 10^-3.

c)

Metric	Score	C	gamma
accuracy	0.8165	100	0.01
F1-score	0.8763	100	0.01
AUROC	0.7545	100	0.01
precision	0.8583	100	0.01
specificity	0.6047	100	0.01
sensitivity	1.0	0.001	1

3.4a) For the linear kernel I chose a value of 100 for C as it has the highest scores for all performance metrics except for sensitivity. For the RBF kernel I chose a value of 100 for C and 0.01 for gamma as it has the highest scores for all performance metrics

```
c)
Classifier: Linear
Metric:  accuracy Score:  0.742857142857
Metric:  f1-score Score:  0.4375
Metric:  auroc Score:  0.625850340136
Metric:  precision Score:  0.636363636364
Metric:  specificity Score:  0.918367346939
Metric:  sensitivity Score:  0.333333333333
Classifier: RBF
Metric:  accuracy Score:  0.757142857143
Metric:  f1-score Score:  0.451612903226
Metric:  auroc Score:  0.636054421769
Metric:  precision Score:  0.7
Metric:  specificity Score:  0.938775510204
Metric:  sensitivity Score:  0.333333333333
```

The RBF-Kernel SVM classifier performs better across all metrics than the Linear SVM classifier.