CM146 HW3

Stephanie Chen

November 2019

1. - Expand the cubic.

$$K_B(x,z) = (1+\beta x^T z)^3 = 1 + 3\beta x^T z + 3(\beta x^T z)^2 + (\beta x^T z)^3$$

Note:

$$x^T z = x_1 x_2 + z_1 z_2$$

- Corresponding feature map $\phi_\beta(x)$ and $\phi_\beta(z) are$:

$$\phi_\beta(x) = [1 \quad \sqrt{3\beta} x_1 x_2 \quad \sqrt{3}\beta(x_1 x_2)^2 \quad \beta^{\frac{3}{2}}(x_1 x_2)^3]$$

$$\phi_\beta(z) = [1 \quad \sqrt{3\beta} z_1 z_2 \quad \sqrt{3}\beta(z_1 z_2)^2 \quad \beta^{\frac{3}{2}}(z_1 z_2)^3]$$

β is a scaling factor that can be used to increase/decrease distance between samples. When $\beta = 1$, $K_B(x,z) = K(x,z) = (1+x^T z)^3$. When $0 < \beta < 1$, less weights are put onto higher degree elements. In contrast, when $\beta > 1$, more weights are put onto higher degree elements compare to lower degree elements. When β approaches infinity or zero, all terms except the constant term approaches infinity or zero correspondingly, so only the constant term really matters in shaping the feature map. β is basically a parameter that allows more flexibility for the feature map of this kernel.

2. a) Given the two training examples, we need to satisfy the constraints:

$$y_1([w1 \quad w2][1 \quad 1]^T) \geq 1, y1 = 1$$

, which means w1 + w2 $\geq$ 1 and

$$y_2([w1 \quad w2][1 \quad 0]^T) \geq 1, y2 = -1$$

, which means w1 $\geq$ -1. We want minimum w that satisfies these 2 contraints, which means we can set $w_1 = -1$ and $w_2 = 2$, then

$$w* = [-1 \quad 2]^T \frac{1}{\sqrt{(-1)^2 + (2)^2}} = \frac{1}{\sqrt{5}}[-1 \quad 2]^T$$

and the margin is $\frac{1}{\sqrt{5}}$

b) For non-zero offset parameter b, the new constraints are:

$$y_1([w1 \quad w2 \quad b][1 \quad 1 \quad 1]^T) \geq 1, y1 = 1$$

, which means w1 + w2 + b $\geq$ 1 and

$$y_2([w1 \quad w2 \quad b][1 \quad 0 \quad 1]^T) \geq 1, y2 = -1$$

, which means b $\leq$ -1 - w1. We can plug the second constraint to the first constraint to get

$$w1 + w2 + b = w1 + w2 + (-1 - w1) = w2 - 1 \geq 1, w2 \geq 2$$

We want to minimize w, so we can set w1 = 0, so b $\leq$ -1 and w2 = 2 and get

$$w* = [0 \quad 2]^T \frac{1}{\sqrt{2^2}} = [0 \quad 1]^T$$

We use b* = -1 to satisfy the original inequality w1 + w2 + b $\geq$ 1 with w1 = 0, w2 = 2. With w1* = 0, w2* = 2, b* = -1, margin = $\frac{1}{2}$

3. 3.1 d) Dimension of the feature matrix is (630, 1811).

3.2 b) By maintaining class proportions across fold, we can ensure cross validation score we compute in the end, which is the average of all 5 folds, is more representative of the actual data. If for instance, a training set for a fold consists of all positive or all negative labels only, the test error will be more inaccurate.

3.2 d) The best C for all three metrics is C = 10 and C = 100 for all three metrics. The score for C = 10 and C = 100 is not detectable for up to 14 decimal places. CV performance increases as C increases for all 3 metrics, with f1-score metric getting the best score for the given training data.

C	accuracy	f1_score	auroc
0.001	0.708941954	0.8296828227	0.5
0.01	0.7107437558	0.8305628005	0.503125
0.1	0.8060326762	0.875472683	0.7187871596
1	0.8146271113	0.8748648327	0.7531113349
10	0.8181827371	0.8765621529	0.7591719409
100	0.8181827371	0.8765621529	0.7591719409
best C	10	10	10

3.3 a) Trained a linear kernel SVM with C = 10.

3.3 c) report performance of each metric The performance of the linear kernel SVM with C = 10 on test data is:

accuracy	0.7428571429
f1_score	0.4375
auroc	0.6258503401