CS260, Winter 2017
Problem Set 5: Bias-Variance and SVM
Due 3/6/2017

Ray Zhang

March 4, 2017

1 Q1

(a) Problem 1a

Solution: Solution to problem 1a

We are trying to solve for $\hat{\beta_\lambda}$, and to solve it we use the normal equations:

$L(\beta, \lambda) = \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$ for some λ.

Differentiating, we get:

$\nabla_\beta L(\beta, \lambda) = \frac{-2X^Ty + 2X^TX\beta}{n} + 2\lambda\beta = 0.$

Separating for β, we get:

$\hat{\beta_\lambda} = (X^TX + n\lambda I)^{-1}X^Ty$

Since $y_i \sim N(x_i^T\beta^*, \sigma^2)$, we can say that:

$y \sim N(X\beta^*, \Sigma)$, where $\Sigma = diag(\sigma^2)$.

We know that by affine transformations of y, we get $\hat{\beta_\lambda}$. Thus retracing affine transformation steps, we get that:

$\hat{\beta_\lambda} \sim N((X^TX+n\lambda I)^{-1}X^TX\beta^*, (X^TX+n\lambda I)^{-1}X^T\Sigma X(X^TX+n\lambda I)^{-1})$

Since it follows that $(X^TX + n\lambda I)^{-1}$ is symmetric.

(b) Problem 1b

Solution: Solution to problem 1b

To compute the bias, we calculate the expected value of:

$E[x^T\hat{\beta_\lambda}] - x^T\beta^* = x^TE[\hat{\beta_\lambda}] - x^T\beta^* = x^T(E[\hat{\beta_\lambda}] - \beta^*)$

$E[\hat{\beta}_\lambda] = (X^T X + n\lambda I)^{-1} X^T X \beta^*$ as found from part a.

Substituting, we get:

$E[\hat{\beta}_\lambda] - \beta^* = (X^T X + n\lambda I)^{-1} X^T X \beta^* - \beta^*$

$= ((X^T X + n\lambda I)^{-1} X^T X - I)\beta^*.$

Then, plugging in back to our original equation we get:

$x^T((X^T X + n\lambda I)^{-1} X^T X - I)\beta^* = bias$

(c) Problem 1c

Solution: Solution to problem 1c

To solve for variance, we can first see that:

$var((x^T \hat{\beta}_\lambda) = E[x^T \hat{\beta}_\lambda - E[x^T \hat{\beta}_\lambda])^2]$

$= E[(x^T \hat{\beta}_\lambda - x^T E[\hat{\beta}_\lambda])^2]$

$= E[x^T(\hat{\beta}_\lambda - E[\hat{\beta}_\lambda])^2 x]$

$= x^T E[(\hat{\beta}_\lambda - E[\hat{\beta}_\lambda])^2] x$

$= x^T var(\hat{\beta}_\lambda) x$

$= x^T (X^T X + n\lambda I)^{-1} X^T \Sigma X (X^T X + n\lambda I)^{-1}) x$

$variance = \sigma^2 (X(X^T X + n\lambda I)^{-1} x)^T (X(X^T X + n\lambda I)^{-1} x)$

Note: This is a positive semidefinite matrix.

(d) Problem 1d

Solution: Solution to problem 1d

$x^T((X^T X + n\lambda I)^{-1} X^T X - I)\beta^* = bias$

$\sigma^2 (X(X^T X + n\lambda I)^{-1} x)^T (X(X^T X + n\lambda I)^{-1} x) = variance$

If we plug in $\lambda = 0$, then we get that:

$x^T((X^T X)^{-1} X^T X - I)\beta^* = x^T(I - I)\beta^* = 0 = bias$

$\sigma^2 (X(X^T X)^{-1} x)^T (X(X^T X)^{-1} x) =$

$\sigma^2 x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x = \sigma^2 x^T (X^T X)^{-1} x = variance$

If we plug in $\lambda = \frac{k}{n}$, where k is very large, then we get that:

$x^T((kI)^{-1} X^T X - I)\beta^* \approx -x^T \beta^* = bias$

$\sigma^2 x^T (kI)^{-1} X^T X (kI)^{-1} x = \frac{\sigma^2}{k^2} x^T X^T X x \approx 0 = variance$

Thus, from our conclusions, when λ is small, we have that bias is small, and variance is large. Conversely, when λ is large, we have that bias is large, and variance is small.

2 Q2

(a) Problem 2a

Solution: Solution to problem 2a

$K_3 = aK_1 + bK_2$ is our kernel, where $K_1, K_2 \in S^n_{++}$

For any x, $x^T K_3 x = x^T (aK_1 + bK_2) x = ax^T K_1 x + bx^T K_2 x$.

Because K_1, K_2 are positive semi-definite, we can see that we are adding 2 non-negative terms multiplied by non-negative constants, so we have a non-negative result. Thus, K_3 is a kernel.

(b) Problem 2b

Solution: Solution to problem 2b

$k_4(x, x') = f(x)f(x')$ is our kernel.

For our kernel, $K_4 = f(X)f(X)^T$ where $X \in \Re^{nxd}$, n is the samples and d is dimensions of each sample. $f(X)$ is thus just a vector in $\Re^n$.

For any x, then, we have that $x^T K_4 x = x^T f(X) f(X)^T x = (f(X)^T x)^T (f(X)^T x)$. Thus, because $f(X)^T x$ is a scalar, we have that the above expression is equal to $(f(X)^T x)^2 \geq 0$.

(c) Problem 2c

Solution: Solution to problem 2c

$K_5 = K_1 \circ K_2$, which is a hadamard product of two PSD matrices.

We actually saw this in the first homework, but I will prove it one more time:

Denote K_5 as H, K_1 as X, K_2 as Y:

$H = X \circ Y$, and $H_{ij} = X_{ij} Y_{ij}$

We know that $X = \sum_i \alpha_i v_i v_i^T$, $Y = \sum_i \beta_i u_i u_i^T$ via eigendecomposition.

We can see that $H = X \circ Y = \sum_i \alpha_i v_i v_i^T \circ \sum_i \beta_i u_i u_i^T$.

Using foil, we get that:

$H = \sum_{ij} \alpha_i v_i v_i^T \circ \beta_j u_j u_j^T$

$= \sum_{ij} \alpha_i \beta_j v_i v_i^T \circ u_j u_j^T$

The resulting matrix of any i, j foil of $vv^T \circ uu^T$ would be of the format:

$X_{ij} = v_i v_j u_i u_j = v_i u_j u_i v_j = vu^T uv^T = (uv^T)^T (uv^T)$

This gives us an outer product resulting in:

$= \sum_{ij} \alpha_i \beta_j v_i u_j^T \circ u_j v_i^T$

$= \sum_{ij} \alpha_i \beta_j (u_j v_i^T)^T (v_j u_j^T)$

This is PSD because the eigenvalues are $\alpha_i, \beta_i \geq 0$, and any outer product combination gives a PSD matrix. Thus, H, or K_5 is a kernel.

3 Q3

(a) Problem 3a

Solution: Solution to problem 3a

If we change our slack variable sum from $\sum_{i=1}^n \xi_i$ to $\sum_{i=1}^n \xi_i^p$, we would get the following quadratic program:

$min_{w,b,\xi} ||w||^2 + C \sum_{i=1}^n \xi_i^p$

subject to $\xi \geq 0$ and $y_i(w^T x_i + b) \geq 1 - \xi_i$

Putting this into the generalized lagrangian form:

$L(w, b, \xi, \lambda, \alpha) = ||w||^2 + C \sum_{i=1}^n \xi_i^p - \lambda^T \xi + \alpha^T \{1 - y(Xw + b) - \xi\}$

Solving for the dual, we first find the infimum with respect to the primal variables, and then solve for the dual variables, which is an affine map:

$sup_{\lambda,\alpha} g(\lambda, \alpha)$

$= sup_{\lambda,\alpha} inf_{w,b,\xi} ||w||^2 + C\sum_{i=1}^{n} \xi_i^p - \lambda^T \xi + \alpha^T\{1 - y(Xw + b) - \xi\}$

Our dual formulation thus requires the infimum first, so we solve for it:

$\nabla_w L = w - \alpha \circ X^T y = 0$. Thus $w = \alpha \circ X^T y$.

$\nabla_b L = -\alpha^T y = 0$.

$\nabla_\xi L = Cp\xi^{p-1} - \lambda - \alpha = 0$ Thus, $\xi_i = (\frac{\alpha_i + \lambda_i}{Cp})^{\frac{1}{p-1}}$

Plugging in our w values, we see that our infimum is:

$\frac{1}{2}(\sum_i \alpha_i y_i x_i)^T(\sum_i \alpha_i y_i x_i) +$

$\sum_i \alpha_i + (\sum_i \alpha_i y_i)b -$

$\sum_i \alpha_i y_i (\sum_j \alpha_j y_j x_j)^T x_i +$

$\sum_i (c\xi_i^{p-1} - \lambda_i - \alpha_i)\xi_i$

Plugging the term in for ξ, we get:

$\sum_i (\frac{\alpha_i + \lambda_i}{p} - \lambda_i - \alpha_i)(\frac{\alpha_i + \lambda_i}{Cp})^{\frac{1}{p-1}}$

$= \sum_i (\alpha_i + \lambda_i)(\frac{1}{p} - 1)(\frac{\alpha_i + \lambda_i}{Cp})^{\frac{1}{p-1}}$

After simplifying:

$g(\lambda, \alpha) = -\frac{1}{2}\sum_{i,j} \alpha_i y_i \alpha_j y_j x_i^T x_j + \sum_i \alpha_i + \sum_i (\alpha_i + \lambda_i)(\frac{1}{p} - 1)(\frac{\alpha_i + \lambda_i}{Cp})^{\frac{1}{p-1}}$

subject to $\alpha \geq 0$, $\alpha^T y = 0$

The function above is our dual formulation.

(b) Problem 3b

Solution: Solution to problem 3b

There are various issues with this dual formulation.

According to Slater's conditions, we must have that

$||w||^2 + C\sum_{i=1}^{n} \xi_i^p$ must be convex, which may no longer be the case (namely, the polynomial ξ sum is not convex.)

In addition, the term that we introduced is much more complex than the original linear ξ formulation for dual SVM. Optimizing this function will naturally take larger overheads.

Thus, in my opinion the general expression is a lot more complicated than the original expression.

4 Q4

(a) Problem 4.1

Solution: Solution to problem 4.1

The third feature has mean: 2.540000e+00 and std: 1.074430e+00. The tenth feature has mean 2.527000e+00 and std: 1.123953e+00.

We don't use the test data's information in our training whatsoever. This is because if we use test data's statistical variables, we are skewing our data towards the test data and can affect our test time predictions in a biased way. We want to assume that the test data was not introduced until test time, and thus we don't use the test data's statistics to normalize.

(b) Problem 4.2

Solution: Solution to problem 4.2

The solution is in trainsvm.m and testsvm.m.

(c) Problem 4.3

Solution: Solution to problem 4.3

a) The values are printed as below:

$C = 4^{-6}$ average time : 0.264872 and average accuracy : 0.545000
$C = 4^{-5}$ average time : 0.211933 and average accuracy : 0.776000
$C = 4^{-4}$ average time : 0.210790 and average accuracy : 0.804000
$C = 4^{-3}$ average time : 0.243938 and average accuracy : 0.806000
$C = 4^{-2}$ average time : 0.306947 and average accuracy : 0.792000
$C = 4^{-1}$ average time : 0.376549 and average accuracy : 0.792000
$C = 4^{0}$ average time : 0.468904 and average accuracy : 0.790000
$C = 4^{1}$ average time : 0.331488 and average accuracy : 0.788000
$C = 4^{2}$ average time : 0.408153 and average accuracy : 0.787000

The value of C is linearly correlated with time : $C \propto time$. It seems for higher C's, it is harder to optimize the function in a short amount of time. It makes sense because if C is bigger, we have a bigger range

of ξ's to consider. With a smaller range of ξ's, the answer does not deviate far from the zero vector.

The accuracy peaks when C is around 4^{-4} to 4^{-2}. It is likely because for this dataset, penalizing the slack variables too much(When C is large) will make the margin too large, and thus introduce too much uncertainty to new points. However, when we have a small C, the margin is too small, and we no longer value the accuracy of our predictions as much. If this data was linearly separable we would have no problem, but it isnt.

b) I chose the accuracy $C = 4^{-3}$, with average accuracy : 8.060000e-01 on our cross validation, which is the highest point we have.

c) Accuracy of test : 0.846897.

(d) Problem 4.4

Solution: Solution to problem 4.4

a) The LibSVM's performance is shown below:

$C = 4^{-6}$ average accuracy : 51.700000 and time : 0.405893

$C = 4^{-5}$ average accuracy : 78.800000 and time : 0.363642

$C = 4^{-4}$ average accuracy : 80.500000 and time : 0.275093

$C = 4^{-3}$ average accuracy : 79.800000 and time : 0.241107

$C = 4^{-2}$ average accuracy : 79.600000 and time : 0.231782

$C = 4^{-1}$ average accuracy : 79.400000 and time : 0.359879

$C = 4^{0}$ average accuracy : 79.400000 and time : 0.850305

$C = 4^{1}$ average accuracy : 79.300000 and time : 2.226188

$C = 4^{2}$ average accuracy : 79.200000 and time : 9.807180

I believe the accuracies are very similar, and is only offset slightly because the cross validation split was chosen differently. Solving the dual and solving the primal makes no difference because this formulation of SVM has strong duality and thus allows for us to solve either equation.

b) The time factor is very different though. Instead of a minor increase in time in quadprog, LibSVM's time for $C = 4^2$ is a whopping 10 seconds of operation which is very slow.

(e) Problem 4.5

Solution: Solution to problem 4.5

a) For polynomial kernels, the result is shown below:

$C = 4^{-4}$, deg $= 1$, average accuracy : 51.700000 and time : 0.450927

$C = 4^{-4}$, deg $= 2$, average accuracy : 51.700000 and time : 0.450927

$C = 4^{-4}$, deg $= 3$, average accuracy : 51.700000 and time : 0.450927

$C = 4^{-3}$, deg $= 1$, average accuracy : 51.700000 and time : 0.597869

$C = 4^{-3}$, deg $= 2$, average accuracy : 51.700000 and time : 0.597869

$C = 4^{-3}$, deg $= 3$, average accuracy : 51.700000 and time : 0.597869

$C = 4^{-2}$, deg $= 1$, average accuracy : 78.700000 and time : 0.399379

$C = 4^{-2}$, deg $= 2$, average accuracy : 51.700000 and time : 0.399379

$C = 4^{-2}$, deg $= 3$, average accuracy : 51.700000 and time : 0.399379

$C = 4^{-1}$, deg $= 1$, average accuracy : 80.400000 and time : 0.319223

$C = 4^{-1}$, deg $= 2$, average accuracy : 65.400000 and time : 0.319223

$C = 4^{-1}$, deg $= 3$, average accuracy : 62.300000 and time : 0.319223

$C = 4^{0}$, deg $= 1$, average accuracy : 79.500000 and time : 0.286002

$C = 4^{0}$, deg $= 2$, average accuracy : 72.800000 and time : 0.286002

$C = 4^{0}$, deg $= 3$, average accuracy : 79.200000 and time : 0.286002

$C = 4^{1}$, deg $= 1$, average accuracy : 79.300000 and time : 0.270160

$C = 4^{1}$, deg $= 2$, average accuracy : 70.900000 and time : 0.270160

$C = 4^{1}$, deg $= 3$, average accuracy : 79.000000 and time : 0.270160

$C = 4^{2}$, deg $= 1$, average accuracy : 79.500000 and time : 0.425980

$C = 4^{2}$, deg $= 2$, average accuracy : 70.900000 and time : 0.425980

$C = 4^{2}$, deg $= 3$, average accuracy : 79.200000 and time : 0.425980

$C = 4^{3}$, deg $= 1$, average accuracy : 79.400000 and time : 0.863529

$C = 4^{3}$, deg $= 2$, average accuracy : 70.900000 and time : 0.863529

$C = 4^{3}$, deg $= 3$, average accuracy : 79.200000 and time : 0.863529

$C = 4^4$, deg = 1, average accuracy : 79.300000 and time : 2.750375

$C = 4^4$, deg = 2, average accuracy : 70.900000 and time : 2.750375

$C = 4^4$, deg = 3, average accuracy : 79.200000 and time : 2.750375

$C = 4^5$, deg = 1, average accuracy : 79.200000 and time : 10.955213

$C = 4^5$, deg = 2, average accuracy : 70.900000 and time : 10.955213

$C = 4^5$, deg = 3, average accuracy : 79.200000 and time : 10.955213

$C = 4^6$, deg = 1, average accuracy : 79.300000 and time : 42.942284

$C = 4^6$, deg = 2, average accuracy : 70.900000 and time : 42.942284

$C = 4^6$, deg = 3, average accuracy : 79.200000 and time : 42.942284

$C = 4^7$, deg = 1, average accuracy : 78.600000 and time : 69.890624

$C = 4^7$, deg = 2, average accuracy : 70.900000 and time : 69.890624

$C = 4^7$, deg = 3, average accuracy : 79.200000 and time : 69.890624

The test accuracy of the polynomial kernel with degree 1 with C = 0.25, which gave the highest accuracy in cross validation stage (80%), has the accuracy:

Highest accuracy is: 0.854253 on test data

b) The accuracies of RBF kernel is shown below:

$C = 4^{-4}$, $gamma = 4^{-7}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-6}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-5}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-4}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-3}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-2}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{-1}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{0}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{1}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-4}$, $gamma = 4^{2}$, average accuracy : 51.700000 and time : 0.487770

$C = 4^{-3}$, $gamma = 4^{-7}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-6}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-5}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-4}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-3}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-2}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{-1}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{0}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{1}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-3}$, $gamma = 4^{2}$, average accuracy : 51.700000 and time : 0.390937

$C = 4^{-2}$, $gamma = 4^{-7}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-6}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-5}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-4}$, average accuracy : 54.500000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-3}$, average accuracy : 61.200000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-2}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{-1}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{0}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{1}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-2}$, $gamma = 4^{2}$, average accuracy : 51.700000 and time : 0.406940

$C = 4^{-1}$, $gamma = 4^{-7}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-6}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-5}$, average accuracy : 73.200000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-4}$, average accuracy : 80.300000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-3}$, average accuracy : 83.600000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-2}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{-1}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{0}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{1}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{-1}$, $gamma = 4^{2}$, average accuracy : 51.700000 and time : 0.387069

$C = 4^{0}$, $gamma = 4^{-7}$, average accuracy : 51.700000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-6}$, average accuracy : 76.200000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-5}$, average accuracy : 79.900000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-4}$, average accuracy : 81.700000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-3}$, average accuracy : 87.200000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-2}$, average accuracy : 76.600000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{-1}$, average accuracy : 57.300000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{0}$, average accuracy : 55.800000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{1}$, average accuracy : 55.200000 and time : 0.384719

$C = 4^{0}$, $gamma = 4^{2}$, average accuracy : 55.200000 and time : 0.384719

$C = 4^{1}$, $gamma = 4^{-7}$, average accuracy : 77.000000 and time : 0.512324

$C = 4^{1}$, $gamma = 4^{-6}$, average accuracy : 79.600000 and time : 0.512324

$C = 4^{1}$, $gamma = 4^{-5}$, average accuracy : 80.000000 and time : 0.512324

$C = 4^{1}$, $gamma = 4^{-4}$, average accuracy : 84.700000 and time : 0.512324

$C = 4^1$, $gamma = 4^{-3}$, average accuracy : 88.600000 and time : 0.512324

$C = 4^1$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.512324

$C = 4^1$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.512324

$C = 4^1$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.512324

$C = 4^1$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.512324

$C = 4^1$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.512324

$C = 4^2$, $gamma = 4^{-7}$, average accuracy : 79.500000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-6}$, average accuracy : 79.700000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-5}$, average accuracy : 81.100000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-4}$, average accuracy : 88.000000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.322966

$C = 4^2$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.322966

$C = 4^2$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.322966

$C = 4^2$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.322966

$C = 4^2$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.322966

$C = 4^3$, $gamma = 4^{-7}$, average accuracy : 79.800000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-6}$, average accuracy : 79.600000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-5}$, average accuracy : 85.700000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-4}$, average accuracy : 87.200000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.259741

$C = 4^3$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.259741

$C = 4^3$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.259741

$C = 4^3$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.259741

$C = 4^3$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.259741

$C = 4^4$, $gamma = 4^{-7}$, average accuracy : 79.000000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-6}$, average accuracy : 81.700000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-5}$, average accuracy : 88.100000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-4}$, average accuracy : 87.200000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.243093

$C = 4^4$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.243093

$C = 4^4$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.243093

$C = 4^4$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.243093

$C = 4^4$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.243093

$C = 4^5$, $gamma = 4^{-7}$, average accuracy : 79.400000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-6}$, average accuracy : 85.400000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-5}$, average accuracy : 87.400000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-4}$, average accuracy : 87.200000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.276290

$C = 4^5$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.276290

$C = 4^5$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.276290

$C = 4^5$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.276290

$C = 4^5$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.276290

$C = 4^6$, $gamma = 4^{-7}$, average accuracy : 81.800000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-6}$, average accuracy : 87.800000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-5}$, average accuracy : 87.400000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-4}$, average accuracy : 87.200000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 0.422648

$C = 4^6$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 0.422648

$C = 4^6$, $gamma = 4^0$, average accuracy : 56.200000 and time : 0.422648

$C = 4^6$, $gamma = 4^1$, average accuracy : 55.200000 and time : 0.422648

$C = 4^6$, $gamma = 4^2$, average accuracy : 55.200000 and time : 0.422648

$C = 4^7$, $gamma = 4^{-7}$, average accuracy : 85.300000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-6}$, average accuracy : 87.400000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-5}$, average accuracy : 87.400000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-4}$, average accuracy : 87.200000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-3}$, average accuracy : 88.200000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-2}$, average accuracy : 77.900000 and time : 1.442553

$C = 4^7$, $gamma = 4^{-1}$, average accuracy : 57.900000 and time : 1.442553

$C = 4^7$, $gamma = 4^0$, average accuracy : 56.200000 and time : 1.442553

$C = 4^7$, $gamma = 4^1$, average accuracy : 55.200000 and time : 1.442553

$C = 4^7$, $gamma = 4^2$, average accuracy : 55.200000 and time : 1.442553

The best performing RBF kernel is with $C = 4^1$, $gamma = 4^{-3}$ with an average accuracy of 88.600000.

Running it on the test gives us:

Highest accuracy is: 0.904368 on test data

In conclusion, RBF kernel is the best, with the hyperparameters of $C = 4, gamma = 4^{-3}$, and it gives us around 90% accuracy on our test data.