# Homework - III

Siddharth Chandrasekaran

November 4, 2017

**Answer 1.a.** In order to find the optimal w*, we need to minimize the loss function L.

$$L = \sum_{i=1}^{N}(y_i - w^T x_i)^2 + \lambda(||w||)^2$$

Now to find the value of w* that minimizes L, we set the gradient $\frac{dL}{dw} = 0$

$$\frac{dL}{dw} = 0 = \sum_{i=1}^{N} 2x_i(x_i^T w - y_i) + 2\lambda w$$
$$2\lambda w = \sum_{i=1}^{N} 2x_i(y_i - x_i^T w)$$
$$\lambda w = \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i x_i^T w$$
$$\sum_{i=1}^{N}(x_i x_i^T + \lambda I)w = \sum_{i=1}^{N} x_i y_i$$
$$w = (\sum_{i=1}^{N} x_i x_i^T + \lambda I)^{-1} \sum_{i=1}^{N} x_i y_i = w* \text{ (by definition)}$$

Hence proved.

**Answer 1.b.** Solution in the case of basis expansion x to $\varphi(x)$ being used will be,

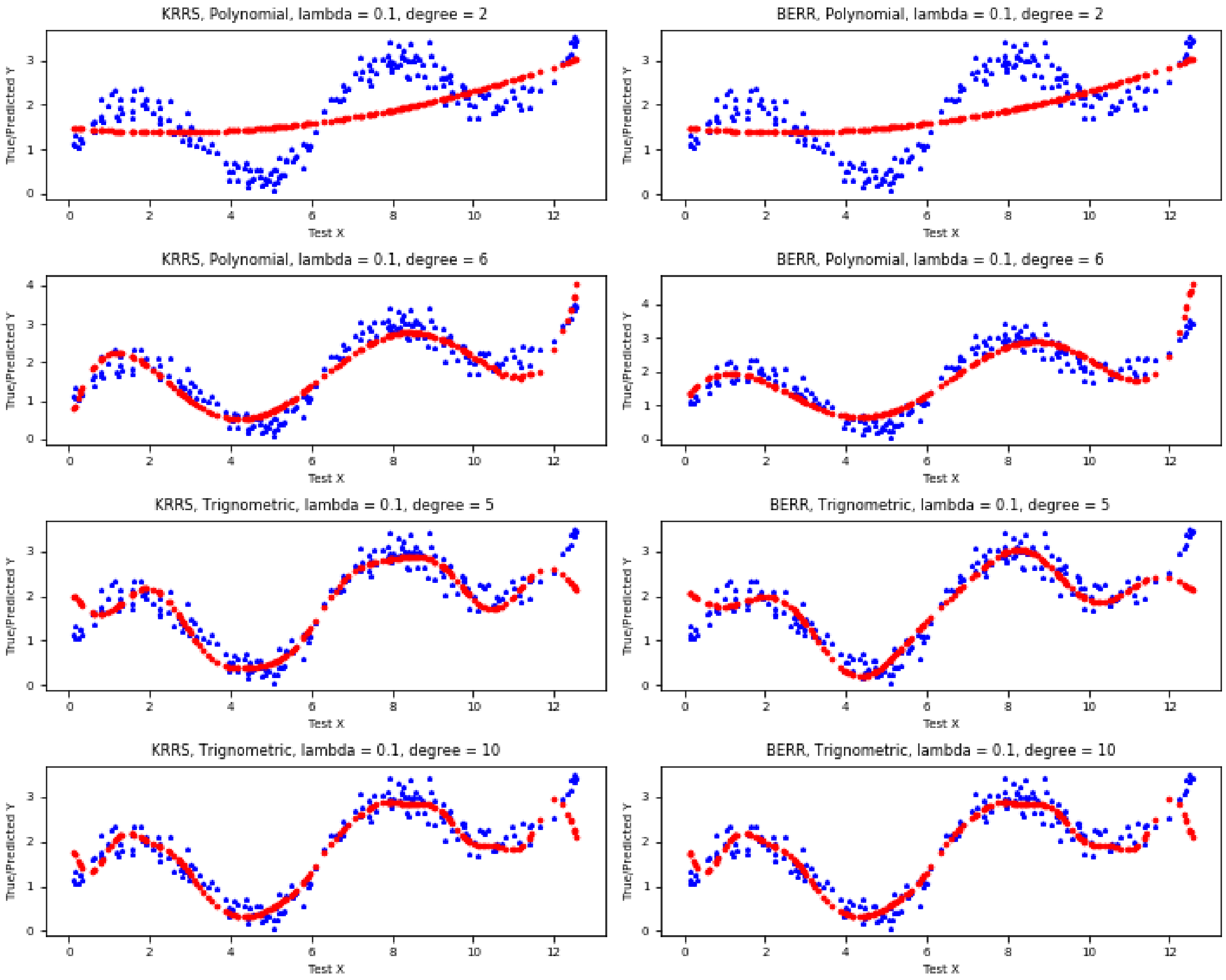
$$w* = (\sum_{i=1}^{N} \varphi(x_i)\varphi(x_i)^T + \lambda I)^{-1} \sum_{i=1}^{N} \varphi(x_i) y_i$$

**Answer 1.c.** Let w* be the optimal beta upon learning, X denote training data, $\alpha = (K+\lambda I)^-1Y$ and $K_{ij} = \varphi(x_i)^T\varphi(x_j)$
For any new point $x_{new}$, we have

$$y_{new} = (w*)^T \varphi(x_{new})$$
$$y_{new} = (\varphi(X)^T \alpha)^T \varphi(x_{new})$$
$$y_{new} = \sum_{i=1}^{N} \alpha_i \varphi(x_i)^T \varphi(x_{new})$$

Thus we can prove that, given a new point $x_{new}$ the expression for $y_{new}$ depends only on the inner products of the samples $x_i$.

**Answer 1.d.1.** The plot for the KRRS and BERR is as follows



Inferences from the graph:

1. BERR and KRRS's predictions are almost the same. (There are very minute differences between the predictions.)
2. Polynomial kernel with degree 6 seems to be fitting the data best.

**Answer 1.d.2.** For the synthetic dataset, KRRS and BERR were implemented

| | Kernel $k(x_1, x_2)$ | Basis Expansion $\varphi(x)$ |
|---|---|---|
| Polynomial Degree 1 | 0.582340444821 | 0.582447550275 |
| Polynomial Degree 2 | 0.536822781304 | 0.536850611277 |
| Polynomial Degree 4 | 0.441506852453 | 0.441917855053 |
| Polynomial Degree 6 | 0.102524426496 | 0.126113336564 |
| Trignometric Degree 3 | 0.165771469732 | 0.166136360724 |
| Trignometric Degree 5 | 0.118826043104 | 0.130578295205 |
| Trignometric Degree 10 | 0.0973803176496 | 0.097819516656 |

Table 1 : Models and out of sample error (MSE) for the synthetic dataset

**Answer 1.e.** Kernel ridge regression was run on the credit card dataset.

|  | RBF | Poly with Degree 3 | Linear |
|---|---|---|---|
| $\alpha = 1, \gamma = def$ | 15.8570879506 | 3341966.24644 | 10.7713543585 |
| $\alpha = 1, \gamma = 1$ | 82.8101122643 | 47046670.5357 | – |
| $\alpha = 1, \gamma = 0.001$ | 9.47763118023 | 6.44547708007 | – |
| $\alpha = 0.0001, \gamma = def$ | 2.49400839079 | 5894221809.63 | 10.4017332213 |
| $\alpha = 0.0001, \gamma = 1$ | 23.244262647 | 47046670.5357 | – |
| $\alpha = 0.0001, \gamma = 0.001$ | 2.42246477144 | 2.02149366546 | – |

Table 2 : Models and out of sample error (MSE) for the credit card dataset

The model with least estimated out of sample error was for the following parameters $\alpha = 0.0001, \gamma = 0.001$, polynomial Kernel for degree 3.
Best model selection was done using grid search using cross validation with train test split of 4-1 (5 fold cross validation).
**The kaggle test error for this model was 1.64554 and the out of sample error was 2.02149366546.**
Kaggle Display Name - Siddharth Chandrasekaran
Kaggle test error was lower in this case than the estimated out of sample error which implies that the test data follows this model (generalization) better. Small positive values of alpha improve the conditioning of the problem and reduce the variance of the estimates and conforming to this, the model with the lowest alpha works the best. Even though RBF kernels give better flexibility than the polynomial kernel, the latter seems to fit the true model better.

**Answer 2.a.** SVM was performed on the tumor dataset:

|  | RBF | Poly Deg:3 | Poly Deg 5 | Linear |
|---|---|---|---|---|
| $C = 1, \gamma = 1$ | 0.875606019914 | 0.944290431997 | 0.944267056588 | 0.959298439318 |
| $C = 1, \gamma = 0.01$ | 0.967831458327 | 0.959275561258 | 0.950626162554 | – |
| $C = 1, \gamma = 0.001$ | 0.965657545284 | 0.91427193061 | 0.746704067321 | – |
| $C = 0.01, \gamma = 1$ | 0.652374842092 | 0.944290431997 | 0.944267056588 | 0.965703798753 |
| $C = 0.01, \gamma = 0.01$ | 0.933513373718 | 0.952799578248 | 0.952800075597 | – |
| $C = 0.01, \gamma = 0.001$ | 0.65237484209 | 0.652374842092 | 0.652374842092 | – |
| $C = 0.0001, \gamma = 1$ | 0.652374842092 | 0.946440969632 | 0.944267056588 | 0.946370843405 |
| $C = 0.0001, \gamma = 0.01$ | 0.652374842092 | 0.701655675251 | 0.875513512976 | – |
| $C = 0.0001, \gamma = 0.001$ | 0.652374842092 | 0.652374842092 | 0.652374842092 | – |

Table 3 : Models and out of sample accuracy for the tumor dataset

The model that worked best for was with the following parameters 'C': 1, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf'
Best model selection was done using grid search using cross validation with train test split of 4-1 (5 fold cross validation).
**The kaggle test accuracy for this model was 0.95689 and the out of sample accuracy was 0.967831458327.**
Kaggle Display Name - Siddharth Chandrasekaran
Kaggle test accuracy was lower in this case than the estimated out of sample accuracy. RBF Kernel seems to give the best fit with a higher C and the median value of gamma. Gamma defines how much influence a single training point has on the model. This implies that assigning moderate importance to each data point fits our true model the most. Higher value of C reduces the smoothness of the curve and thus gives more flexibility to fit the data.

**Answer 3.** For this question a wide range of parameter space was experimented on.
Range of parameters experimented :
alpha : 0.00001 to 1
gamma : 0.00001 to 1
kernels : ’rbf’, ’linear’, ’poly’ (Degrees 3,4,5), ’cauchy’ (*Cauchy kernel is not a part of sklearn so it was custom implemented in the run_me file)
The best MSE is for the model ’alpha’: 1e-05, ’degree’: 3, ’gamma’: 0.0009, ’kernel’: ’poly’ Out of sample error was 0.0269760155785 and kaggle error was 1.49

**Answer 4.** For this question a wide range of parameter space was experimented on.
Range of parameters experimented :
C : 0.00001 to 1
gamma : 0.00001 to 1
kernels : ’rbf’, ’linear’, ’poly’ (Degrees 3,4,5), ’sigmoid’, cauchy’ (*Cauchy kernel is not a part of sklearn so it was custom implemented in the run_me file).

The best accuracy was for the model C = 0.014, gamma = 0.018, ’kernel’: ’rbf’ Out of sample accuracy was 0.972086777476 and kaggle accuracy was 0.95689