Homework 1

Cody Grogan
A02313514

September 2023

Problem 1

1.

MLP's are good at classification tasks or when regressing onto a function given a data set. An example of a good use for an MLP would MDP's be where spatial relationships are assumed to be unimportant like reinforcement learning.

2.

Convolutional Neural Networks use spatial relationships to determine its output. An example of a good use are the use of CNN's in cancer identification.

3.

Recurrent neural networks are good when the sequence of inputs is important. An example of an application of these are nlp's or language processing.

4.

Autoencoders attempt to compress data into the smallest for that will reproduce the same output. An example of a use for this is states that are observed using observations for reinforcement learning. For atari games, the state of the agent is usually described by an auto encoder then a decision is made based on that data.

5.

Generative adversarial networks are good at outputting new data from a set of data. An example of a use for this would be for determining cheaters in a video game. Trainers could give a generated video and real gameplay and have the discriminator determine whether the player is cheating or not.

6.

Deep reinforcement learning is for learning an optimal decision process by taking samples from the environment. An example of an application of deep reinforcement learning is teaching a robot how to walk with its own unique configuration.

7.

Large Language models are good at generating text given a series of inputs. An example of a good use for this is summarizing a large passage of text into a more concise format to cut down on time.

Problem 2

1.

The definition for a PD matrix is the following,

$$\{A \quad | \quad x^T Ax > 0 \quad \forall \quad x \neq 0\} \tag{1}$$

Adding together two PD matrices A and B we get,

$$x^T Ax + x^T Bx > 0 \tag{2}$$

$$x^T(A+B)x > 0 \implies A + B \succ 0 \tag{3}$$

2.

Adding a PD and PSD matrix always results in a PD matrix. This is evident because any positive number plus 0 is still greater than zero. So if we choose some x where $x^T Bx = 0$ the PD matrix A will still be greater than 0.

$$x^T(A+B)x > 0 \tag{4}$$

$$x^T Ax + x^T Bx > 0 \tag{5}$$

$$x^T Ax + 0 > 0 \tag{6}$$

$$x^T Ax > 0 \tag{7}$$

This is just the definition of a PD matrix showing the sum of a PD and PSD matrix is a PD matrix.

3.

a) A is neither because it is not a square matrix the the operation $x^T Ax$ cannot be done

b) $A^T A$ is PD because its eigen values are all greater than 0

c) AA^T is PSD because its eigen values are all greater than or equal to 0

d) B is PD because all eigen values are 1

e) -B is neither because all eigen values are -1

f) C is PD because all of its eigen values are greater than 0

g) C - .1B is PD because all of its eigen values are greater than 0

h) C - $.1AA^T$ is neither because all of its eigenvalues are not greater than 0

Problem 3

1.

a) The gradient of f_1 is

$$\nabla f_1 = \begin{bmatrix} \frac{\partial}{\partial v_1} f_1 \\ \frac{\partial}{\partial v_2} f_1 \end{bmatrix} \tag{8}$$

$$\nabla f_1 = \begin{bmatrix} 2v_1 + 3e^{v_2} \\ 3v_1 e^{v_2} \end{bmatrix} \tag{9}$$

The Hessian of f_1 is

$$H_1 = \nabla^2 f_1 \tag{10}$$

$$H_1 = \nabla \begin{bmatrix} 2v_1 + 3e^{v_2} & 3v_1 e^{v_2} \end{bmatrix} \tag{11}$$

$$H_1 = \begin{bmatrix} 2 & 3e^{v_2} \\ 3e^{v_2} & 3v_1 e^{v_2} \end{bmatrix} \tag{12}$$

b) The gradient of f_2 is,

$$\nabla f_2 = \begin{bmatrix} \frac{\partial}{\partial v_1} f_2 \\ \frac{\partial}{\partial v_2} f_2 \end{bmatrix} \tag{13}$$

$$\nabla f_2 = \begin{bmatrix} 12v_1^2 v_2 - v_2 \log(v_2) \\ 4v_1^3 - v_1 \log(v_2) - v_1 \end{bmatrix} \tag{14}$$

The Hessian of f_2 is,

$$H_1 = \nabla^2 f_2 \tag{15}$$

$$H_1 = \nabla \begin{bmatrix} 12v_1^2 v_2 - v_2 \log(v_2) & 4v_1^3 - v_1 \log(v_2) - v_1 \end{bmatrix} \tag{16}$$

$$H_1 = \begin{bmatrix} 24v_1 v_2 & 12v_1^2 - \log(v_2) - 1 \\ 12v_1^2 - \log(v_2) - 1 & -\frac{v_1}{v_2} \end{bmatrix} \tag{17}$$

c) The Jacobian of f is just the combination of the gradients,

$$J_f = \begin{bmatrix} 2v_1 + 3e^{v_2} & 3v_1 e^{v_2} \\ 12v_1^2 v_2 - v_2 \log(v_2) & 4v_1^3 - v_1 \log(v_2) - v_1 \end{bmatrix} \tag{18}$$

2.

First we look at the partial with respect to w_1 of the first term and defining the softmax of $x_1 w_1 = S_1$,

$$\frac{\partial}{\partial w_1} \frac{e^{x_1 w_1}}{\sum_{i=1}^d e^{x_i w_i}} \tag{19}$$

$$\frac{x_1 e^{x_1 w_1} (\sum_{i=1}^d e^{x_i w_i}) - x_1 e^{2x_1 w_1}}{(\sum_{i=1}^d e^{x_i w_i})^2} \tag{20}$$

$$\frac{x_1 e^{x_1 w_1} (\sum_{i=1}^d e^{x_i w_i} - e^{x_1 w_1})}{(\sum_{i=1}^d e^{x_i w_i})^2} \tag{21}$$

$$x_1 \frac{e^{x_1 w_1}}{\sum_{i=1}^d e^{x_i w_i}} (1 - \frac{e^{x_1 w_1}}{\sum_{i=1}^d e^{x_i w_i}}) \tag{22}$$

$$x_1 S_1 (1 - S_1) \tag{23}$$

Now looking at the off terms,

$$\frac{\partial}{\partial w_2} \frac{e^{x_1 w_1}}{\sum_{i=1}^d e^{x_i w_i}} \tag{24}$$

$$\frac{(0(\sum_{i=1}^d e^{w_i}) - x_2 e^{x_1 w_1} e^{x_2 w_2}}{(\sum_{i=1}^d e^{x_i w_i})^2} \tag{25}$$

$$-x_2 S_1 S_2 \tag{26}$$

Now that we know this, we can form the Jacobian of $w \otimes x$ is a matrix of the form,

$$\begin{bmatrix} x_1 S_1 (1 - S_1) & -x_2 S_1 S_2 & \cdots & -x_d S_1 S_d \\ -x_1 S_1 S_2 & x_2 S_2 (1 - S_2) & & \vdots \\ \vdots & & \ddots & -x_d S_{d-1} S_d \\ -x_1 S_1 S_d & \cdots & -x_{d-1} S(w_{d-1}) S(w_d) & x_d S_d (1 - S_d) \end{bmatrix} \tag{27}$$

Problem 4

a)

The bias of the estimator $\bar{X}$ is

$$Bias[\bar{X}] = E[\bar{X}] - \mu \tag{28}$$
$$= E[\frac{1}{n}\sum_{i=1}^{n} X_i] - \mu \tag{29}$$
$$= \frac{1}{n}E[X_1 + ... + X_n] - \mu \tag{30}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] - \mu \tag{31}$$
$$= \mu - \mu = 0 \tag{32}$$

b)

The variance of the estimator is,

$$Var[\bar{X}] = Var[\frac{1}{n}\sum_{i=1}^{n} X_i] \tag{33}$$
$$= \frac{1}{n^2}Var[\sum_{i=1}^{n} X_i] \tag{34}$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \tag{35}$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \tag{36}$$
$$= \frac{\sigma^2}{n} \tag{37}$$

c)

The MSE of an estimator is,

$$MSE(\bar{X}) = E[(\bar{X} - \mu)^2] \tag{38}$$
$$= E[\bar{X}^2 - 2\bar{X}\mu + \mu^2] \tag{39}$$

Now using the equation for variance,

$$Var[\bar{X}] = E[\bar{X}^2] - E[\bar{X}]^2 \tag{40}$$
$$Var[\bar{X}] + E[\bar{X}]^2 = E[\bar{X}^2] \tag{41}$$

Now substituting this in for $E[\bar{X}^2]$

$$MSE(\bar{x}) = Var[\bar{X}] + E[\bar{X}]^2 - 2\mu E[\bar{X}] + \mu^2 \tag{42}$$
$$= Var[\bar{X}] + (E[\bar{X}] - \mu)^2 \tag{43}$$
$$= \frac{\sigma^2}{n} \tag{44}$$

d)

Starting with the equation for bias of the estimator,

$$Bias[\hat{s}^2] = E[\hat{s}^2] - \sigma^2 \tag{45}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \bar{X})^2] - \sigma^2 \tag{46}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \bar{X})^2] - \sigma^2 \tag{47}$$
$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i^2 - 2X_i\bar{X} + \bar{X}^2] - \sigma^2 \tag{48}$$
$$= \frac{1}{n}\sum_{i=1}^{n} (E[X_i^2] - 2E[X_i]E[\bar{X}] + E[\bar{X}^2]) - \sigma^2 \tag{49}$$
$$= \frac{1}{n}\sum_{i=1}^{n} (E[X_i^2] - 2E[\bar{X}^2] + E[\bar{X}^2]) - \sigma^2 \tag{50}$$
$$= \frac{1}{n}\sum_{i=1}^{n} (E[X_i^2] - E[\bar{X}^2]) - \sigma^2 \tag{51}$$
$$= \frac{1}{n}\sum_{i=1}^{n} (E[X_i^2]) - E[\bar{X}^2] - \sigma^2 \tag{52}$$
$$= \frac{1}{n}[\sum_{i=1}^{n} (E[X_i^2]) - n(Var[\bar{X}] + E[\bar{X}]^2)] - \sigma^2 \tag{53}$$
$$= \frac{1}{n}[\sum_{i=1}^{n} (E[X_i^2]) - n(\frac{\sigma^2}{n} + \mu^2)] - \sigma^2 \tag{54}$$
$$\tag{55}$$

Now using the property of variance for a random variable X_i,

$$Var[X_i] = E[X_i^2] - E[X_i]^2 \tag{56}$$
$$\sigma^2 = E[X_i^2] - \mu^2 \tag{57}$$
$$\sigma^2 + \mu^2 = E[X_i^2] \tag{58}$$

Now substituting this in for $E[X_i^2]$,

$$Bias[\hat{s}^2] = \frac{1}{n}[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)] - \sigma^2 \tag{59}$$

$$= \frac{1}{n}[n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)] - \sigma^2 \tag{60}$$

$$= \frac{1}{n}[n\sigma^2 - \sigma^2] - \sigma^2 \tag{61}$$

$$= \frac{1}{n}[\sigma^2(n-1)] - \sigma^2 \tag{62}$$

$$= \frac{\sigma^2(n-1)}{n} - \sigma^2 \tag{63}$$

$$= -\frac{\sigma^2}{n} \tag{64}$$

Yes, the way to do this would be to divide by (n-1) and define the variance estimator as,

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{65}$$

2.

For this problem we are trying to prove that,

$$I(X;Y) = -\int f_{XY}(x,y) log(\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)}) dxdy = 0 \tag{66}$$

From the definition of independent random variables, we know the following,

$$f_{XY}(x,y) = f_X(x)f_Y(y) \tag{67}$$

Substituting this in we get,

$$I(X;Y) = -\iint_{-\infty}^{\infty} f_X(x)f_Y(y) log(\frac{f_X(x)f_Y(y)}{f_X(x)f_Y(y)}) dxdy \tag{68}$$

$$= \iint_{-\infty}^{\infty} f_X(x)f_Y(y) log(1) dxdy \tag{69}$$

$$= \iint_{-\infty}^{\infty} 0 dxdy \tag{70}$$

$$= 0 \tag{71}$$

Problem 5

Starting with the equation for MSE where w is the weights concatenated with the offset and X is the n x d+1 matrix.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i w - y_i)^2 \tag{72}$$

From this we know that we remove the sum using the X matrix and Y matrix as the column vector of the targets,

$$\begin{aligned} MSE &= \frac{1}{n}(Xw - Y)^T(Xw - Y) && (73)\\ &= \frac{1}{n}(w^T X^T - Y^T)(Xw - Y) && (74)\\ &= \frac{1}{n}(w^T X^T Xw - Y^T Xw - w^T X^T Y + Y^T Y) && (75) \end{aligned}$$

From the definition of the transpose we find $Y^T Xw = (w^T X^T Y)^T$. In addition, we know the result is a scalar so we can add the two center terms together.

$$MSE = \frac{1}{n}(w^T X^T Xw - 2w^T X^T Y + Y^T Y) \tag{76}$$

Now taking the gradient with respect to w of the MSE,

$$\begin{aligned} \nabla_w MSE &= \nabla_w \frac{1}{n}(w^T X^T Xw - 2w^T X^T Y + Y^T Y) && (77)\\ &= \frac{1}{n}(\nabla_w w^T X^T Xw - \nabla_w 2w^T X^T Y + \nabla_w Y^T Y) && (78) \end{aligned}$$

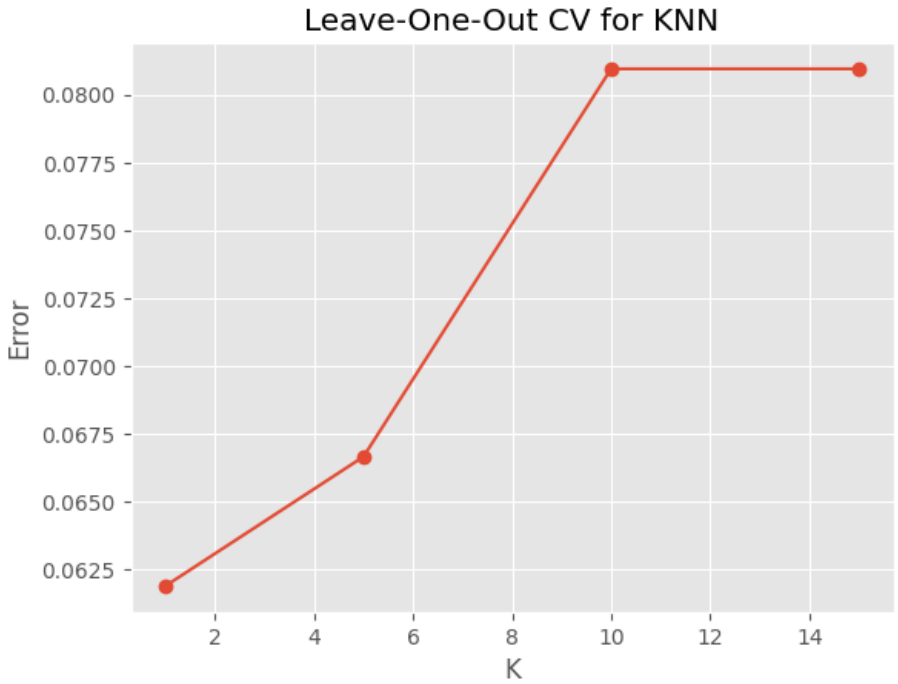
We also know for symmetric matrices, such as $X^T X$, that $\nabla_x w^T Aw = 2Aw$. So substituting this in, setting the equation equal to 0 to get $\hat{w}$.

$$\begin{aligned} 0 &= \frac{1}{n}(2X^T X\hat{w} - 2X^T Y) && (79)\\ 0 &= X^T X\hat{w} - X^T Y && (80)\\ X^T Y &= X^T X\hat{w} && (81)\\ \hat{w}s &= (X^T X)^{-1} X^T Y && (82) \end{aligned}$$

Problem 6

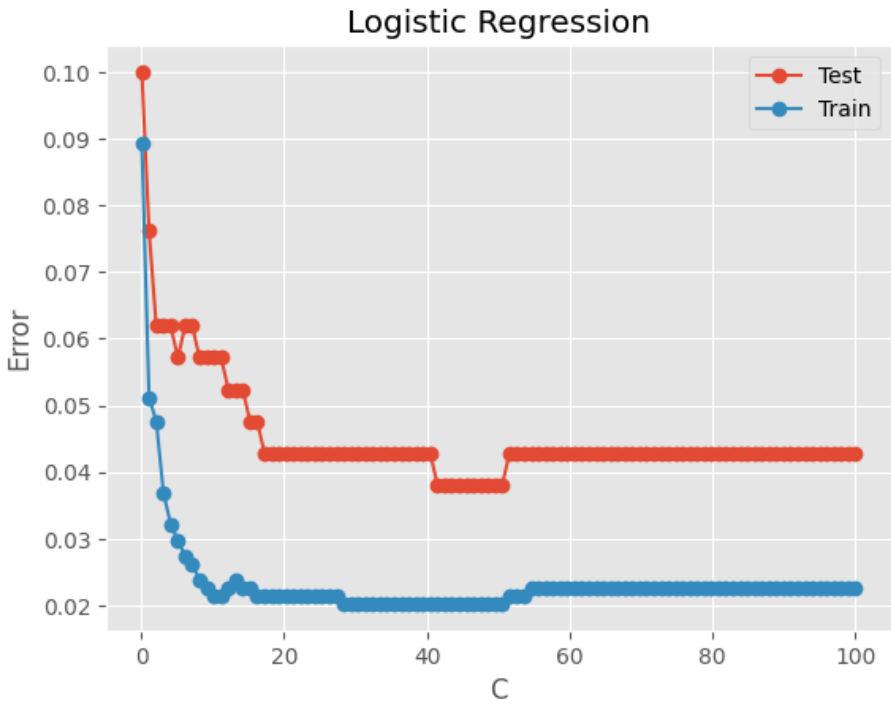
1.

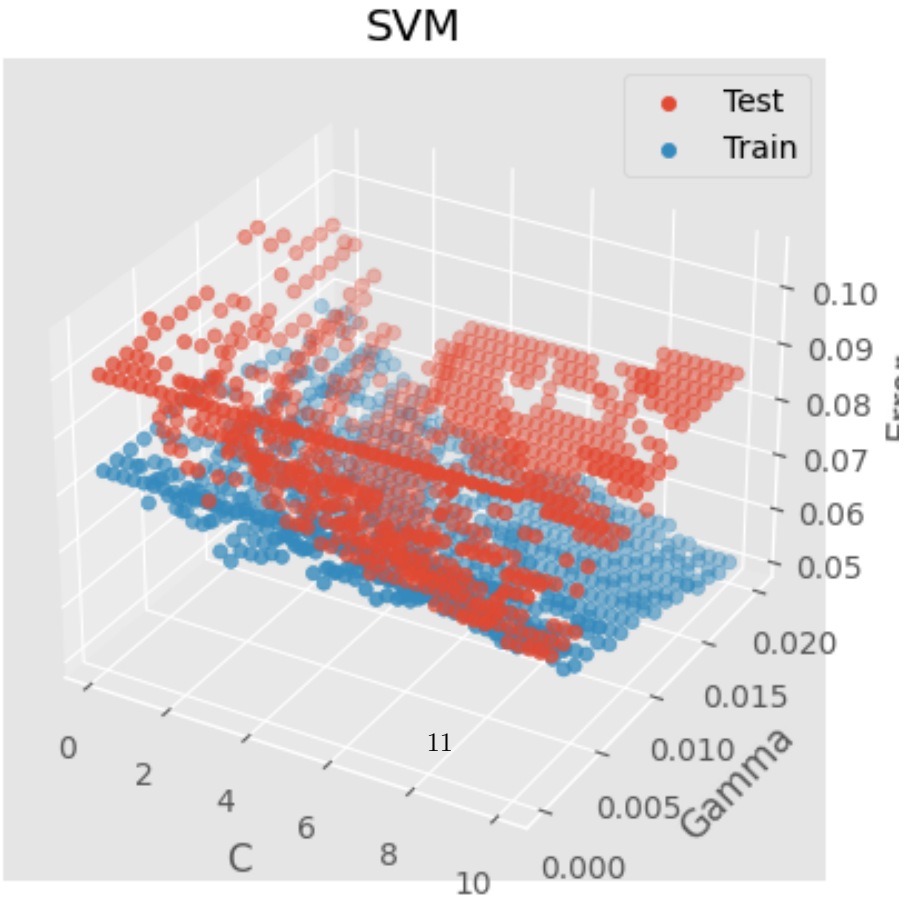
For the 5 fold cross validation is the low values of k like 1 and 5 are under fitted to the data. This is because the performance increases as the number of neighbors goes up. From this figure we see the best values of k are either 10 or 15

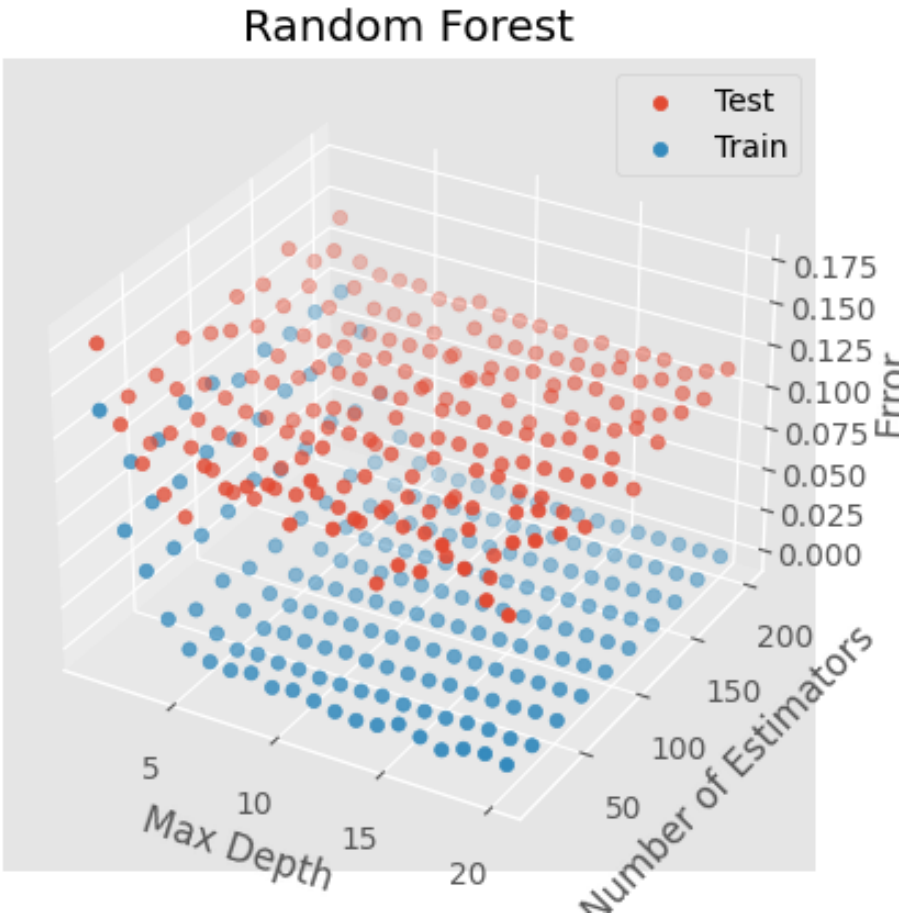


This graphic shows that the KNN classifier is actually overfitting to the data with increasing k. This is contrary to the results to the 5 fold validation but because LOO is a better approximation of expected test error we can say with certainty that overfitting is occuring. This is especially evident for values of 10 and 15 which performed the best in the previous section. The best value of k for this CV is k=1.

2.


Logistic Regression
Test
Train
Error
C
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0
20
40
60
80
100


SVM
Test
Train
C
Gamma
0
2
4
6
8
10
0.10
0.09
0.08
0.07
0.06
0.05
0.020
0.015
0.010
0.005
0.000



The classifier that performed the best was the Logistic Regression classifier with a test error of 3.8% and a training error of 2%.

In the parameter sweep, we can see the logistic regression classifier would over fit for regularization coefficients greater than 50 and under fit for values lower than 40. For the SVM, the parameter sweep yielded a variety of under and overfitted parameters. And, in most cases, the optimal parameters would have a higher training error than testing error indicating overfitting. The random forest classifier had a much lower chance of overfitting to the data. This is evident because the training error is always higher than the testing error.

The overall best classifier was the logistic regression classifier. Followed by the SVM, KNN, and random forest. However the SVM and KNN classifier may be about even with the overfitting of the SVM to the data.