Name:

Karan Pandya

Netid:

karandp2

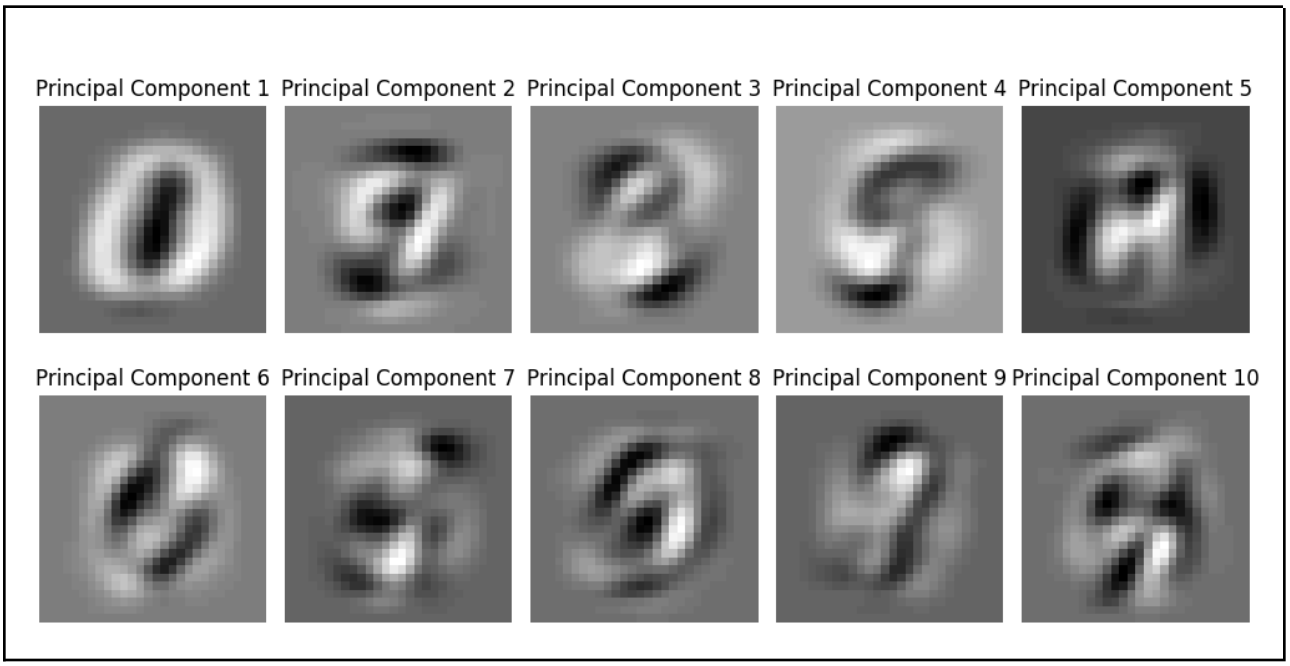
CS 441 - HW2: PCA and Linear Models

Complete the sections below. You do not need to fill out the checklist.

Total Points Available	**[ ] / 160**
1. PCA on MNIST	
a. Display 10 principal component vectors	[ ] / 5
b. Display scatterplot	[ ] / 5
c. Plot cumulative explained variance	[ ] / 5
d. Compression and 1-NN experiment	[ ] / 15
2. MNIST Classification with Linear Models	
a. LLR / SVM error vs training size	[ ] / 20
b. Error visualization	[ ] / 10
c. Parameter selection experiments	[ ] / 15
3. Temperature Regression	
a. Linear regression test	[ ] / 10
b. Feature selection results	[ ] / 15
4. Stretch Goals	
a. PR and ROC curves	[ ] / 10
b. Visualize weights	[ ] / 10
c. Other embeddings	[ ] / 15
d. One city is all you need	[ ] / 15
e. SVM with RBF kernel	[ ] / 10

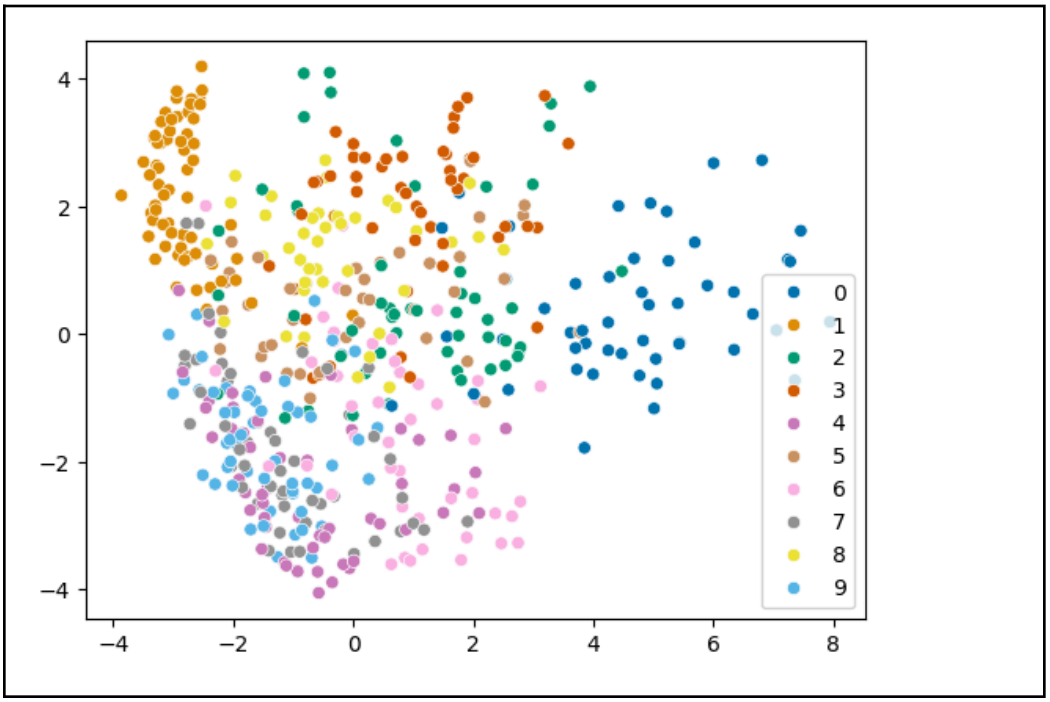
1. PCA on MNIST

a. Display 10 principal component vectors

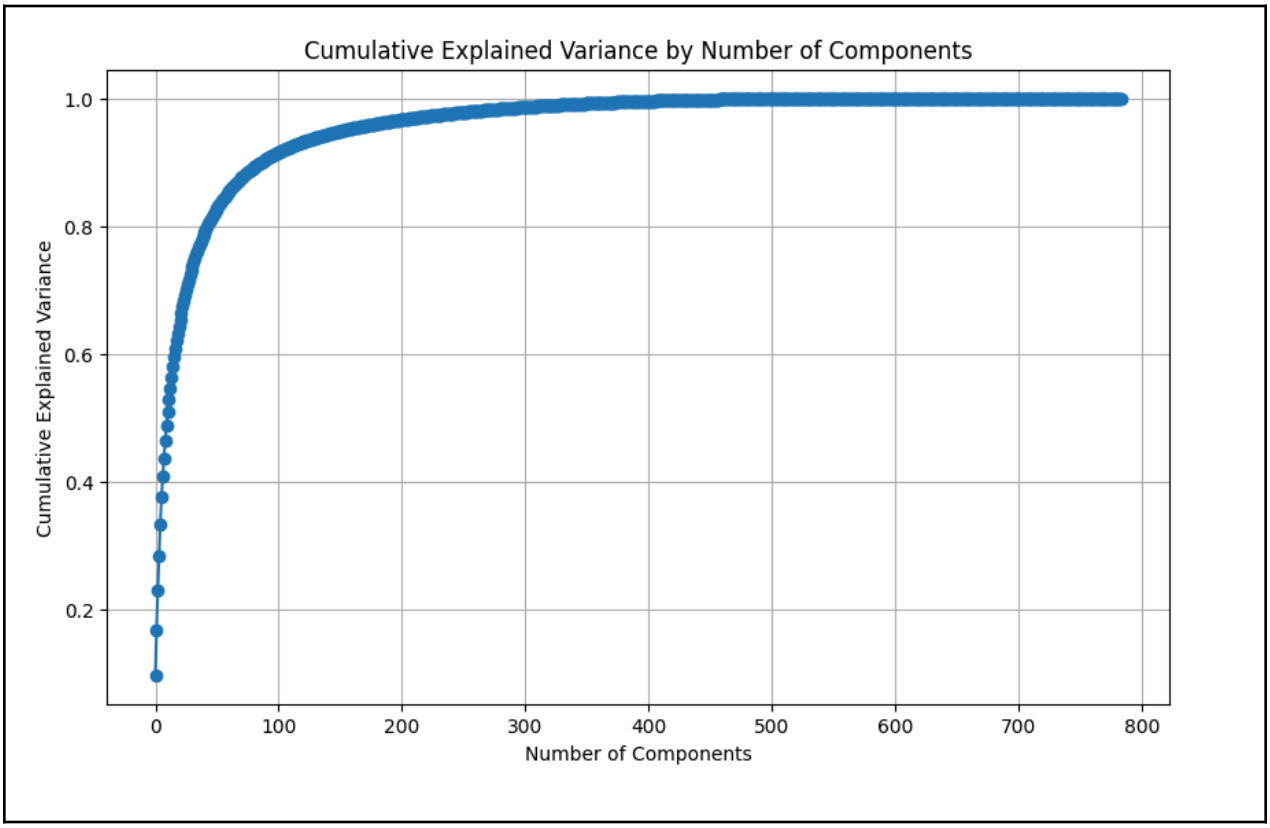


b. Display scatterplot

Scatterplot `x_train[:500]` for the first two PCA dimensions. Show a different color for each label.



c. Plot cumulative explained variance



d. Compression and 1-NN experiment

Number of components selected = 100

	Total Time (s)	Test Error (%)	Dimensions
Brute Force (PCA)	7.81	2.68	87
Brute Force	23.26	3.09	784

2. MNIST Classification with Linear Models

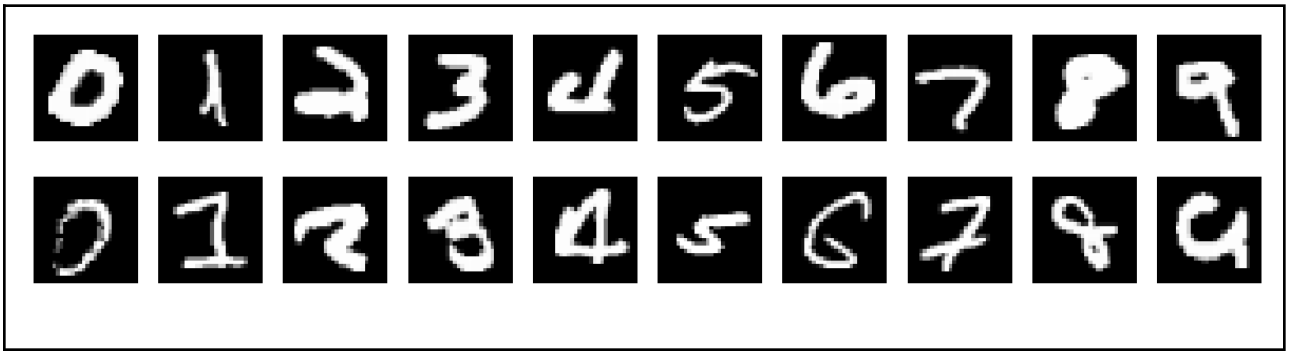
a. LLR / SVM error vs training size

Test error (%)

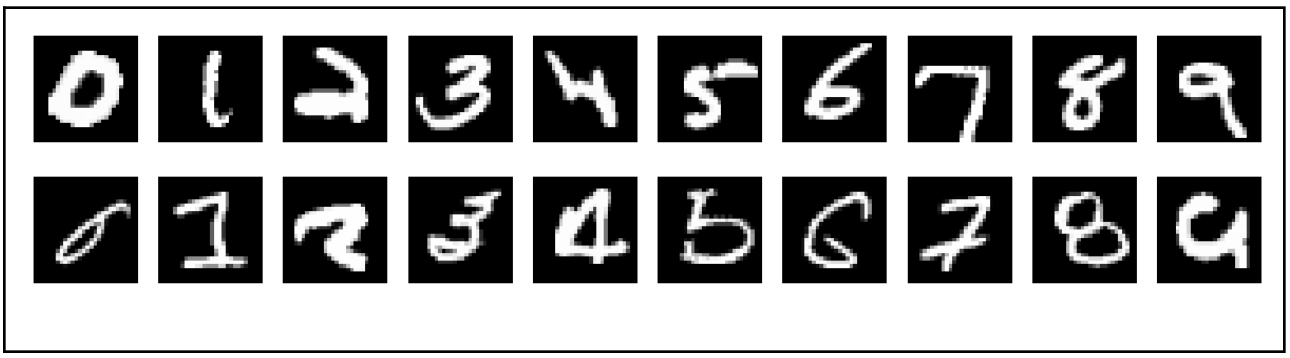
# training samples	LLR(%)	SVM(%)
100	32.5	32.4
1,000	13.78	16.12
10,000	9.49	11.11
60,000	7.44	8.17

b. Error visualization

LLR



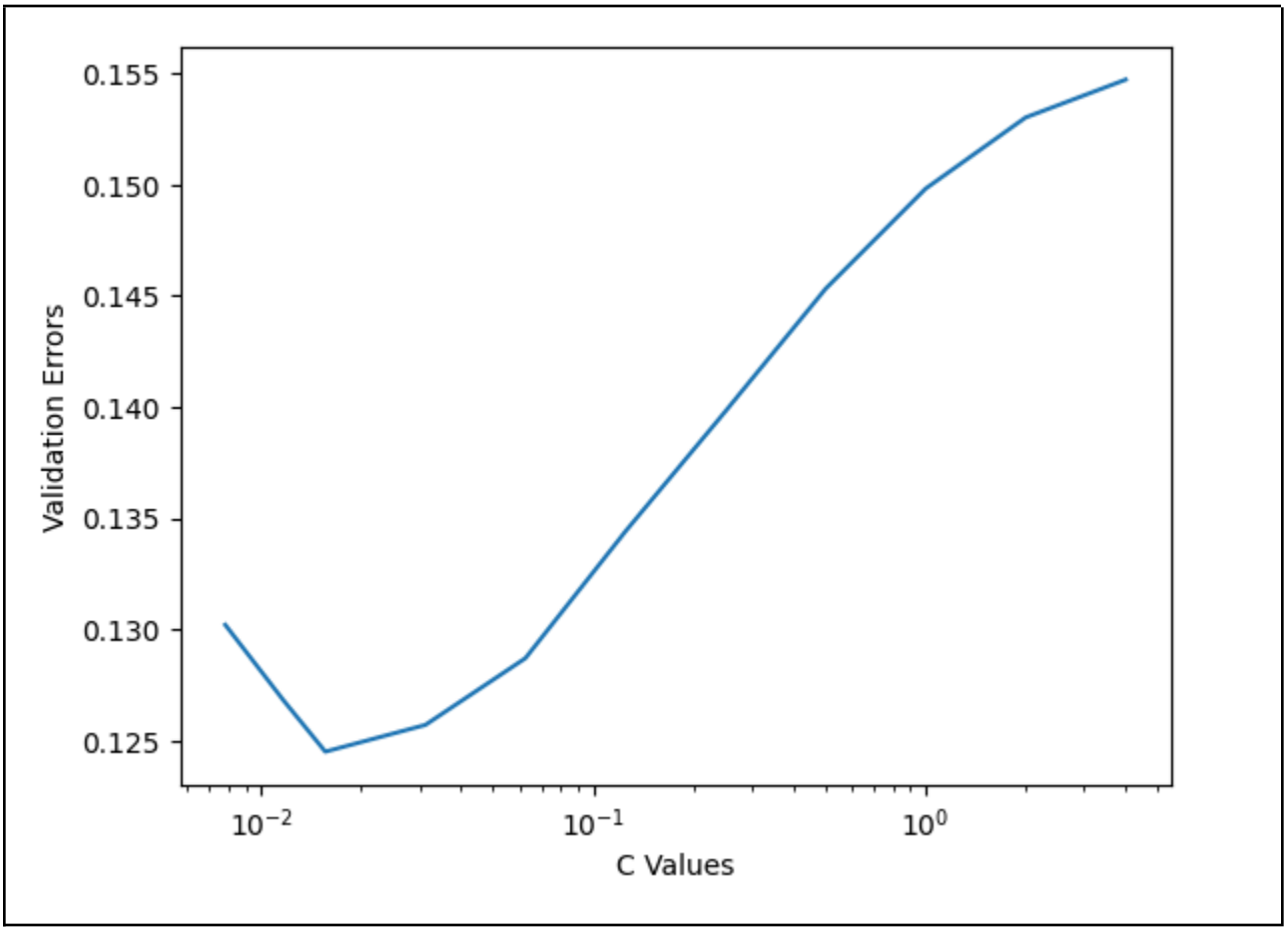
SVM



c. Parameter selection experiments

	Logistic Regression
Best C value	0.015625
Validation error (%)	12.45
Test error (%)	13.62

Plot C value vs validation error for values tested



3. Temperature Regression

a. Linear regression test

Test RMSE

	Linear regression
Original features	2.16

Normalized features	2.163

Why might normalizing features in this way not be as helpful as it is for KNN?

As Ridge regression is a linear model, normalizing the features does not change the linear relationship between the variables hence it is not as effective as for KNN

b. Feature selection results

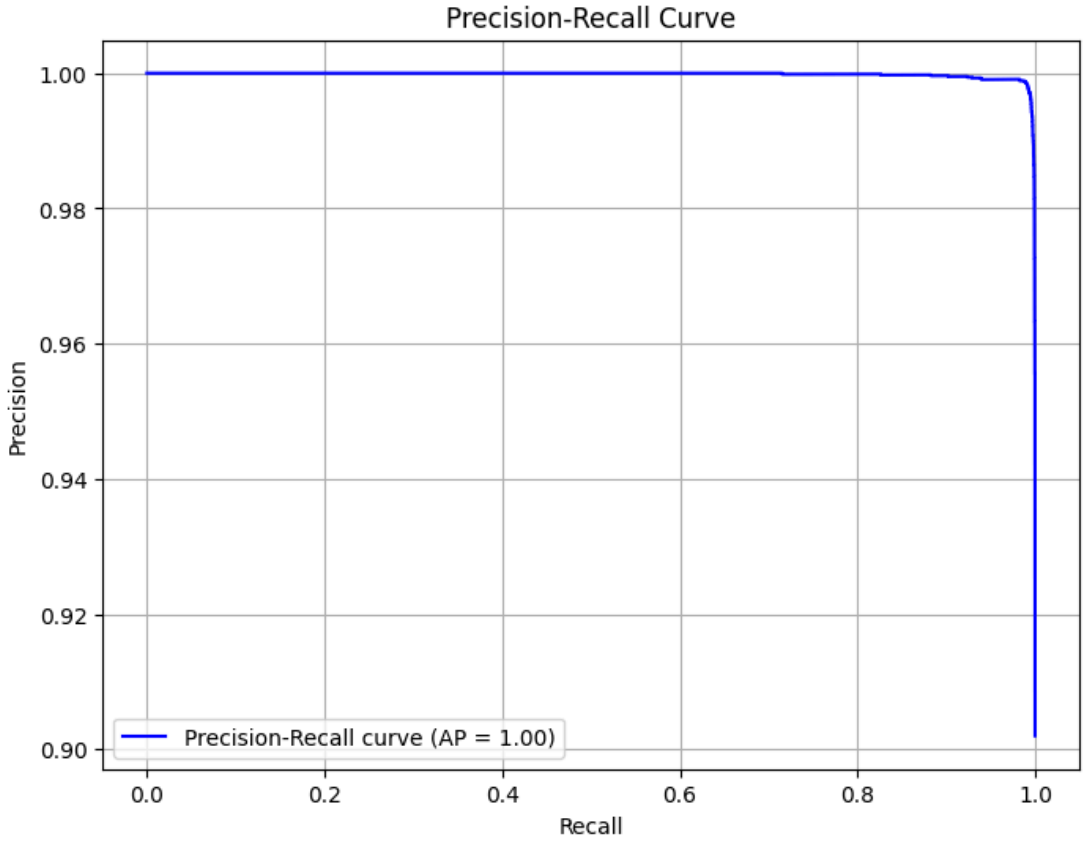
Feature Rank	Feature number	City	Day
1	334	Chicago	-1
2	347	Minneapolis	-1
3	405	Grand Rapids	-1
4	366	Kansas City	-1
5	361	Cleveland	-1
6	307	Omaha	-2
7	367	Indianapolis	-1
8	264	Minneapolis	-2
9	9	Boston	-5
10	236	SpringField	-3

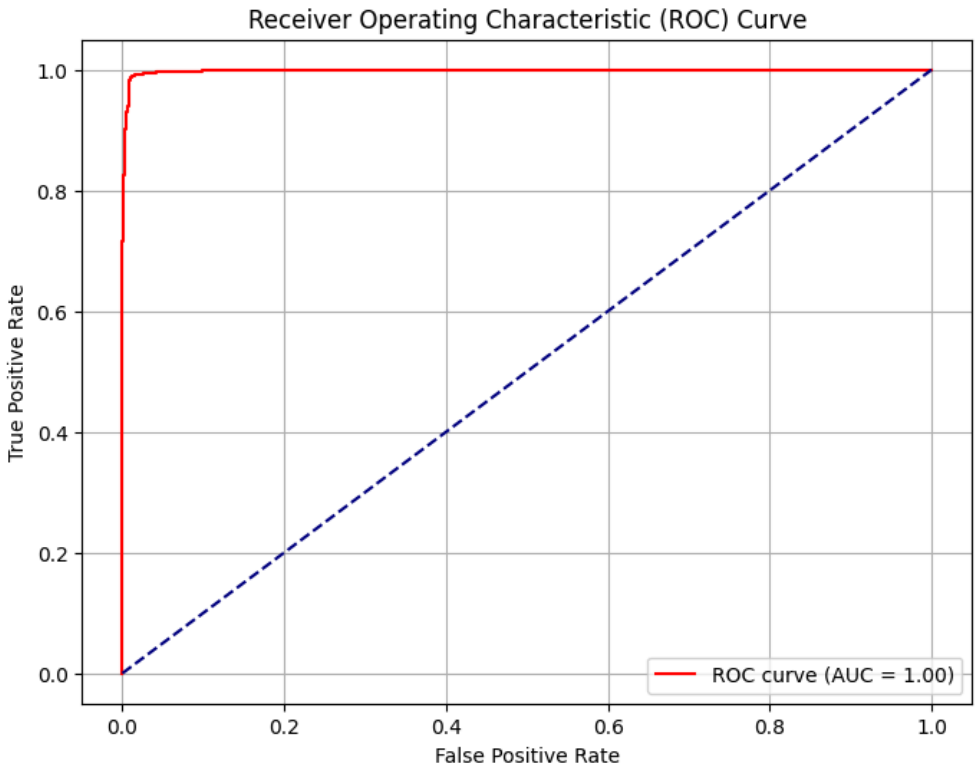
Test error using only the 10 most important features for regression

	Linear Regression
RMS Error	2.0621

4. Stretch Goals

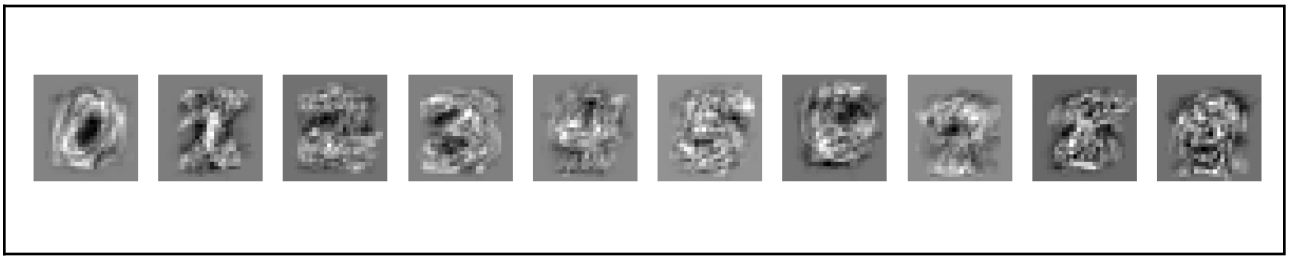
a.


Precision-Recall Curve
Precision
Recall
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
1.0
Precision-Recall curve (AP = 1.00)

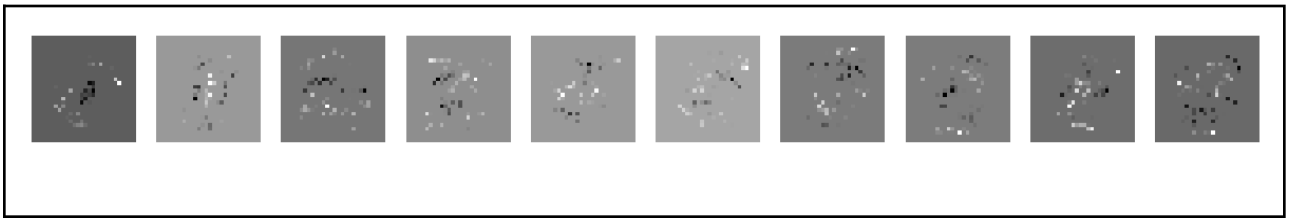


b. Visualize weights

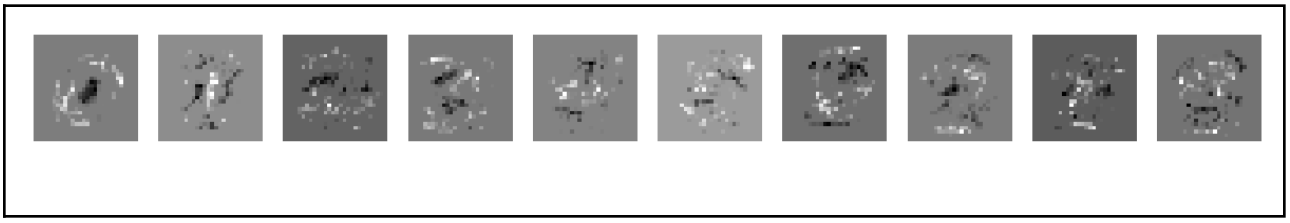
LLR - L2



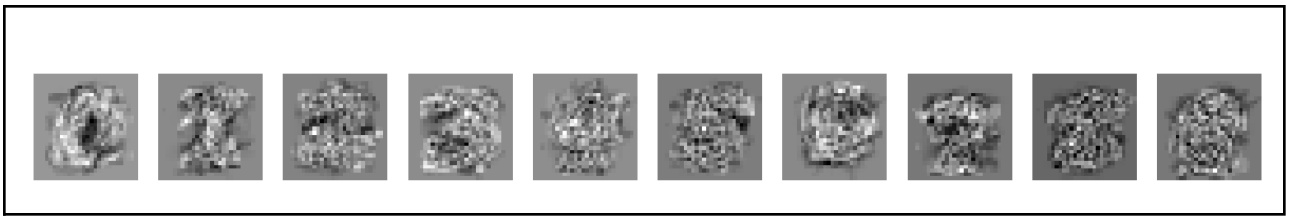
LLR - L1



LLR - elastic



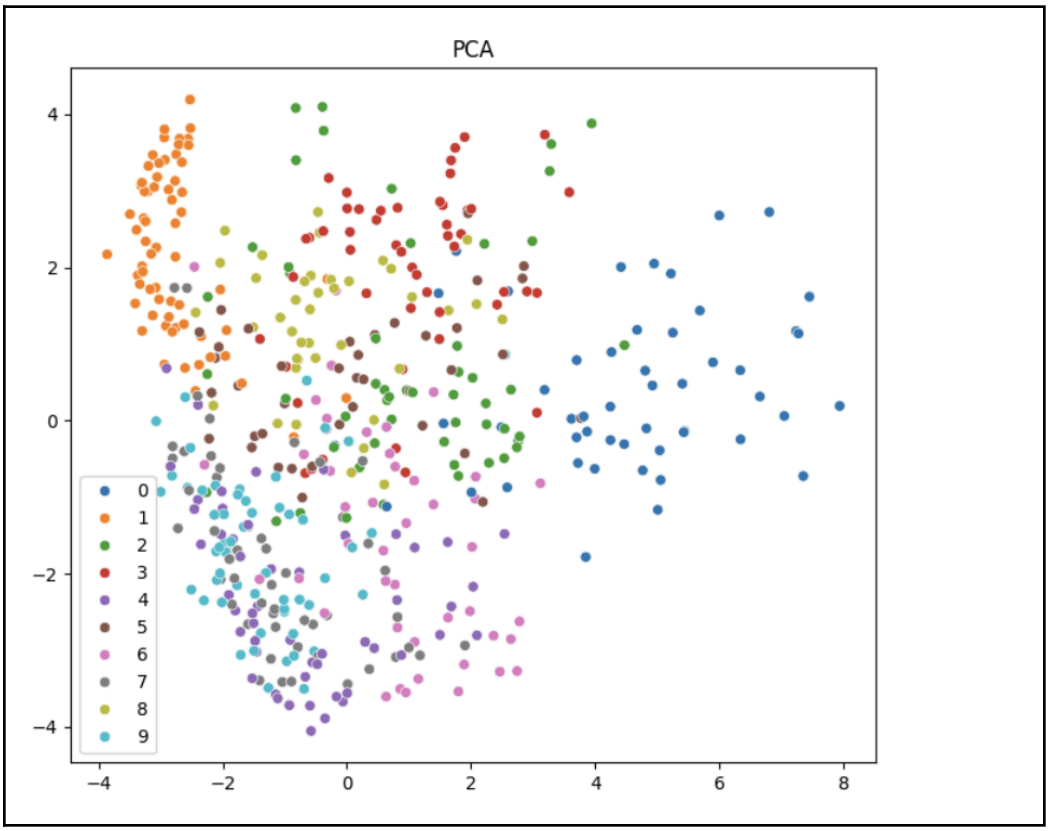
SVM



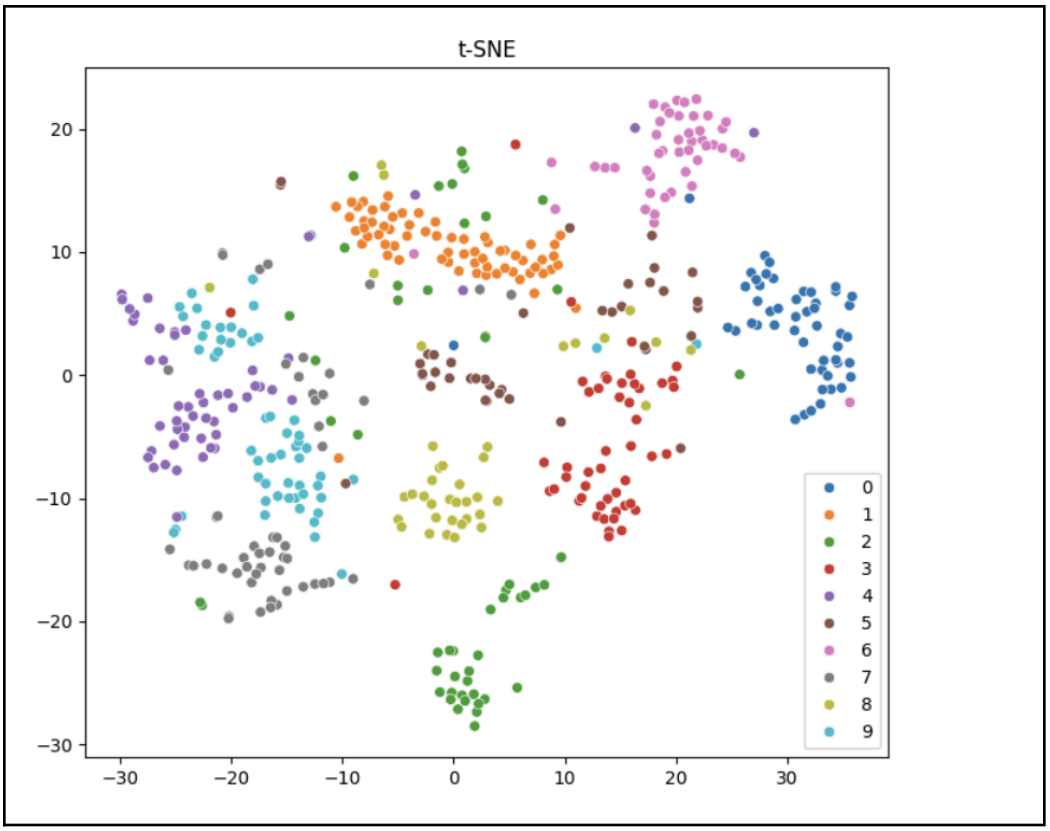
c. Other embeddings

Display 2+ plots for TSNE, MDA, and/or LDA, and copy PCA plot from 1b here.

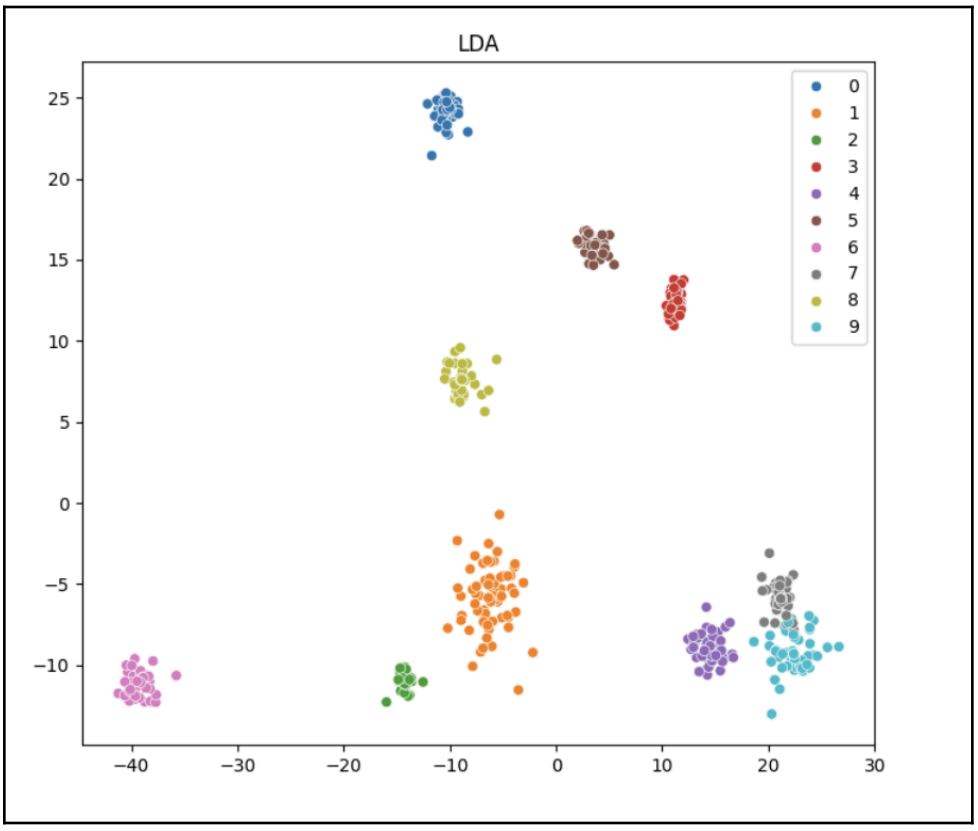
PCA



[t-SNE]


t-SNE
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
0
1
2
3
4
5
6
7
8
9

LDA


LDA
25
20
15
10
5
0
−5
−10
−40
−30
−20
−10
0
10
20
30
0
1
2
3
4
5
6
7
8
9

d. One city is all you need

City

St. Louis

Test error using features only from that city

4.979

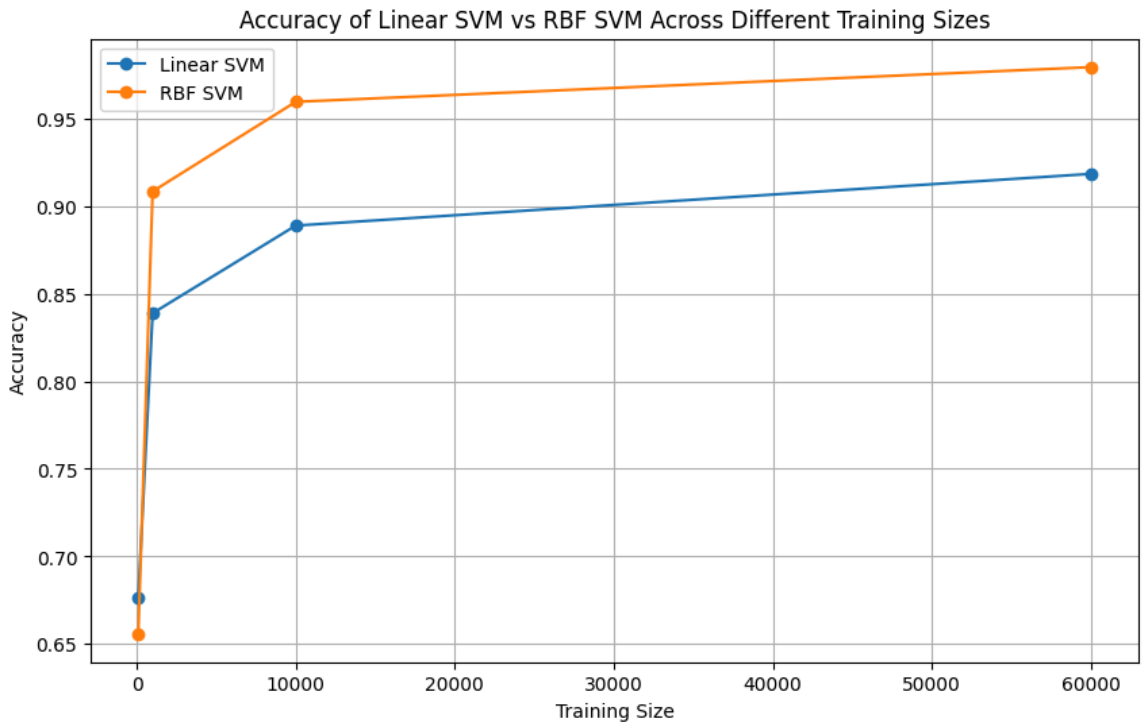
Explain your process (in words):

1) I calculated RMSE using only the temperature of the past 5 days of one city and did this one by one for all cities.
2) To do this I used Ridge and Lasso models, and for both models St. Louis gave the lowest RMSE(2.7256) on the validation set.
3) Then I calculated the RMSE on the test set using the temperature of the past 5 days of St. Louis only and got an RMSE of 4.979.

e. Compare linear SVM and SVM with RBF kernel

Test accuracy (%)

# training samples	SVM-Linear	SVM-RBF
100	32.4%	34.4%
1,000	16.2%	9.17%
10,000	11.11%	4.05%
60,000	8.17%	2.08%


Accuracy of Linear SVM vs RBF SVM Across Different Training Sizes
Linear SVM
RBF SVM
Accuracy
Training Size
0.65
0.70
0.75
0.80
0.85
0.90
0.95
0
10000
20000
30000
40000
50000
60000