Name:

Darian Irani

Netid:

irani2

CS 441 - HW1: Instance-based Methods

Complete the claimed points and sections below. *There is a 5 point penalty for failing to complete this section.*

Total Points Claimed	**[145] / 145**
1. Retrieval, K-means, 1-NN on MNIST	
a. Retrieval	[5] / 5
b. K-means	[15] / 15
c. 1-NN	[10] / 10
2. Make it fast	
a. K-means plot	[15] / 15
b. 1-NN error plots	[8] / 8
c. 1-NN time plots	[7] / 7
d. Most confused label	[5] / 5
3. Temperature Regression	
a. RMSE Tables	[20] / 20
4. Conceptual questions	[15] / 15
5. Stretch Goals	
a. Evaluate effect of K for MNIST	[15] / 15
b. Evaluate effect of K for Temp Reg.	[15] / 15
c. Compare Kmeans more iterations vs. restarts	[15] / 15

1. Retrieval, K-means, 1-NN on MNIST

a. What index is returned for x_test[1]?

28882

b. Paste the display of clusters after the 1st and 10th iteration for K=30.

1st iteration:

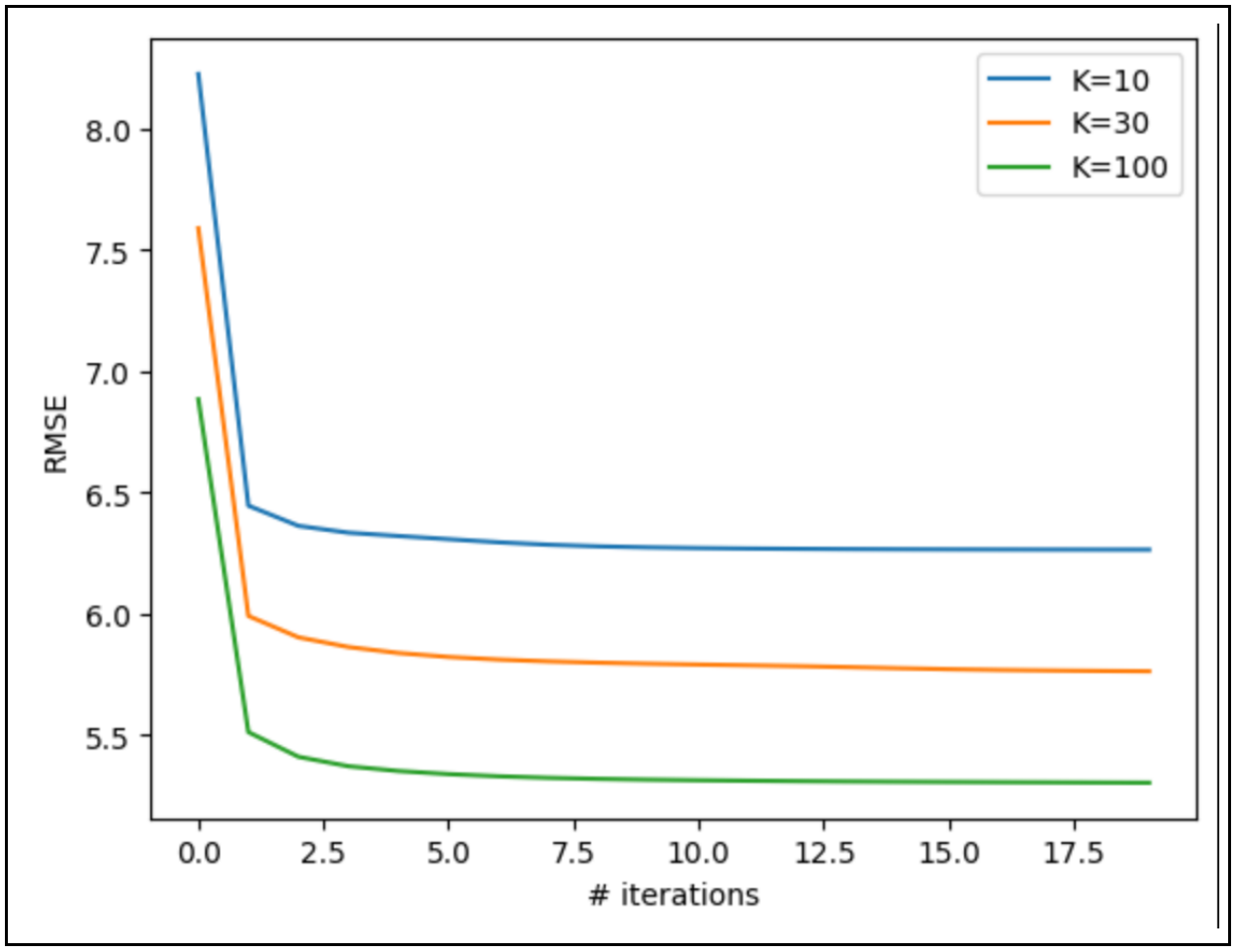
10th iteration:

c. Error rate for first 100 test samples, using first 10,000 training samples (x.x)

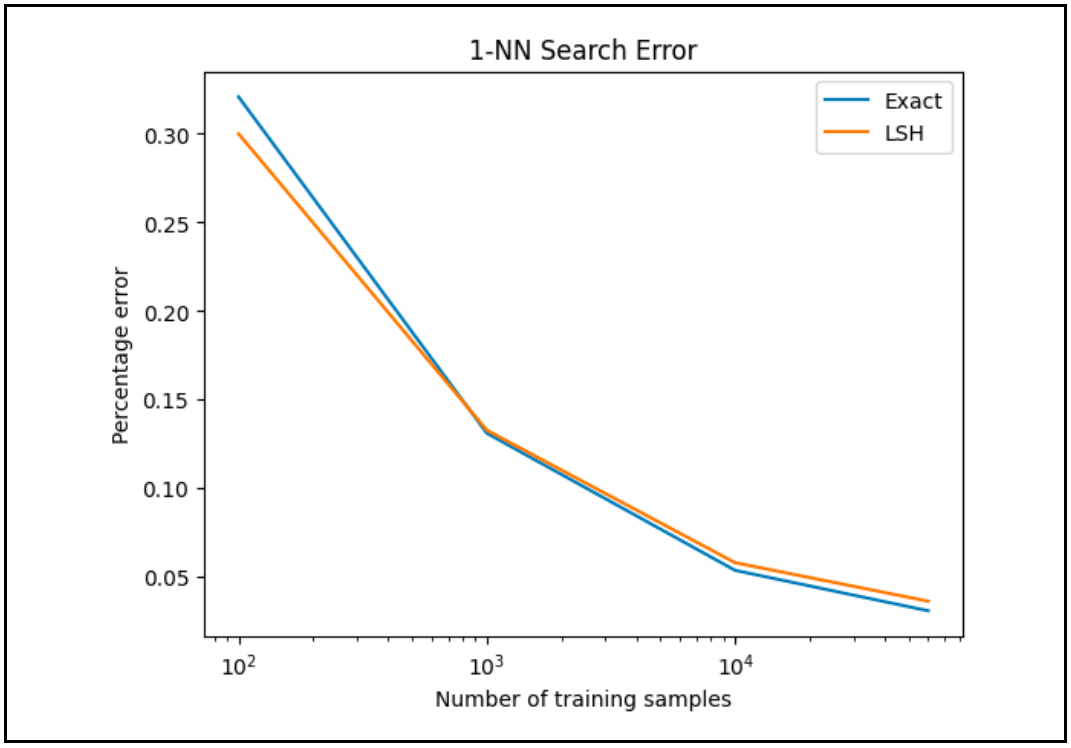
0.1

2. Make it fast

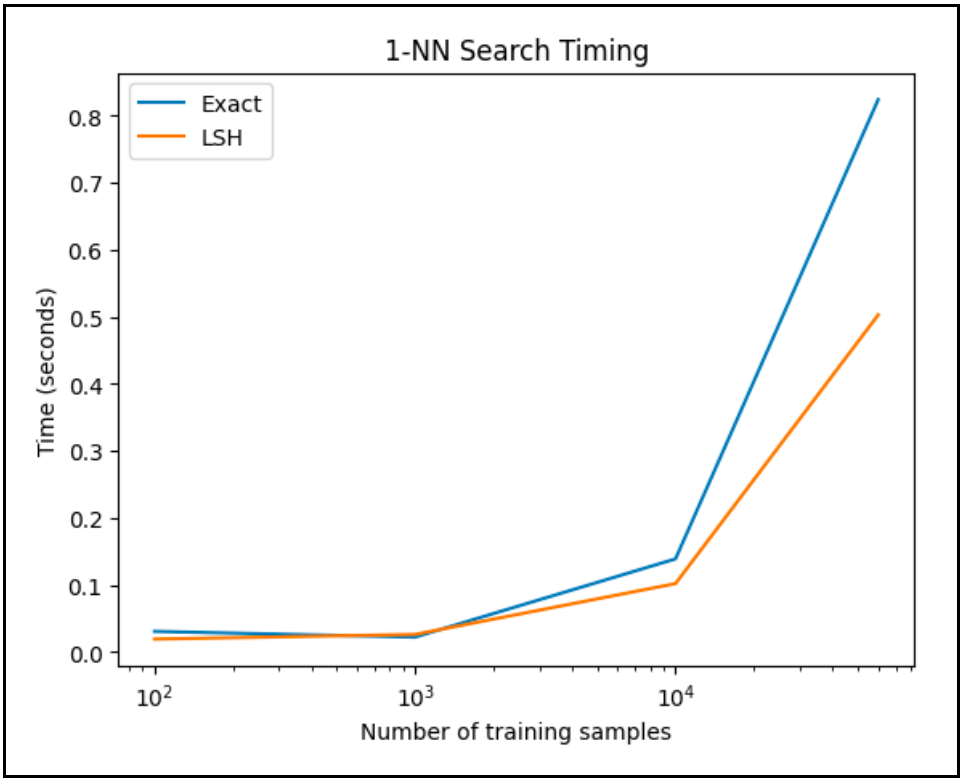
a. KMeans plot of RMSE vs iterations for K=10, 30, 100



b. Nearest neighbor error vs training size plot



c. Nearest neighbor time vs training size plot



d. What label is most commonly confused with ‘2’?

7

3. Temperature Regression

a. Table of RMSE for KNN with K=5 (x.xx)

	KNN (K=5)
Original Features	3.25
Normalized Features	3.17

4. Test your understanding

Fill in the letter corresponding to the answer. If you’re not sure, you can sometimes run small experiments to check.

1. Is K-means guaranteed to decrease RMSE between nearest cluster and samples at each iteration until convergence?
    a. Yes
    b. No

Yes, in the update step the recalibration of cluster centers to the mean of assigned samples ensures distances are minimized for future iterations.

2. If you increase K, is K-means expected or guaranteed to achieve lower RMSE?
    a. Guaranteed
    b. Expected but not guaranteed
    c. Not expected

Guaranteed, each additional cluster provides an opportunity to reduce the distance between points and their nearest center. Therefore, average squared distance between points and centroids is reduced resulting in a lower RMSE.

3. In K-NN regression, for training labels y, what is the lowest target value that can possibly be predicted for any query?
    a. Min(y)
    b. Mean(y)
    c. Can’t be determined

Min(y), the theoretical lower bound for a K-NN regression is the minimum value of the training label y, nothing else in KNN results in a value less than Min(y)

4. Would you expect the “training error” for 1-NN to be higher or lower than 3-NN for classification? Training error is the error if you test on the training data.
    a. Higher
    b. Lower
    c. It’s problem-dependent

Lower, the training error for 1-NN is zero since its neighbor is its own point (has no neighbors).

5. Would you expect the test error for 1-NN to be higher or lower than for 3-NN for regression?
    a. Higher
    b. Lower
    c. It’s problem-dependent

It’s problem-dependent, 3-NN would normally decrease test error due to its relative insensitivity to noisy neighbors however for some datasets a smaller K might capture trends more accurately – depending on the kind of data and its variance.

5. Stretch Goals (optional)

a. Select best K parameter for K-NN MNIST classification in K=1, 3, 5, 11, 25.

(x.xx)

Validation Set Performance	K=1	K=3	K=5	K=11	K=25
% error	2.88	2.80	2.82	3.08	3.82

Best K:

3

Test % error (x.xx)

2.83

b. Select best K parameter for K-NN temperature regression in K=1, 3, 5, 11, 25.

Validation Set RMSE	K=1	K=3	K=5	K=11	K=25
Original Features	4.33	3.23	3.10	3.06	3.06
Normalized Features	3.87	3.17	3.03	2.89	2.91

Best Setting (K, feature type):

11, normalized

Test RMSE (x.xx)

2.77

c. Kmeans: compare average and standard deviation RMSE based on number of iterations and number of restarts

(4 digit precision)

K=30	RMSE avg	RMSE std
20 iterations, 1 restart	5.7863	0.0076
4 iterations, 5 restarts	5.8228	0.0121
50 iterations, 1 restart	5.7771	0.0055
10 iterations, 5 restarts	5.7876	0.0037

Acknowledgments / Attribution

https://www.w3schools.com/python/python_ml_k-means.asp

https://stackoverflow.com/questions/33458834/k-means-clustering-in-python

https://codereview.stackexchange.com/questions/154609/knn-algorithm-implemented-in-python