**Name (netid):** Your Name (Your Netid)

**CS 441 - HW1: Intro to Classification and Regression**

Complete the claimed points and sections below.

**Total Points Claimed [ ] / 160**

1. MNIST Classification
   1. Main Results
      1. KNN [10] / 10
      2. Naive Bayes [10] / 10
      3. Logistic Regression [10] / 10
      4. Confusion Analysis [10] / 10
   2. comparison over data size [10] / 10
   3. parameter selection [10] / 10
2. Temperature Regression
   1. Main Results
      1. KNN [10] / 10
      2. Naive Bayes [10] / 10
      3. Linear Regression [10] / 10
   2. Most Important Features [10] / 10
3. Stretch Goals
   1. Improve MNIST classification [ ] / 20
   2. Improve Temperature regression [ ] / 20
   3. Create a dataset where NB beats NN/LR [ ] / 20
4. **MNIST**
5. **Main Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naive Bayes | Logistic Regression |
| Val Error | 3.04% | 36.25% | 7.28% |
| Training Time (s) | 5003 | 37 | 300 |
| Inference Time (ms) | 500.3 | 3.7 | 30 |

Using your confusion matrix for KNN, a report which label is most commonly confused with each true label. For example, if the most common mistake for “3” images is to assign them to “8”, then put “8” in the second row under “3” and the percent of “3”s that are classified as “8” in the third row.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **Most common mistaken label** | **6** | **7** | **1** | **5** | **9** | **3** | **0** | **9** | **1** | **7** |
| **% of times assigned to that label** | **0.7%** | **0.5%** | **2.6%** | **2.6%** | **3.8%** | **1.7%** | **0.8%** | **1.5%** | **2.66%** | **1.34%** |

1. **Plot Validation Error vs. Training Size**

|  |  |  |  |
| --- | --- | --- | --- |
| # training samples | KNN | Naive Bayes | Logistic Regression |
| 50 | 33.82% | 67.33% | 34.75% |
| 500 | 16.17% | 52.83% | 15.21% |
| 5,000 | 6.87% | 35.19% | 10.76% |
| 50,000 | 3.1% | 35.1% | 7.17% |

1. **Parameter selection**

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN (K) | Naive Bayes (alpha) | Logistic Regression (C) |
| Best parameter | 1 | 1 | 1 |
| Validation error | 12.22% | 40.05% | 10.01% |
| Test error | 11.11% | 37.81% | 9.91% |

**2. Temperature Prediction**

1. **Main Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naive Bayes | Linear Regression |
| RMS Error | 3.227 | 3.785 | 2.445 |
| Median Abs Error | 1.443 | 1.569 | 1.282 |

1. **Most Important Features**

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

Using only the 10 most important features

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naive Bayes | Linear Regression |
| RMS Error | 2.262 | 2.79 | 2.27 |
| Median Abs Error | 1.114 | 1.229 | 1.115 |

**3. Stretch Goals**

1. **Improve MNIST classification performance**

Report the classification val and test errors and details of your best method. Describe your approach and parameters.

1. **Improve Temperature regression performance**

Report the RMS val and test errors and details of your best method. Describe your approach and parameters.

1. **Generate a train/test classification dataset in which Naive Bayes outperforms 1-NN and Logistic Regression**

Explain how you generated the data and report test performance for each method.

**Acknowledgments / Attribution**

List any outside sources for code or improvement ideas or “None”.