My Name: Qi Zhao

Major resources: A laptop, java libraries.

Programming Language: java.

Dataset details:

Pima Indians Diabetes Database:

I really don’t know how go depict these datas

Digit Recognizer:

I choose some examples to show what the data look like:



Training data: 80%

Validation data: 10%

Testing data: 10%

Algorithm description:

K-NN algorithms.

Feature scaling: I calculate a mediate value for each feature. mediate value = (max + min)/2

For each feature: the scaling value = original value/mediate value;

|  |  |  |  |
| --- | --- | --- | --- |
|  | (actual)Non-diabetes | (actual)diabetes | total |
| (predict)Non-diabetes | TN=57 | FN=18 | 75 |
| (predict)diabetes | FP=5 | TP=20 | 25 |
| total | 62 | 38 | 100 |

Confusion matrix

Accuracy: (TP + TN)/total = 77%;

Misclassification Rate: (FP + FN)/total = 23%;

True Positive Rate (sensitivity): TP/(TP + FN) = 53%;

False Positive Rate: FP/(TN + FP) = 8%;

Specificity: TN/(TN + FP) = 92%;

Precision: TP/(FP + TP) = 80%;

Prevalence: (FN + TP)/total = 38%;

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual  Predict | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | total |
| 0 | 398 | 0 | 2 | 1 | 0 | 0 | 2 | 0 | 2 | 1 | 406 |
| 1 | 0 | 440 | 10 | 2 | 3 | 1 | 0 | 7 | 6 | 1 | 470 |
| 2 | 0 | 1 | 403 | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 409 |
| 3 | 0 | 0 | 1 | 358 | 0 | 5 | 0 | 0 | 12 | 4 | 380 |
| 4 | 0 | 0 | 0 | 0 | 376 | 0 | 1 | 0 | 0 | 4 | 381 |
| 5 | 0 | 0 | 1 | 3 | 0 | 365 | 0 | 0 | 5 | 1 | 375 |
| 6 | 1 | 2 | 0 | 0 | 0 | 4 | 410 | 0 | 3 | 1 | 421 |
| 7 | 1 | 1 | 8 | 2 | 1 | 0 | 0 | 401 | 0 | 7 | 421 |
| 8 | 0 | 1 | 4 | 1 | 0 | 0 | 0 | 0 | 348 | 0 | 354 |
| 9 | 0 | 1 | 1 | 1 | 9 | 4 | 0 | 1 | 5 | 361 | 383 |
| total | 400 | 446 | 430 | 370 | 389 | 379 | 413 | 409 | 383 | 381 | 4000 |

Choose number 9 as the measure number.

For number 9, the accuracy of the classifier: 361/381 = 94.75%;

K = 1 : accuracy = 360/381 = 94.49%

K =2: accuracy = 340/381 = 89.24%

K =4: accuracy = 358/381 = 93.96%

K=8: accuracy = 361/381 = 94.75%

K=16: accuracy = 358/381 = 93.96%

K =32: accuracy = 354/381 = 92.91%

K=64: accuracy = 351/381 = 92.13%

Runtime

For Pima Data Set : K =5; run Time = 16ms;

For Mnist, k =10; run time = 2.6 min