**CAP6676 Assignment 1- Decision Tree Report**

**Kaiqiang Song**

**Oct.10 2017**

**[Dataset Selection]**

Source: UCI Datasets

Dataset: CoverType

URL:<http://archive.ics.uci.edu/ml/datasets/Covertype>

Including 7 categories, 581012 instances with 12 features (10 continuous, 2 discrete), no missing value.

**[k-fold-split]**

K = 5 in my experiment

**[Implementation of Decision Tree]**

Code can be found at:

<https://github.com/KaiQiangSong/CAP6676_KnowledgeRepresentation>

**[Performance]**

Overview:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Fold | Missing Rate | Error Rate | Accuarcy | Macro\_F1 |
| 1 | 0.00018 | 0.26395 | 0.73586 | 0.599642 |
| 2 | 0.00012 | 0.25504 | 0.74482 | 0.590535 |
| 3 | 0.00009 | 0.25420 | 0.74570 | 0.596133 |
| 4 | 0.00005 | 0.25496 | 0.74498 | 0.616383 |
| 5 | 0.00015 | 0.28257 | 0.71726 | 0.523546 |
| Average | 0.000118 | 0.262144 | 0.737724 | 0.5852478 |

(\*) Missing Rate is which decision tree can not decide which category it is due to the lack of training data (not enough to cover some corner case).

Detail:

Precision, Recall and F1\_Score of Each label and fold:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Fold-1 | Class-1 | Class-2 | Class-3 | Class-4 | Class-5 | Class-6 | Class-7 |
| Recall | 0.71829689 | 0.79709071 | 0.80891806 | 0.55681819 | 0.19463807 | 0.24120751 | 0.62610567 |
| Precision | 0.71919703 | 0.75499773 | 0.68672127 | 0.68055558 | 0.60000002 | 0.63749033 | 0.80411422 |
| F1\_Score | 0.71874666 | 0.77547342 | 0.74282777 | 0.61250001 | 0.29392713 | 0.34998935 | 0.70403224 |
| Fold-2 | Class-1 | Class-2 | Class-3 | Class-4 | Class-5 | Class-6 | Class-7 |
| Recall | 0.72071797 | 0.81138766 | 0.83321565 | 0.45303866 | 0.10337663 | 0.35652173 | 0.59161931 |
| Precision | 0.72666365 | 0.7634728 | 0.7184608 | 0.67213112 | 0.66112959 | 0.63369399 | 0.78683877 |
| F1\_Score | 0.72367859 | 0.78670132 | 0.77159488 | 0.5412541 | 0.17879605 | 0.45631605 | 0.67540538 |
| Fold-3 | Class-1 | Class-2 | Class-3 | Class-4 | Class-5 | Class-6 | Class-7 |
| Recall | 0.69857645 | 0.82812226 | 0.84091228 | 0.42490843 | 0.14959785 | 0.33055714 | 0.59381843 |
| Precision | 0.73791176 | 0.75648433 | 0.71473014 | 0.70948011 | 0.71355498 | 0.62249863 | 0.79869068 |
| F1\_Score | 0.71770561 | 0.79068393 | 0.77270377 | 0.53150058 | 0.24734041 | 0.43181393 | 0.6811837 |
| Fold-4 | Class-1 | Class-2 | Class-3 | Class-4 | Class-5 | Class-6 | Class-7 |
| Recall | 0.70089424 | 0.82269466 | 0.83620089 | 0.5324232 | 0.18622848 | 0.31586361 | 0.63045633 |
| Precision | 0.73604399 | 0.75678706 | 0.71268481 | 0.75362319 | 0.72857141 | 0.64499426 | 0.77194047 |
| F1\_Score | 0.71803916 | 0.78836578 | 0.76951796 | 0.62399995 | 0.29663479 | 0.424059 | 0.6940614 |
| Fold-5 | Class-1 | Class-2 | Class-3 | Class-4 | Class-5 | Class-6 | Class-7 |
| Recall | 0.71204519 | 0.77157098 | 0.85647446 | 0.37132353 | 0.04060385 | 0.14622913 | 0.61887938 |
| Precision | 0.69459248 | 0.74729055 | 0.64457333 | 0.70877194 | 0.70909089 | 0.67823768 | 0.71175617 |
| F1\_Score | 0.70321053 | 0.75923663 | 0.73556697 | 0.48733407 | 0.07680945 | 0.24058725 | 0.66207647 |

**[About Overfitting]**

Some overfitting appears because of the unbalanced data. But generally, they are strong results. Since the distance between training error and test error is not so large (<1% among ~25% error rate). See also <https://github.com/KaiQiangSong/CAP6676_KnowledgeRepresentation/blob/master/Running_Report>

to get a better understanding of that.