**Feature Selection – Program Flow**

|  |  |
| --- | --- |
| **File Description** | |
| RunFS.cpp | 處理程式主要流程，控制參數初始化設定、演算法、檔案讀寫等。透過呼叫FeatureSelect Class、Regression Class的函式執行相對應之演算法。main函式也位於此。 |
| FeatureSelection.h  FeatureSelection.cpp  FeatureSelectionAlgo.cpp  FeatureSelectionDisct.cpp | 實作FeatureSelection Class，先抓取CSV檔案資料，並選擇離散化方式，其後由外部呼叫函式執行MI演算法。 |
| Regression.h  Regression.cpp  LinearRegression.h  LinearRegression.h  Lars.h  Lars.cpp | 處理 Regression相關之正規化與演算法，其中回歸演算法使用外部函式庫mlpack、armadillo來實作(詳見README)，使用方式詳見官網。 |
| FEAST/\*.\* | MI演算法細節實作，使用FEAST與MIToolbox。  <http://mloss.org/software/view/386/>  <http://mloss.org/software/view/325/> |
| MatrixOp.h | 自定義的簡易vector矩陣操作。 |

**Program Flow:**

* 1. Parameter setting(input filename、algorithm parameter).
  2. FeatureSeletction initialization (by constructor). It will read file and store all data in “vector<vector<double> > featureData”. Since the initialization succeeded, we can use data access function to get the data.
  3. Exclude some undesired feature (e.g., dp\_filter). Any feature name containing the string pattern will be marked as “not used”, and later you call disct\_ew() or disct\_ew\_cycle() will get only the feature not excluded.

1. Discretize the input raw data. You can choose whether use each cycles’ max and min to discretize. Or just use global max and min. It has a significant effect on MI result.
   1. Choose the target column.(e.g. dp\_filter\_max)
   2. Discretize target column. Can choose disct\_col\_manual () or ew disct\_col\_ew\_cycle().
   3. Run MI algorithms. Save all results for later use.
   4. MI score and rank. You can go to FeatureSelection.cpp for detail implementation. Currently it is judged by every MI algorithm’s rank (equal-weight). So if one feature gets 100% in final score, it means that it is ranked as No.1 in all MI algorithms.
   5. Run regression algorithm. Save both ranking and coefficients.
   6. Regression score and rank. Currently sum up all coefficients as the ranking index, but exclude Least Square method, because its coefficients are too large.