

# Distance to containing variety: The UCI Adult dataset II

An experimental demonstration, using the UCI **adult** dataset, that the failures of the independent stream evaluator for binary classifiers can be used to find nearly independent trios on unlabeled data.

We read some common code for the three datasets

```
In[*]:= nbDir = NotebookDirectory[];  
SetDirectory[nbDir];  
NotebookEvaluate[FileNameJoin[{nbDir, "CommonCode.nb"}]]
```

---

## Getting the data and taking a look at the available features

```
In[*]:= {tsvHeader, benchmarkData} = ImportPennMLBenchmarksDataset["adult"];  
  
In[*]:= tsvHeader  
Out[*]:= {age, workclass, fnlwgt, education, education-num,  
marital-status, occupation, relationship, race, sex, capital-gain,  
capital-loss, hours-per-week, native-country, target}  
  
In[*]:= uciAdultFeatures  
Out[*]:= {{1, Numerical}, {2, Nominal}, {3, Numerical}, {4, Nominal}, {5, Nominal},  
{6, Nominal}, {7, Nominal}, {8, Nominal}, {9, Nominal}, {10, Nominal},  
{11, Numerical}, {12, Numerical}, {13, Numerical}, {14, Nominal}}  
  
In[*]:= Length /@ benchmarkData  
Out[*]:= <| 1 → 37 155, 0 → 11 687 |>
```

## Experiments

```

In[*]:= manyFeaturePartitions = Table[RandomSample[uciAdultFeatures] //
    Partition[#, UpTo@3] & //
    Take[#, 3] &, {1000}] //
    Map[Sort, #, {2}] & // Sort /@# & // DeleteDuplicates //
    Partition[#, 20] & //
    Take[#, 3] &;

For[i = 1, i < 4, i++,
  nTrain = 200;
  globalTrainSize = 600;
  nClassifiers = 3;
  classifierTypes = Table[{"LogisticRegression", "L1Regularization" → 0.1,
    "L2Regularization" → 0.5, "OptimizationMethod" → "Newton"}, {nClassifiers}];
  (* We separate a held-out test set that is never seen in any of the training
    runs below *)
  aTrainIndices = RandomSample[Range@Length@benchmarkData[0], globalTrainSize];
  bTrainIndices = RandomSample[Range@Length@benchmarkData[1], globalTrainSize];
  globalTrainTestSplit =
    MakeTrainTestSplit[benchmarkData, aTrainIndices, bTrainIndices];
  globalTrainData = SelectTrainDataFromSplit[globalTrainTestSplit];
  globalH0Data = SelectTestDataFromSplit[globalTrainTestSplit];
  nTest = 2000;
  runs = Table[
    {classifiersFeatures,
      Table[
        (* Select random, disjoint partitions of the training data,
          grouped by classifier *)
        trainIndices = Transpose@{
          (* The disjoint partition of a subset of the alpha training data *)
          ClassifiersSampleIndices[globalTrainData[0], nTrain],
          (* The disjoint partition of a subset of the beta training data *)
          ClassifiersSampleIndices[globalTrainData[1], nTrain]};
        (* Prepare the training data for the current feature partitions *)
        trainClassifiersData = Table[Map[#[[First@Transpose@features]] &,
          globalTrainData, {2}], {features, classifiersFeatures}];
        classifiers = TrainClassifiersDisjoint[
          trainClassifiersData, classifierTypes, trainIndices,
          Map[(Last@Transpose@#) &, classifiersFeatures], "TrainingSpeed"];

        (* Now evaluate *)
        evalIndices = {

```

```

(* The disjoint partition of a subset of the alpha training data *)
SampleIndices[globalH0Data[0], nTest],
(* The disjoint partition of a subset of the beta training data *)
SampleIndices[globalH0Data[1], nTest]]];
testClassifiersData = Table[Map[#[[First@Transpose@features]] &,
  SelectFromData[globalH0Data, First@evalIndices, Last@evalIndices],
  {2}], {features, classifiersFeatures}];
vcbl = LabelCounts[classifiers, testClassifiersData];
Which[
  Not@HasAllDecisionFrequenciesQ@vcbl,
  Nothing,
  Not@IndependentEvaluationIdealIsNonEmptyQ@vcbl,
  Nothing,
  sols = AlgebraicallyEvaluateClassifiers[vcbl];
  Not@HasOnlyRealNumbersQ@sols,
  Nothing,
  Not@InsideUnitCubeQ@sols,
  Nothing,
  True,
  vcbl],
{10}]],
{classifiersFeatures, manyFeaturePartitions[[i]]}];
runs >> "./adult-ni-" ~~ ToString[i] ~~ ".m"
]

```