

| This data was extracted from the census bureau database found at
| <http://www.census.gov/ftp/pub/DES/www/welcome.html>
| Donor: Ronny Kohavi and Barry Becker,
| Data Mining and Visualization
| Silicon Graphics.
| e-mail: ronnyk@sgi.com for questions.
| Split into train-test using MLC++ GenCVFiles (2/3, 1/3 random).
| 48842 instances, mix of continuous and discrete (train=32561, test=16281)
| 45222 if instances with unknown values are removed (train=30162, test=15060)
| Duplicate or conflicting instances : 6
| Class probabilities for adult.all file
| Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)
| Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)
|
| Extraction was done by Barry Becker from the 1994 Census database. A set of
| reasonably clean records was extracted using the following conditions:
| ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))
|
| Prediction task is to determine whether a person makes over 50K
| a year.
|
| First cited in:
| @inproceedings{kohavi-nbtree,
| author={Ron Kohavi},
| title={Scaling Up the Accuracy of Naive-Bayes Classifiers: a
| Decision-Tree Hybrid},
| booktitle={Proceedings of the Second International Conference on
| Knowledge Discovery and Data Mining},
| year = 1996,
| pages={to appear}}
|
| Error Accuracy reported as follows, after removal of unknowns from

| train/test sets):

| C4.5 : 84.46+-0.30

| Naive-Bayes: 83.88+-0.30

| NBTree : 85.90+-0.28

|

|

| Following algorithms were later run with the following error rates,

| all after removal of unknowns and using the original train/test split.

| All these numbers are straight runs using MLC++ with default values.

|

| Algorithm Error

| -- -----

| 1 C4.5 15.54

| 2 C4.5-auto 14.46

| 3 C4.5 rules 14.94

| 4 Voted ID3 (0.6) 15.64

| 5 Voted ID3 (0.8) 16.47

| 6 T2 16.84

| 7 1R 19.54

| 8 NBTree 14.10

| 9 CN2 16.00

| 10 HOODG 14.82

| 11 FSS Naive Bayes 14.05

| 12 IDTM (Decision table) 14.46

| 13 Naive-Bayes 16.12

| 14 Nearest-neighbor (1) 21.42

| 15 Nearest-neighbor (3) 20.35

| 16 OC1 15.04

| 17 Pebls Crashed. Unknown why (bounds WERE increased)

|

| Conversion of original data as follows:

| 1. Discretized agrossincome into two ranges with threshold 50,000.

| 2. Convert U.S. to US to avoid periods.

| 3. Convert Unknown to "?"

| 4. Run MLC++ GenCVFiles to generate data,test.

|

| Description of fnlwgt (final weight)

|

| The weights on the CPS files are controlled to independent estimates of the

| civilian noninstitutional population of the US. These are prepared monthly

| for us by Population Division here at the Census Bureau. We use 3 sets of

| controls.

| These are:

| 1. A single cell estimate of the population 16+ for each state.

| 2. Controls for Hispanic Origin by age and sex.

| 3. Controls by Race, age and sex.

|

| We use all three sets of controls in our weighting program and "rake" through

| them 6 times so that by the end we come back to all the controls we used.

|

| The term estimate refers to population totals derived from CPS by creating

| "weighted tallies" of any specified socio-economic characteristics of the

| population.

|

| People with similar demographic characteristics should have

| similar weights. There is one important caveat to remember

| about this statement. That is that since the CPS sample is

| actually a collection of 51 state samples, each with its own

| probability of selection, the statement only applies within

| state.

>50K, <=50K.

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.