Stat 882 Project

Bob Patterson
Hui Chen

1. Description

The purpose of the project is to classify an adult dataset by salary. The dataset comes from the 1994 Census database.

2. Number of Classes

Two classes: Class 1: Salary ≤ $50,000
Class 2: Salary > $50,000.

3. Number of Instances

- 48842 instances, a mix of continuous and discrete (train=32561, test=16281)
- 45222 if instances with at least one unknown (missing) value are removed (train=30162, test=15060)
- We intend to make use of all 48842 instances

4. Number of Attributes

There are fourteen total attributes. Six are continuous and eight are nominal attributes.

5. Attribute Information:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. 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.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. 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.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. 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, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

6. Missing Attribute Values:

7% of the instances have missing values.

7. Past Algorithm Work

	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)

8. Previous Class Distribution:

Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)
Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)