Prediction of Income Level based on Census Data

Objective:

To use machine learning techniques to carry out prediction of income level of a citizen based on parameters like age, working class, education, marital status, occupation, relationship, race, gender, capital loss, capital gain, number of working hours per week and nationality.

Description of data set:

Task: Classification
Number of instances: 48842
Number of training examples: 32561
Number of test examples: 16281
Classification: Income below $50000 and income above $50000.

Attributes:

1) Age: continuous.
2) Working class: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3) 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.
4) Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
5) 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.
6) Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
7) Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
8) Sex: Female, Male.
9) Capital-gain: continuous.
10) Capital-loss: continuous.
11) Hours-per-week: continuous.
12) 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.

Training and encoding of features:

Two different methods were used for training: logistic regression and neural network training. Continuous variables were normalised to make their values lie between 0 and 1. The ratio of difference between the value and the minimum of all the values to the difference between maximum and minimum value is taken as the normalised value.

$$x_{norm} = (x-x_{min})/(x_{max}-x_{min})$$

Discrete variables are encoded by converting them to several binary variables. For example, ‘marital status’ is a variable that may contain one of seven different states. Each of these states was converted into a binary variable which takes value 1 or 0 according to the state of the variable.

This gives rise to a total input vector of 102 elements.

Logistic Regression:

Functions and scripts used:

1) logisticreg.m – Loads the data and runs logistic regression on it. Calls all the other functions in order to carry out this task. First the data is loaded and all the categories in the categorical variables are identified and sorted in alphabetical order. Then the data is encoded into a feature vector and logistic regression is carried out to get the classification parameters. After this the parameter predictions are tested for accuracy against the training and test set .
2) features.m – Encodes the categorical variables into binary variables and the continuous variables into their normalised forms.
3) gradientDescent.m – Runs gradient descent algorithm on the data to compute the logistic regression parameters.
4) cost.m – Calculates the cost of parameters with respect to the training set.
5) sigmoid.m – Calculates the sigmoid function of the input value.
6) predict.m – Predicts the result using the parameters and data input points.
7) crossval.m – Tests the quality of parameters against a test data set.

Neural network training:

1) nnlearning.m – Loads data and runs neural network learning algorithm. Calls all the other functions in order to carry out this task. First the data is loaded and all the categories in the categorical variables are identified and sorted in alphabetical order. Then the data is encoded into a feature vector and neural network learning is carried out to get the classification parameters. After this the parameter predictions are tested for accuracy against the training and test set . One hidden layer of 25 members are used for this neural network.
2) features.m – Encodes the categorical variables into binary variables and the continuous variables into their normalised forms.
3) deletefeatures.m – Deletes features as specified by the user.
4) randparams.m – Randomly initializes parameters before starting the optimisation.
5) fmincg – Function that optimises parameters.
6) costfunction.m – Function that compute cost and gradient by forward propagation and backtracking.
7) predict.m – Predicts the result using the parameters and data input points.
8) crossval.m – Tests the quality of parameters against a test data set.

Results:

1) Logistic regression, all features:

	Accuracy	Precision	Recall	F1score
Training set	83.2%	64.5%	66.9%	65.7%
Test set	82.9%	63.4%	65.9%	64.6%

2) Neural network, all features:

	Accuracy	Precision	Recall	F1score
Training set	85.4%	67.6%	75.2%	71.2%

Test set	84.4%	65.3%	72.1%	68.5%

3) Income prediction using marital status

Model: Neural networks
Training quality:

	Accuracy	F1score
Training set	71.0%	58.7%
Test set	71.3%	58.5%

Relationship between marital status and income:

Marital status	Income	Probability of income >50k
Divorced	<50K	0.11
Married-AF-spouse	>50K	0.46
Married-civ-spouse	>50K	0.45
Married-spouse-absent	<50K	0.08
Never-married	<50K	0.05
Separated	<50K	0.06
Widowed	<50K	0.09

4) Income prediction using education

Model: Neural networks
Training quality:

	Accuracy	F1score
Training set	75.2%	50.0%
Test set	75.0%	48.5%

Relationship between marital status and income:

Education	Income	Probability of income greater than 50K
10th	<50K	0.07
11th	<50K	0.05
12th	<50K	0.08
1st-4th	<50K	0.05
5th-6th	<50K	0.05
7th-8th	<50K	0.06
9th	<50K	0.05
Assoc-acdm	<50K	0.25
Assoc-voc	<50K	0.26
Bachelors	>50K	0.41
Doctorate	>50K	0.74
HS-grad	<50K	0.16
Masters	>50K	0.56
Preschool	<50K	0.08
Prof-school	>50K	0.73
Some-college	<50K	0.19

Note: The threshold probability for the classifier has been set at below 0.5 to ensure better results.

Conclusions:

1) Married people have the highest probability of having a good income. There is a probability of 45% that a married man/woman earns above 50000 dollars.
2) Divorced, separated, widowed and never married people are less than 10% probable to have an income of above 50000 dollars.
3) A person holding a Bachelors degree has a good pay with probability 40% while a Master degree holder has chances of 56%.
4) Doctorate degree holders and professional school graduates have a chance of above 70% to earn above 50000 dollars.