Machine Learning Engineer Nanodegree

Using Supervised Classification Algorithms to Predict Income from Census Data

Fabiano Shoji Yoschitaki
June 28th, 2018

Proposal

Domain Background

The domain background of this project is about predicting money income from the US Census dataset, which is an extraction done by Barry Becker from the 1994 Census database [1]. This dataset was donated to the University of California Irvine's Machine Learning Repository in 1996. The Census dataset is composed of 48842 instances, 14 attributes (or features) and the target: whether or not a person's annual income exceeds 50k.

The task of this problem is to determine whether someone's income is over 50k a year based on the 14 variables. Since the prediction output may be true or false, this is a binary classification problem which can be solved using supervised machine learning classification techniques [2][3].

The personal reason to work on this domain background comes from the fact that I've worked on a census tool which could not provide much useful statistical information about the local economy because there was a lack of data related to people's wages and earnings. So I believe that applying machine learning techniques could have helped us to take the data we had as the training set and use the predicted output to fill the gaps in the data.

Problem Statement

Given the US Census dataset, a supervised binary classification model has to be created and trained using a subset of the entire dataset holding all the 15 features related to people (including the target output: whether or not the person makes over 50k a year) with the objective of predicting unseen people's wage based on the same features.

Datasets and Inputs

The dataset chosen for this project is an extraction done by Berry Becker from the 1994 US Census database. It was obtained by exploring the University of Carolina Irvine's Machine Learning Repository [4]. The dataset is in CSV format and contains 48842 instances with 15 columns each. The last column is the target: whether or not the person makes over 50k a year. The probability for the label '>50k' is 23.93% and for '<=50k' is 76.07%. The description of the columns follow:

- **age**: the age of the person (continuous).
- **workclass**: the workclass of the person (categorical). Possible values: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: the final weight of the person (continuous).

- **education**: the education level of the person (categorical). Possible values: 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**: the education level of the person (continuous).
- **marital-status**: the marital status of the person (categorical). Possible values: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: the job occupation of the person (categorical). Possible values: 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**: the relationship status of the person (categorical). Possible values: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: the race of the person (categorical). Possible values: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: the sex of the person (categorical). Possible values: Female, Male.
- **capital-gain**: the income from investment sources, apart from wages or salary (continuous).
- **capital-loss**: the losses from investment sources, apart from wages or salary (continuous).
- **hours-per-week**: number of working hours per week (continuous).
- **native-country**: the native country of the person (categorical). Possible values: 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.

And finally the target label:

- **income**: whether or not the person makes more than 50k a year (categorical). Possible values: >50k, <=50k.

Solution Statement

The main goal of this project is to apply different machine learning techniques, including supervised classification algorithms such as Gaussian Naive Bayes, Linear Discriminant Analysis, K-Nearest Neighbors, Logistic Regression, Decision Tree, Extreme Gradient Boosting, SVM and Random Forest, to predict from the US Census dataset whether or not money income exceeds 50k a year (so it is a binary classification task). After applying the chosen supervised algorithms to the dataset, their prediction scores will be evaluated by accuracy metric and the best one (maybe more than one if we have similar high scores) will be selected so that its hyper-parameters are tuned using Grid Search technique and eventually apply the newly tuned model on the same training and testing data to compare before and after scores.

Benchmark Model

The US Census dataset is a supervised learning classification problem, so I propose to apply - without tuning - all the supervised classification algorithms described in the Solution Statement section and choose the one with the best accuracy score as the benchmark model for this project. Then, after we tune the hyper-parameters of the model, it is expected that the newly tuned model may overcome the benchmark model accuracy score.

Evaluation Metrics

The evaluation metric that will be used in this project is accuracy. It is an appropriate metric for supervised classification problems. Accuracy is a metric that takes considers both correct and incorrect classifications in its formula:

(TP + TN)/(TP + TN + FP + FN)

Where:

- **TP**: True Positive, in our case a person that makes >50k a year and is correctly classified as >50k a year.
- **TN**: True Negative, in our case a person that makes <=50k a year and is correctly classified as <=50k a year.
- **FP**: False Positive, in our case a person that makes <=50k a year and is incorrectly classified as >50k a year.
- **FN**: False Negative, in our case a person that makes >50k a year and is incorrectly classified as <=50k a year.

Project Design

The project is designed in the following steps:

- **Data and Library Loading:** the first step is to load the US Census dataset in the CSV format from the UCI's Machine Learning Repository and all the libraries needed for the project.
- **Data Exploration:** in this step, we'll do some tasks like: visualize the data, print some samples, check its dimensions, check the most relevant features, show its statistical summary.
- **Data Preparation:** after exploring the data, pre-processing tasks will be done: data cleaning, remove null values, convert categorical features into dummy/indicator variables and split the data into training and testing datasets.
- **Model Selection:** with the prepared data, various supervised classification algorithms will be experimented in order to find compare their results and choose the best one (taking into account the accuracy score) for model tuning.
- **Model Tuning:** after we choose the best model, grid search cross validation will be applied with the objective to tune the hyper-parameters of the model.
- **Final Evaluation:** in this step, the accuracy score of the tuned model will be evaluated by applying it to the testing dataset.

References

[1] https://archive.ics.uci.edu/ml/support/census+income

[2] Ke Wang and Shiyu Zhou and Ada Wai-Chee Fu and Jeffrey Xu Yu. Mining Changes of Classification by Correspondence Tracing. SDM. 2003.

[3] Bart Hamers and J. A. K Suykens. Coupled Transductive Ensemble Learning of Kernel Models. Bart De Moor. 2003.

[4] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].