Capstone Project Proposal

Emma Weng

August 23, 2020

Domain Background

This project is derived from cancer classification and prediction by gene expression at the molecular level. With a background in molecular biology and research experience in cancer treatment study, topics related to genetic engineering and biotechnology have always attracted my attention.

The original dataset of this project comes from a research study by Professor Golub. It described a generic method for automatically determining the type of cancer between acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML). The published paper showed the possibility of cancer classification based only on the gene expressions without relevant biological knowledge.

Acute lymphocytic leukemia (ALL) is a rare cancer in adults, but it is the most common form of leukemia in children. Acute myeloid leukemia (AML) is one of the most common leukemias in adults. They are developed from different types of white blood cells and involve different treatment strategies. Better classification can provide useful information for medical staff to identify cancer.

Problem Statement

The goal of the project is to find a good classification method to classify leukemia patients into acute lymphocytic leukemia (ALL) and acute myeloid leukemia (AML).

The input data format is a table of 7129 gene expressions from 72 patients with ALL or AML. The expression of each gene is a numeric value measured from DNA microarray.

The expected output of the classification model will be a single categorical value, showing the cancer type ALL or AML of each patient.

Datasets and Inputs

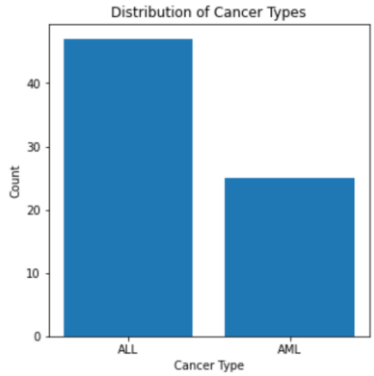
The dataset is provided by Kaggle (https://www.kaggle.com/crawford/gene-expression/download). It contains three files which has been converted to comma separated value files so that are great for classification problem.

1. actual.csv: each patients cancer type
2. data_set_ALL_AML_independent.csv: test data
3. data_set_ALL_AML_train.csv: training data

Patients file:

The first column is the patient id number; the second column is the cancer type. The total number of patients is 72; the cancer type is ALL or AML. The distribution of cancer types is 47 ALL patients and 25 AML patients.

```
cancer
ALL    47
AML    25
dtype: int64
```




```
labels.describe()
```

	cancer
count	72
unique	2

Training data/test data file:

The table has 7129 rows representing 7129 different genes. The first column is "gene description" and the second column is "gene accession number". The gene accession number has a unique value, but the gene description does not.

The rest of columns represent the gene expression value of each patient. Each patient has two columns showing the value and category of each gene.

The training data contains patients 1 to 38. The test data includes patients 39 to 72.

```
train_data.loc[:, ['Gene Description', 'Gene Accession Number']].describe()
```

	Gene Description	Gene Accession Number
count	7129	7129
unique	6627	7129

Reference:

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression

Science 286:531-537. (1999). Published: 1999.10.14

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander

Solution Statement

The project will build a machine learning model to classify cancer types based on the gene expression of each patient.

Since each patient has labeled his own cancer type, each data point is composed of gene expression values as input and cancer type as output. This is a discrete classification problem and supervised learning will be the solution. In the project, the model algorithm Naive Bayes, Logistic Regression, Support Vector Machine, XGboost and Neural Network will be used for comparison.

In addition, each patient has 7129 different gene values within dataset, which means 7129 dimensions of the input data. For higher-dimensional data like this, classification algorithms will have difficulty distinguishing noise from important features. Therefore, PCA analysis will be the first step to reduce the number of features and select the top principal component to use in a classify model.

In this dataset, the ratio of the two types of cancer is about 2:1, which is not a very high imbalance. However, it is still worth trying to see if SageMaker's built-in algorithm, linear learner with positive_example_weight_mult parameter and Hyperparameter Tuner can improve the model.

Benchmark Model

The dataset has labeled the cancer type of each patient. In the test data, there are 20 ALL patients and 14 AML patients. If the binary classification predicts all patients have ALL type cancer, the accuracy rate is 58.8%. This is a simple benchmark for modeling.

In addition, an unsupervised clustering approach K-Means algorithm will be tried as a benchmark model. Supervised learning models should have better performance than unsupervised learning models.

Evaluation Metrics

The supervised learning will compare the actual output and predicted output of the model and use an error matrix with true positives, false positives, true negatives, false negatives to quantify its performance. The accuracy, precision, and recall rate will be calculated to evaluate the model.

Project Design

The workflow of this project:

1. Data Loading and Exploration
    - Data Loading and Exploratory Data Analysis
    - Data Cleaning and Pre-processing
2. Feature Engineering and Data Transformation
    - Normalization: use sklearn preprocessing to standardize the scale of data
    - Dimensionality reduction: use Sagemaker PCA model to reduce the number of input features
3. Defining and Training Model
    - Define and Create Estimator:

        a. Benchmark Model
        b. Sagemaker built-in binary classification
        c. Custom PyTorch Neural Network Classifier Train the estimator
    - Modeling tuning
4. Deploy the trained model
5. Evaluate the Performance of model: use error matrix to see the accuracy
6. Clean up Resources