POLITECNICO DI TORINO

THIRD REPORT ON THE ICT for HEALTH LABORATORY

A.Y. 2018-2019

Professor:

Visintin Monica

Student:

Bellone Lorenzo

Summary

Introduction

The kidney is an essential organ that provides functions without which a human being could not survive. The main functions of the kidney are:

- the balance of the plasma volume;
- the plasma filtering, in order to get rid of waste products from the organism;
- the production of hormones, such as Vitamin D and erythropoietin.

A person suffers of a chronic kidney disease (CKD) if his kidney is no longer able to guarantee some of these functions. CKDs are still not curable, so the aim of the main treatments is to slow the progression of the illness.

The most common solutions are dialysis or kidney transplants, but these kinds of procedures are very uncomfortable and expensive. It frequently happens that a patient refuses these treatments with a consequent risk of decease caused by the CKD.

One of the principal methods to detect a CKD is the measure of the Glomerular Filtration Rate (GFR), which is an estimation of how much blood the kidney is able to filter in a minute (ml/min). This measure can be done through the Creatinine Clearance, which means the evaluation of the amount of Creatinine filtered by the kidney in a given time interval.

Another sign of a possible kidney disease is the presence of an abnormal quantity of albumin in the urine (known as Albuminuria). This symptom can be simply evaluated performing the urine test. It has been demonstrated that GFR and Albuminuria are correlated in the diagnosis and the staging of a chronical kidney disease.

During the first stages of a CKD the patient is asymptomatic, so the illness can not be treated in time to avoid a premature dialysis, a kidney transplant or, even worse, the death of the patient. Therefore, it is very important to make an early prediction by exploiting other information such as the blood pressure, the age or some features simply evaluated through a blood test.

The medical doctor can follow a decision tree in order to make the right diagnosis. In a decision tree each node represents a feature (GFR, Albuminuria, blood pressure etc…) and each link represents a decision based on the condition of the patient related to the specific feature. The outcome of the decision tree gives the diagnosis, which in this case consists of two classes: the first one if the patient has a CKD and the second one if the patient is healthy. The doctor needs to take the smallest number of decisions as possible, this means that the decision tree must be developed in an optimal way through the application of the **Information Theory**.

Information Theory

Information Theory [1] is the basis of telecommunication systems. The two main concepts of these discipline are the **quantity of information** and the **entropy**.

Given an event whose outcome is a discrete random variable x_i with probability $p_{i,}$ the quantity of information of $x = x_i$ is defined as:

$$I(x_i) = log_2 \frac{1}{p_i}$$

(1.1)

The entropy of the event x can be evaluated as follows:

$$\mathbb{H}(x) = \sum_{i=1}^{N} p_i \, log_2 \frac{1}{p_i}$$

(1.2)

In the equation 1.1 it can be seen that an outcome with a low probability carries more information with respect to another one with a higher probability. The entropy describes an average of the information quantity, that is to say: the lower the entropy related to an event is, the higher its predictability will be.

These main concepts are exploited in a classification algorithm, named **C4.5**, which allows to build an efficient decision tree. A dataset of N already classified samples is given, each sample has D features. C4.5 chooses the feature with the higher **mutual information** in order to take the smallest number of decisions. Mutual information is the difference between the entropy of the class and the entropy of the class given the feature:

$$I(Class; Feature) = \mathbb{H}(Class) - \mathbb{H}(Class|Feature)$$

(1.3)

In other words, mutual information returns how much the knowledge of an attribute of the training set increases the predictability of the class. Starting from the feature that ensures the maximum mutual information, C4.5 splits the dataset into subsets according to all the possible values that the feature can take. The algorithm is repeated for each single subset until a stop condition occurs; it is stopped when all the attributes have been exploited or when the class can be totally predicted, which means that the entropy of the class for the specific subset is zero.

The Dataset

The aim of the third lab is to build an optimal decision tree, given a dataset of features related to different patients, in order to establish whether a patient has a CKD.

The dataset is provided by the University of California Irvine on the website: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. The file contains twenty-four features plus the class (CKD or NOTCKD), involving four hundred patients. These features are: Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia.

The first part of the lab concerns an initial cleaning of the data, since each feature must be written in the same way for every patient. This cleaning performance is needed in order to process the data, not only in this specific lab but in every data analysis application. Moreover, it is important to detect the missing attributes (identified with a question mark "?"). If a sample has too many invalid features, more than five in this case, it can not be used for the classification since it brings not enough information; thus, it must be removed from the dataset.

The Regression of Missing Data

The aim of the next task is to regress the remaining missing features through a linear regression using the Ridge Regression [2]. The data train is represented by a matrix that contains only the patients with twenty-five valid features, including the class. Instead, the data test is the entire dataset with at least twenty valid features for each patient, minus the column that contains the invalid attribute. The optimum weight vector is evaluated with the training set as:

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} \cdot \boldsymbol{X}^T\boldsymbol{y} \tag{2.1}$$

In the relation 2.1: **X** is the matrix that contains only valid data minus the column to be regressed, **y** is the missing column of **X** and λ is the Langrangian Multiplier (set equal to 10), which is added to the diagonal entries of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ in order to avoid numerical problems due to the matrix inversion.

After that, the test vector is evaluated from the product between the test set and the weight vector. Thus, the missing value is replaced with the corresponding entry of the test vector.

If the missing values are more than one for a specific patient, it is necessary to remove all the columns to be regressed from the matrix **X** and from the test set. Then, the regression can be performed column by column by exploiting always the same training and test set.

As usual, the data have been normalized before the regression, in order to avoid numerical problems. Since there are both numerical and categorical features in the dataset, it has been necessary to convert the second ones into numbers. After the regression, the data must be denormalized to let the medical doctor understand the output information.

The Decision Tree

A matrix with 337 number of patients without invalid data is given after the cleaning of the dataset and the regression of the missing features.

The C4.5 algorithm has been applied to this dataset through the Python's library "*Scikit Learn*", by using the method "*tree.DecisionTreeClassifier( )*". The decision tree generated is shown in figure 1.

DECISION TREE

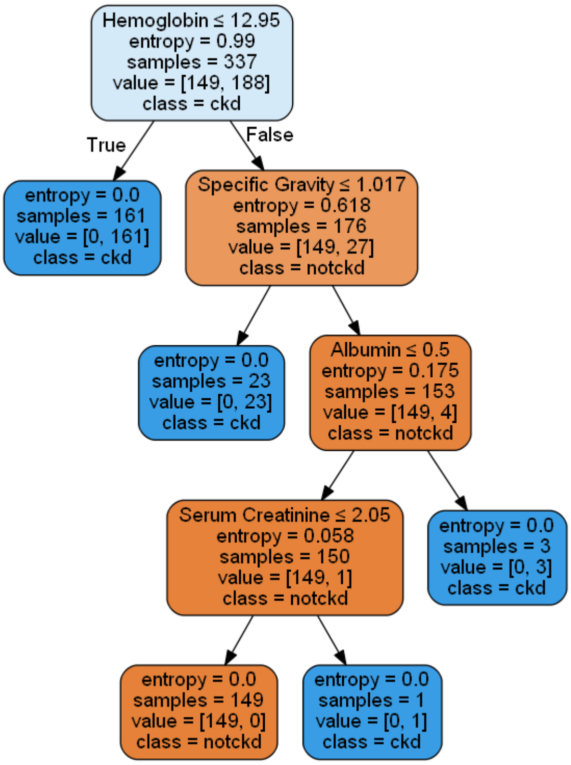

Figure 1 - The decision tree generated from the cleaned dataset of patients

Conclusions

In figure 1 it is shown the working principles of the C4.5. In each internal node four aspects are represented: the threshold of the feature on which the decision has to be made, the entropy of the class given the knowledge of all the previous features, the number of samples that have to be classified and the chosen class. Furthermore, near the first two arrows it is specified the decision rule, which consists in a true-false decision based on the identified threshold.

At each step it is considered the feature that has the higher mutual information. In this case, four decisions at most are enough to classify the entire dataset. The very first decision is made on the haemoglobin levels. The production of erythropoietin is compromised when the kidney is damaged, hence the blood has fewer red blood cells and the level of haemoglobin decreases.

Then, other three features are involved: Specific Gravity, Albumin and Serum Creatinine. Their importance in the decision of the diagnosis is progressively smaller. The importance of each feature can be computed on Python through the attribute "*feature_importance_*".

It seems clear that this method works properly on the given dataset. However, the number of patients is not enough in order to extend this decision tree to a whole different set of patients.

References

[1] Wikipedia Contributors. (2018, December 7). *Information Theory*. Retrieved December 11, 2018, from Wikipedia, The Free Encyclopedia.: https://en.wikipedia.org/w/index.php?title=Information_theory&oldid=872560886

[2] Wikipedia Contributors. (2018, September 11). *Tikhonov regularization*. Retrieved December 10, 2018, from Wikipedia, The Free Encyclopedia.: https://en.wikipedia.org/w/index.php?title=Tikhonov_regularization&oldid=859557845