Strasbourg University / French-Azerbaijani University

Artificial Intelligence

Computer Science track - Year 3

Lab : Predicting the risk of heart disease in patients

Instructions – You will work in pairs and submit your work (source code and lab report) on Moodle. Not only you must include the answers to the questions asked in this lab sheet in your report, but your report must also be analytic : you are expected to have a critical view of your work to show you grasp the problems you can face while working on a simple machine learning problem. Using machine learning framework and libraries is prohibited : every functionality must be implemented from scratch. You should, however, re-use what you implemented during the previous labs.
The deadline for submitting your work is XXX.

1 The "Cleveland Heart Disease" dataset

You must train models on the "Cleveland Heart Disease" dataset in order to **predict the risk of heart disease in patients**. This dataset contains 303 instances and each instance is described by 14 attributes. Each of these attributes are phylosiological measurements. They are presented in table 1.

Attributs	**Description**
age	The person's age in years
sex	The person's sex (1 = male, 0 = female)
chest_pain_type	The chest pain experienced (Value 1 : typical angina, Value 2 : atypical angina, Value 3 : non-anginal pain, Value 4 : asymptomatic)
resting_blood_pressure	The person's resting blood pressure (mm Hg on admission to the hospital)
cholesterol	The person's cholesterol measurement in mg/dl
fasting_blood_sugar	The person's fasting blood sugar (> 120 mg/dl, 1 = true ; 0 = false)
rest_ecg	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
max_hear_rate_achieved	The person's maximum heart rate achieved
exercise_induced_angina	Exercise induced angina (1 = yes ; 0 = no)
st_depression	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot
st_slope	the slope of the peak exercise ST segment (Value 1 : upsloping, Value 2 : flat, Value 3 : downsloping)
num_major_blood_vessels	The number of major vessels (0-3)
thalassemia	A blood disorder called thalassemia (3 = normal ; 6 = fixed defect ; 7 = reversable defect)

Attributs	Description
target	Diagnosis of a heart disease (0 = no, 1 = yes)

TABLE *1 – Attributes of the dataset*

Answer the following questions by studying table 1 :

1. What is the name of the attribute we want to predict ?
2. Is that a binary or a multi-class classification ?
3. Which attributes are categorical ?
4. How can we encode categorical attributes ?

Non-categorical attributes are numeric. We know that some optimization algorithms (such as gradient descent) are sensitive to numerical attributes. A quick look in the dataset shows that these attributes have various magnitudes : in order for a gradient descent to be effective, we will have to normalize these attributes.

Attribute normalization

1. Which normalization method will be best adapted so we can preserve the variance of the dataset ?
2. Will you normalize the data *before* or *after* splitting the dataset in training/testing datasets ?
3. Implement a method/function to normalize the attribute using the adequate normalization method.

2 Predictions using a Multi-layer perceptron

As a first step, we will use MLP with 1 hidden layer containing 5 units.

Questions

1. What are the dimensions of the matrices you will use to represent your model (inputs, parameters and outputs) ? How will you integrate the concept of mini-batch training ?
2. How should you check whether or not you should keep training your model ?
3. Draw your network (you can use the following online tool : `http://alexlenail.me/NN-SVG/index.html`)

You will set the mini-batch size to 4 and use a learning rate of 0.01. To stop the training process, you will check if the error on the test set increases on average during 10 consecutive training epochs.

2.1 Model evaluation

In order to evaluate your final model, you will have to compute the following metrics :

1. the precision of your model
2. the accuracy of your model
3. the sensitivity of your model
4. the specificity of your model

Question In the case of predicting the risk of heart disease in patients, would you prefer that your model is sensitive or specific ?

3 Predictions using a Decision Tree

Implement a Decision Tree to predict the risk of heart disease using the Cleveland Heart Disease dataset.

For the technical details of this model, you are on your own (we just instruct you to not go beyond a depth of 4 in your tree).

You are expected to give a thorough description of your methodology (regarding both how you handle you data and how your model performs).

To train your Decision Tree, we strongly advise you keep the same splitting proportion as for MLP training.

4 Comparing MLP and DT

Using the evaluation metrics you defined on both your MLP and your DT, explain which model you would rather use to predict the risk of heart disease in patients.