Machine Learning in Data Science Project Week 5

Eva Aßmann, Paul Vogler

November 2019

Introduction

Decision Tree Classifiers (DTs) are a simple, non-parametric supervised learning method that can be used for both classification and regression. DTs can fit complex and linearly unseparable data sets. They learn decision rules inferred from the data features in order to predict the value of a target variable. In the learning process, a DT splits the data iteratively based on specific feature values by drawing axis-parallel decision boundaries. Each decision boundary separates data instances of different target classes into decision regions. A DT grows until every decision region contains only the instances of one class. Since DTs are prone for overfitting, additional stopping criteria for data splitting are recommended. Generally, DTs are very sensitive to small data variations and data rotation, since they make only axis parallel boundaries. Using dimensionality reduction techniques can help limiting the problem.
A logistic model is used to predict the probabilities of certain target classes for each instance in a data set. Linear Regression (LR) is a linear classification model that estimates the parameters of a logistic function to predict the value of a binary dependent variable for a data instance. By introducing a cutoff value into the logistic model, a binary classifier is build that classifies inputs based on showing a probability greater than or below the cutoff, the logistic model becomes logistic regression.
The Vapnik-Chervonenkis (VP) dimension is a measure for the complexity for the space of functions that can be learned by a classification algorithm. It is defined by the maximum number of points the algorithm shatters, i.e., can divide distinctively into two groups. A classification model shatters a set of data points, if, for all label assignments, there exists a parameter vector such that the model makes no errors when evaluating that set of data points. Thus the VC dimension of a classification model is the maximum number of data points that are shattered by it.
Assuming to have a d dimensional response variable where each item is a binary outcome, there were 2^d possible outcomes in the sample space and 2^d possible response patterns. If one would build a DT with d levels and d leaves, this DT could shatter any pattern of 2^d responses. Outside this constructed scenario, to avoid overfitting, the DT would first be overfitted and then pruned back using cross-validation to get a smaller and simpler tree. The VC dimension measures the complexity for the hypothesis space of a classification model. Since it is difficult to say what the hypothesis space actually is in this case, and because cross-validation directly avoids the creation of over-complex trees, estimating VC dimensions for DTs is unnecessary does not make too much sense.
A DT's complexity is determined by its depth. The DT depth can either be directly defined by setting the respective model parameter or it can be influenced by changing the parameters for number of leafs, minimum number of instances per node, etc..

1 Description of the data and preprocessing steps

For 303 patients referred for coronary angiography at the Cleveland Clinic historical data was collected, invasive and noninvasive clinical tests were run plus angiography was performed. All tests were analyzed and the results recorded independently and without knowledge of other test data meaning no work-up bias. From the collected patient data in the test group 13 variables were derived that were entered into a computerized database. The data was structured in a table with 303 entries and 14 attribute columns. The first 13

attributes contained the clinical and test results, the additional attribute was the target variable describing the CAD diagnosis of each patient. The data set comprised continuous variables (Table 1 and variables with categorical values (Table 2) There were nan entries found in the column storing information on the

attribute	description	values
sex	Sex	1: male, 2: female
cp	Chest pain type	1: typical anginal, 2: atypical, anginal, 3: non-anginal, 4: asymptomatic
fps	Fasting blood sugar	0: fbs <= 120mg/dl, 1: fbs >120mg/dl
restecg	ECG result at rest	0 = normal, 1: ST-T wave abnormality, 2: probable or definite left ventricular hypertrophy
exang	Exercise induced angina	0: no, 1: yes
slope	Slope of the peak exercise ST segment	1: upsloping, 2: flat, 3: downsloping
thal	Exercise thallium scintigraphic defects	3: normal, 6: fixed defect, 7: reversible defect
num	Diagnosis of heart disease	0: <50% vessel diameter narrowing, 1-4: >50% diameter narrowing

Table 1

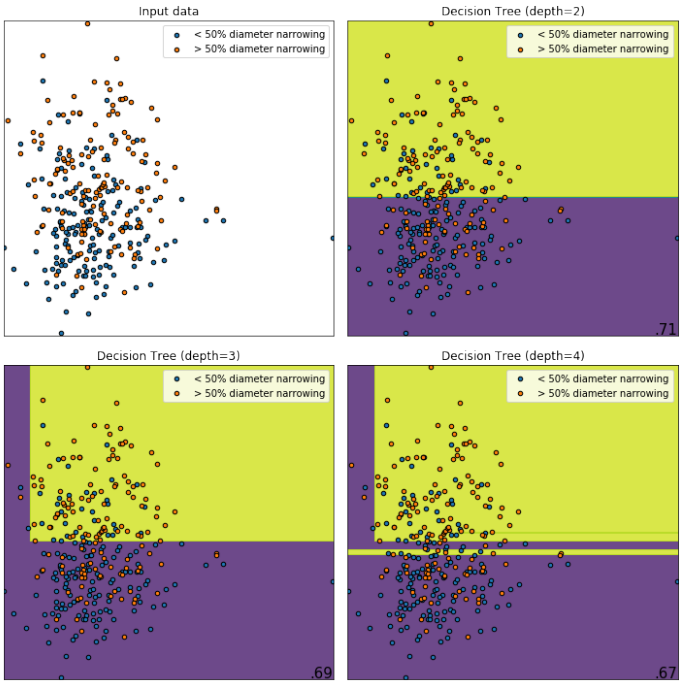
attribute	description	values
age	Age	age in years, range 29-77
trestbps	Resting blood pressure	recorded in mm Hg, range 94-200 mmHg
chol	Serum cholesterol	recorded in mg/dl, range 126-564 mg/dl
thalach	Maximum heart rate achieved during exercise	range 71-202 bpm
oldpeak	ST depression induced by exercise relative to rest	range 0-6.2 mm
ca	Number of major vessel colored by fluoroscopy	values 0-3

Table 2

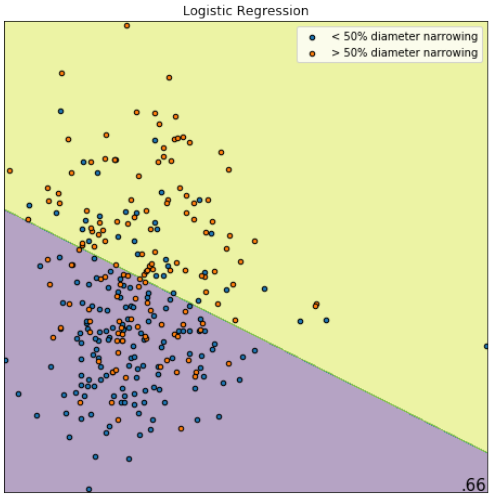
number of major vessels colored by fluoroscopy (ca) and for thallium scintigraphy (thal). Since the nan values comprised around 1.9% of the data set (6 entries), the respective subjects were deleted instead of performing imputation or choosing dummy values. The target attribute (num) encoded the magnitude of artery diameter narrowing in an angiogram with 5 values (0,1,2,3,4). Since this project aimed at simply distinguishing between the presence or absence of CAD the target attribute was redefined as a boolean variable. All values >=1 (1,2,3,4) were changed into 1 to encode the presence of CAD, zero values encoding the absence of CAD remained unchanged. All variables that were read in as floating point numbers but only contained integer values were redefined to integer values.

2 Methods

The dataset with 13 features was reduced to two dimensions using principal component analysis (PCA). The reduced dataset was randomly split into a training (60%) and testing (40%) set for cross-validation. Three decision trees were trained on the training set with depth 2, 3 and 4. Their accuracy was measured by evaluating the predicted target values for the testing set. Additionally, a logistic regression classifier was trained on the same training set and once again cross-validated with the testing set. The accuracy and the decision boundaries for all 4 classifiers were displayed (Figures 1a and 1b), as well as the tree layout of the decision trees (the first one displayed in Figure 2).



(a) The Decision Boundaries of three Decision Trees with depth 2, 3 and 4 on the two PCs of the data-set. The prediction accuracy is displayed in the lower right-hand corner.



(b) The Decision Boundary of the Logistic Regression on the two PCs of the data-set. The prediction accuracy is displayed in the lower right-hand corner.

Figure 1: Decision Boundaries for the decision trees and a logistic regression.

3 Results

Validating the Decision Tree classifiers with the test set split of the data resulted in an accuracy of 71%, 69% and 67% for the decision trees with depth 2, 3 and 4, respectively (Figure 1a, the accuracy can vary between runs due to the random split of the testing and training set). For this data-set, the simplest tree tested here was the one with the highest accuracy, though 71% accuracy is still not ideal. The simplest Tree coming out on top is probably the result of slight overfitting which was performed by the more complex trees. Still, the three DT classifiers showed only slight differences in accuracy, keeping the three trees almost equally accurate at predicting the test set.
The cross-validation of the logistic regression only resulted in an accuracy of 66% (Figure 1b), being lower than all of the three trained DT classifiers. Thus for this data-set, a classifier simpler than a decision tree resulted in a worse prediction. It was generally observed that the target values for the data set's principal components were not easily predictable.

4 Answering the project question

"Let us say our hypothesis class (see lecture slide 11) is a circle instead of a rectangle. What are the parameters? How can the parameters of a circle hypothesis be calculated in such a case? What if it is an ellipse? Why does it make more sense to use an ellipse instead of a circle? How can you generalize your code to K ¿ 2 classes?"
The parameters for a circular hypothesis class could be a center point and a radius. For every point that needs to be classified, a distance measure is used to get a distance to the center point and the point is included in the group if the distance is smaller than or equal to the radius. The ellipse is defined by a center point and two radii, that define the deformed circle shape. All points (x,y) around the center point (h,k) fulfilling this

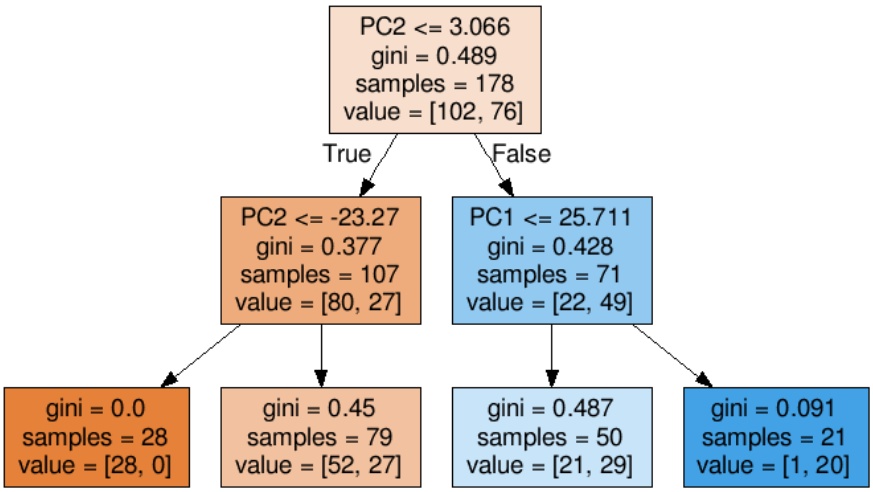


Figure 2: Decision Tree with depth 2.

equation: $\frac{(x-h)^2}{r_x^2} + \frac{(y-k)^2}{r_y^2} \leq 1$ are then in the decision class. An ellipse is a more general geometric shape, with the circle being a special ellipse, so it can be fitted to more data point cloud shapes.
To make the ellipse shape applicable to higher dimensions, the formula can be extended to sum over all n point, center and radius dimensions: $\sum_{x,h,r}^{n} \frac{(x_i-h_i)^2}{r_i^2} \leq 1$.