MAC0459/MAC5865 - Data and Engineering Science

Decision tree exercices - 10/2019

Q1. What is the approximate depth of a Decision Tree trained (without restrictions) on a training set with 10 million instances?

Q2. Is a node's Gini impurity generally lower or greater than its parent? Is it generally lower (greater), or always lower (greater)?

Q3. If a Decision Tree is overfitting the training set, is it a good idea to try decreasing max_depth ? Why?

Q4. If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features? Why?

Q5. If it takes one hour to train a Decision Tree on a training set containing 10 million instances, roughly how much time will it take to train another Decision Tree on a similar training set containing 100 million instances?

Q6. Train and fine-tune a Decision Tree for the Cardiovascular Disease dataset. The dataset has very few information on acquisition protocol except that all of the dataset values were collected at the moment of medical examination.

There are three types of input features:

- Objective: factual information;
- Examination: results of medical examination;
- Subjective: information given by the patient.

Besides that, features are classified according to the following values:

- 1: normal,
- 2: above normal,
- 3: well above normal

Questions or tasks to be addressed:

1. Present an EDA analysis of the dataset.
2. Split it into a training set and a test set using `train_test_split` method from sklearn.
3. Use grid search with cross-validation (with the help of the `GridSearchCV` method from sklearn) to find good hyperparameter values for a `DecisionTreeClassifier`. Hint: try various values for `max_leaf_nodes`.
4. Train it on the full training set using these hyperparameters, and measure your model's performance on the test set.

Table 1: Data description.

Feature	**Type**	**column**	**value**
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1, 2, 3
Glucose	Examination Feature	gluc	1, 2, 3
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary