Project 2016-17

Predicting forest cover type from cartographic variables only (7 class values)

Projects can be conducted by teams of 2. The provided dataset is quite large, so if you use WEKA you may run into problems. You can either use other packages, or reduce the dataset. In both cases, you must carefully explain what you did.
You are asked to use one or more ML algorithms of your choice, to run several experiments with different feature settings, and to perform a careful performance evaluation with confidence intervals.
You must produce a **report** (>=10 pages) with the explanation of what you did: any data preprocessing, the software packages you used, the experiments and the evaluation. Graphs and statistics are welcome.

Selected students can propose their own project.

The project is worth 30% of your final grade this year. I will register your grade on INFOSTUD only when you handle the project, in addition to having passed the written exam (either mid term and second term, or the full exam). Rounding of your final grade is performed on the basis of your homeworks.

Description of the datset

The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

Summary Statistics

Number of instances (observations)	581012
Number of Attributes	54
Attribute breakdown	12 measures, but 54 columns of data (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables)
Missing Attribute Values	None

Variable Information

Given is the variable name, variable type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

Name	Data Type	Measurement	Description
Elevation	quantitative	meters	Elevation in meters
Aspect	quantitative	azimuth	Aspect in degrees azimuth
Slope	quantitative	degrees	Slope in degrees
Horizontal_Distance_To_Hydrology	quantitative	meters	Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology	quantitative	meters	Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways	quantitative	meters	Horz Dist to nearest roadway
Hillshade_9am	quantitative	0 to 255 index	Hillshade index at 9am, summer solstice
Hillshade_Noon	quantitative	0 to 255 index	Hillshade index at noon, summer soltice
Hillshade_3pm	quantitative	0 to 255 index	Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points	quantitative	meters	Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns)	qualitative	0 (absence) or 1 (presence)	Wilderness area designation
Soil_Type (40 binary columns)	qualitative	0 (absence) or 1 (presence)	Soil Type designation
Cover_Type (7 types)	integer	1 to 7	Forest Cover Type designation

Code Designations

```
Wilderness Areas:
    1 -- Rawah Wilderness Area
    2 -- Neota Wilderness Area
    3 -- Comanche Peak Wilderness Area
    4 -- Cache la Poudre Wilderness Area

Soil Types:
    1 to 40 : based on the USFS Ecological Landtype Units for this study area.

Forest Cover Types:
    1 -- Spruce/Fir
    2 -- Lodgepole Pine
    3 -- Ponderosa Pine
    4 -- Cottonwood/Willow
    5 -- Aspen
    6 -- Douglas-fir
    7 -- Krummholz
```

Class Distribution

```
Number of records of Spruce-Fir:          211840
Number of records of Lodgepole Pine:      283301
Number of records of Ponderosa Pine:       35754
Number of records of Cottonwood/Willow:     2747
Number of records of Aspen:                 9493
Number of records of Douglas-fir:          17367
Number of records of Krummholz:            20510
Number of records of other:                    0

Total records:                            581012
```