Dataset analysis

1. Dataset Size and type

* 15120 samples
* 55 dimensions

1-10 dimensions: continuous values

11-54 dimensions: binary values

55th dimension: forest cover type (total 7 types)

1. Statistical description

This part statistically analyses mean, standard deviation, min, max etc. of each feature.

Here list analysis of first three features.

          Elevation        Aspect         Slope

Count 15120.000000 15120.000000  15120.000000

mean    2749.322553    156.676653     16.501587

std      417.678187    110.085801      8.453927

min     1863.000000      0.000000      0.000000

25%     2376.000000     65.000000     10.000000

50%     2752.000000    126.000000     15.000000

75%     3104.000000    261.000000     22.000000

max     3849.000000    360.000000     52.000000

1. Skewness

Skewness measure how asymmetric the distribution can be.

* Values close to 0 show less skew
* some algorithms may benefit if large skew is corrected

Eg. Skewness of 1st to 5th feature in our dataset

Elevation                               0.075640

Aspect                                  0.450935

Slope                            0.523658

Horizontal\_Distance\_To\_Hydrology      1.488052

Vertical\_Distance\_To\_Hydrology          1.537776

1. Class distribution

Check number of instances belonging to each class.

In this dataset:

Cover\_Type # of instances

1    2160

2    2160

3    2160

4    2160

5    2160

6    2160

7    2160

1. Correlation between two features

Below list high correlated pairs (threshold 0.5):

Feature 1 Feature 2 Correlation score

Hillshade\_9am Hillshade\_3pm -0.78

Horizontal\_Distance\_To\_Hydrology Vertical\_Distance\_To\_Hydrology 0.65

Aspect Hillshade\_3pm 0.64

Hillshade\_Noon Hillshade\_3pm 0.61

Slope Hillshade\_Noon -0.61

Aspect Hillshade\_9am -0.59

Elevation Horizontal\_Distance\_To\_Roadways 0.58

The high correlated features can be further processed to decrease the number of dimensions. Thus decrease the complexity of the model.

1. Scatter

Plot highly correlated pairs. Below are scatters of two examples from our dataset.

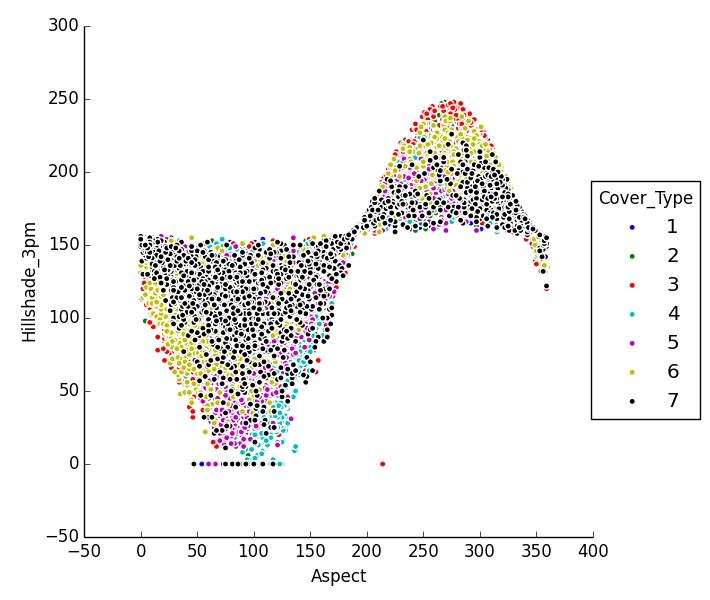


Figure 1 relation between Hillshade\_3pm and Aspect

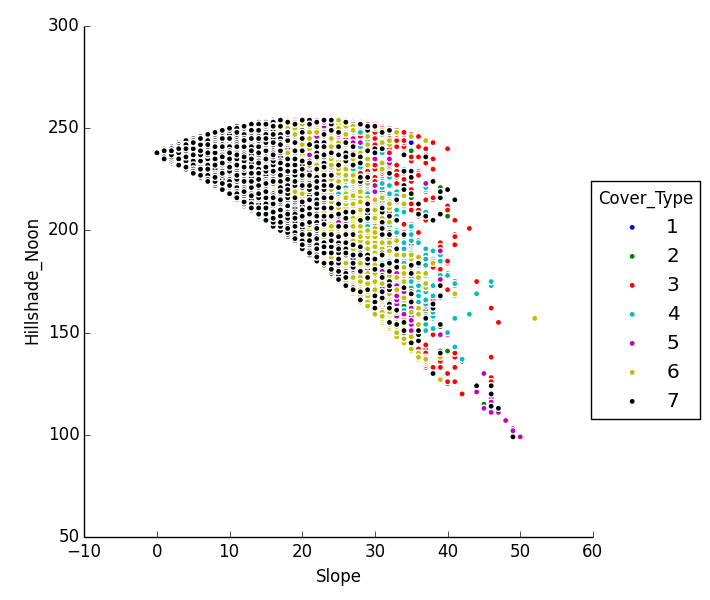


Figure 2 Relation between Hillshade\_Noon and Slope

It can be seen from the figure 1 that cover type No.7 scattered around 150 Hillshade\_3pm. Other features scattered far away from 150 Hillshaed\_3pm. When the aspect is less than 200, all cover types are less than 150 Hillshade\_3pm and vice versa. In figure 2, when the slope is low, it is more likely to be cover type 1. With the increasing of the slope, cover type changes to type 6 and then other types.