For this example, number of trending days for YouTube can be identified as a classification problem. This is binary classification since there are only 2 classes:

* trending greater then 4 days
* or trending 4 days or less.

The classifier utilizes some training data to understand how given input variables relates to the class. In this case the following have been used to train the data:

* known trending days,
* length of the tag,
* YouTube category

When the classifier is trained accurately, it can be used to predict trending days for the testing data.

* Naïve Bayes can be used on a large data set.

This data set for US data only has tags for 21350 records.

|  |  |  |
| --- | --- | --- |
| **Trending Days** | **Input variables** | **Accuracy of model** |
| Greater than 8 days | Tags only | 78.7% |
| Greater than 8 days | Tags, length, category codes | 87.8% |
| Greater than 3 days | Tags only | 72.0% |
| Greater 4 days | Tags, length, category codes | 63.0% |
| Greater 4 days | Tags only | 60.4% |