# **KNN & Naive Bayes part**

# zID: z5235682

Name: Genyuan Liang

# **Nearest Neighbour classification**

K-Nearest Neighbour classification can be used as a machine learning methods. For each unknown data points, calculating the distance of k number of the most adjacent data points and then calculating the average in order to classify.

Unlike most machine learning methods, the KNN algorithm does not have a clear model fitting step to the training data. We only need to achieve predictions through this method. For distance based KNN, due to its sensitive to scale of the features, higher dimensions will lead to the observations in each partition rapidly approach zero, and the distance from one observation point to another tends to be far away. Therefore, in this model, we need to select the features which are more highly correlated with the target to reduce the dimensions.

First look at all the features, after

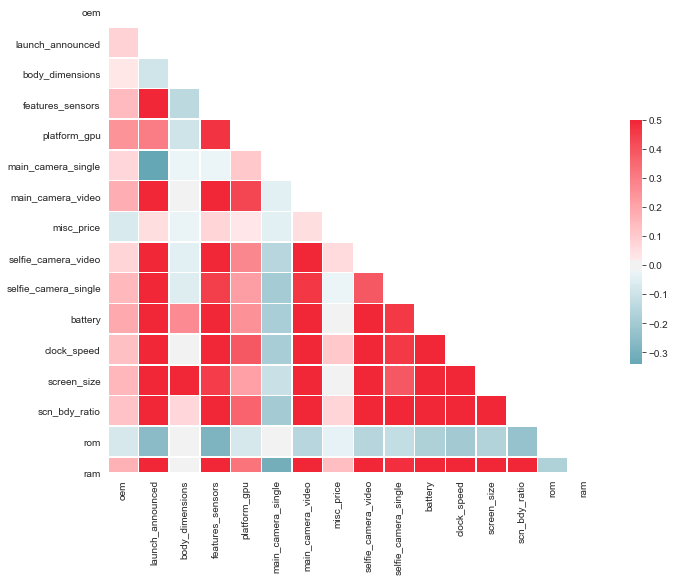


Figure 1: the correlation heatmap of training dataset

Then we select

# **Naive Bayes**

In machine learning, the naive bayes classifier assumes that the features are independent of each other, and calculates the probability according to the Bayes formula to determine the label. This assumption is often not true in practical applications. In this model, according to the correlation heatmap of training dataset above, it can be seen that there may actually be a large relationship between attributes, and the number of attributes is relatively large, so use plain The effect of Bayesian classification is not very good, and the actual accuracy obtained will not be very high.

There are several naive bayes classification algorithm classes in sklearn, such as BernoulliNB, MultinomialNB , GaussianNB and ComplementNB. In this dataset, we use these four models to train the data. Gaussian Naive Bayes which prior is Gaussian distribution can perform better if the distribution of sample features are mostly continuous values. Multinomial Naive Bayes which prior is a polynomial distribution is more appropriate to use when the distribution of sample features is mostly multivariate discrete values. Bernoulli Naive Bayes which prior is Bernoulli distribution is suitable for sample features that are binary discrete values or very sparse multivariate discrete values. ComplementNB is designed to correct the "serious assumption" of the standard polynomial naive Bayes classifier. It is particularly suited to unbalanced data sets.

For our dataset, the accuracy of the each kind of naive bayes classification are showed in table.

Table 1: the accuracy of naive bayes models

|  |  |
| --- | --- |
| **model** | **accuracy** |
| BernoulliNB | 0.4983922829581994 |
| MultinomialNB | 0.42765273311897106 |
| GaussianNB | 0.5305466237942122 |
| ComplementNB | 0.5852090032154341 |