CS235 Poisonous Mushrooms Classification

Kuai Yu
UC Riverside
kyu035@ucr.edu

I-Hsien Huang
UC Riverside
ihuan010@ucr.edu

Yifei Lai
UC Riverside
ylai017@ucr.edu

Introduction

In order to determine whether is possible to classify the toxicity (binary classification) of various types of mushrooms based on various features we have decided to use UCI's Mushroom Dataset which has a total of 22 features as follows:

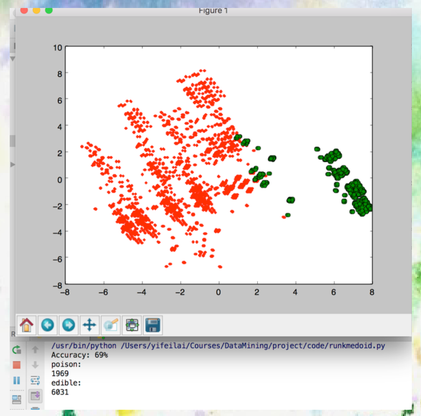
```
['poisonous',
 'cap-shape',
 'cap-surface',
 'cap-color',
 'bruises?',
 'odor',
 'gill-attachment',
 'gill-spacing',
 'gill-size',
 'gill-color',
 'stalk-shape',
 'stalk-root',
 'stalk-surface-above-ring',
 'stalk-surface-below-ring',
 'stalk-color-above-ring',
 'stalk-color-below-ring',
 'veil-type',
 'veil-color',
 'ring-number',
 'ring-type',
 'spore-print-color',
 'population',
 'habitat']
```

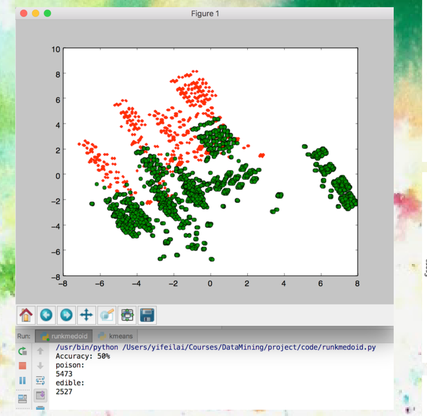
	poisonous	cap-shape	cap-surface	cap-color	bruises?	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type
0	e	x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	p
1	e	b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p
2	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p
3	e	x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	e
4	e	x	y	y	t	a	f	c	b	n	...	s	w	w	p	w	o	p
5	e	b	s	w	t	a	f	c	b	g	...	s	w	w	p	w	o	p
6	e	b	y	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p
7	p	x	y	w	t	p	f	c	n	p	...	s	w	w	p	w	o	p
8	e	b	s	y	t	a	f	c	b	g	...	s	w	w	p	w	o	p
9	e	x	y	y	t	l	f	c	b	g	...	s	w	w	p	w	o	p
10	e	x	y	y	t	a	f	c	b	n	...	s	w	w	p	w	o	p

	poisonous_e	poisonous_p	cap-shape_b	cap-shape_c	cap-shape_f	cap-shape_k	cap-shape_s	cap-shape_x	cap-surface_f	cap-surface_g	...	population_s	population_v
0	1	0	0	0	0	0	0	1	0	0	...	0	0
1	1	0	1	0	0	0	0	0	0	0	...	0	0
2	0	1	0	0	0	0	0	1	0	0	...	1	0
3	1	0	0	0	0	0	0	1	0	0	...	0	0
4	1	0	0	0	0	0	0	1	0	0	...	0	0
5	1	0	1	0	0	0	0	0	0	0	...	0	0
6	1	0	1	0	0	0	0	0	0	0	...	1	0
7	0	1	0	0	0	0	0	1	0	0	...	0	1
8	1	0	1	0	0	0	0	0	0	0	...	1	0
9	1	0	0	0	0	0	0	1	0	0	...	0	0
10	1	0	0	0	0	0	0	1	0	0	...	1	0
11	1	0	1	0	0	0	0	0	0	0	...	1	0
12	0	1	0	0	0	0	0	1	0	0	...	0	1
13	1	0	0	0	0	0	0	1	0	0	...	0	0
14	1	0	0	0	0	0	1	0	1	0	...	0	0

K-Medoids Cluster Analysis

The k-medoids algorithm is similar to k-means algorithm, with the k-medoids algorithm choosing one data point as the center of one cluster as opposed to calculating a theoretical average. Each time an object is injected into the dataset, it will be assigned to a cluster which ensures the total distance between all points and the center of their corresponding clusters is minimized, and the central data point of each cluster may change.



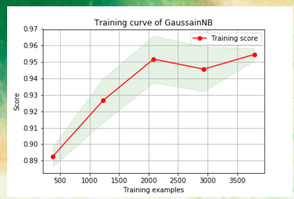


Data Preprocessing & Dataset Analysis

After using OpenRefine to remove null-data, We used factorplots to determine the relationships between features' categorical value frequencies and the output label. Note that several features contribute significantly to the ouput label.

Naive Bayes Classification

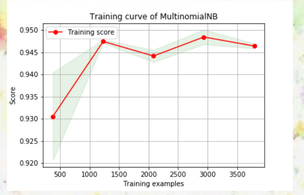
The reason why naïve Bayes is considered for our project is because of its ability to produce accurate classification for complex datasets with a minimal amount of training data. In some instances, naïve Bayes can even outperform more complex models such as logistic regression and decision trees for nominal input values.



$$p(X|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{(\theta-\mu_i)^2}{2\sigma_i^2}}$$

```
Misclassified samples: 127 out of 2437
Accuracy: 0.9479 F1: 0.9487
```

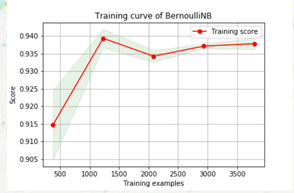
Gasussian Likelihood Distribution



$$p(X|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki^{xi}}$$

```
Misclassified samples: 115 out of 2437
Accuracy: 0.9528 F1: 0.9575
```

Multinomial Likelihood Distribution



$$p(X|C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1-p_{ki})^{(1-x_i)}$$

```
Misclassified samples: 153 out of 2437
Accuracy: 0.9372 F1: 0.9436
```

Bernoulli Liklihood Distribution

Random Forest Classification

The Random Forest classifier constructs a lot of decision trees with random features. Each node is a feature, and which are chosen at random to build the tree. After training, each tree will vote for the label. We were able to determine the most useful features of the dataset for the Random Forest classifier as well as achieve an accuracy of 0.996.