Course Project on Unsupervised Machine Learning

Main objective of the analysis

The objective of this analysis is to develop clustering model which divides tumors into clusters depending on their characteristics. Such model can be used in order to diagnose whether the tumor is benign or malignant. Although most models for diagnosing cancer are classification models, clustering models can be also used, especially, when there data is unlabeled.

Brief description of the data set and a summary of its attributes

The data set used in this project is publicly available from the UCI Machine Learning Repository. It consists of 699 human cell sample records, each of which contains the values of a set of cell characteristics. 458 records are benign tumors and 241 are malignant.

The fields in each record are:

Field name	DescriptionPatient identifier
ID	Patient identifier
Clump	Clump thickness
UnifSize	Uniformity of cell size
UnifShape	Uniformity of cell shape
MargAdh	Marginal adhesion
SingEpiSize	Single epithelial cell size
BareNuc	Bare nuclei
BlandChrom	Bland chromatin
NormNucl	Normal nucleoli
Mit	Mitoses
Class	Benign or malignant

The goal of this project is to build a clustering model which will divide the examples into clusters as much closer to the actual division into classes.

This model is an unsupervised model, but the data set actually contains labels. So, for training the model, I created a training set which included only the features, except the patient identifier which is not needed here.

Then, after the training of the model, I used labels (class) for the testing of the model. I did not divide data into training and testing sets because there was no need to do it since I used the labels for testing model.

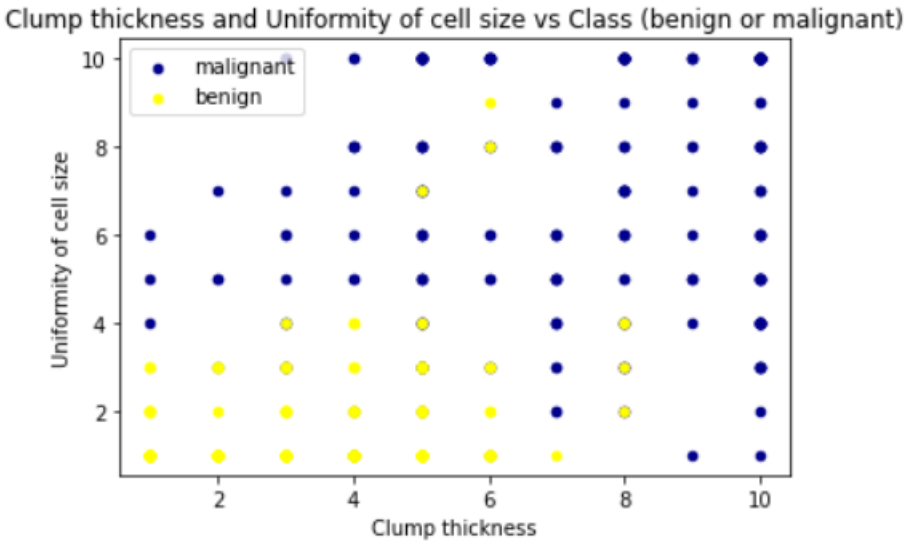
Brief summary of data exploration and actions taken for data cleaning and feature engineering

Data exploration includes data collection, data cleaning, analysis, and feature engineering.

Data Collection. I downloaded the data set from Internet and examined it.

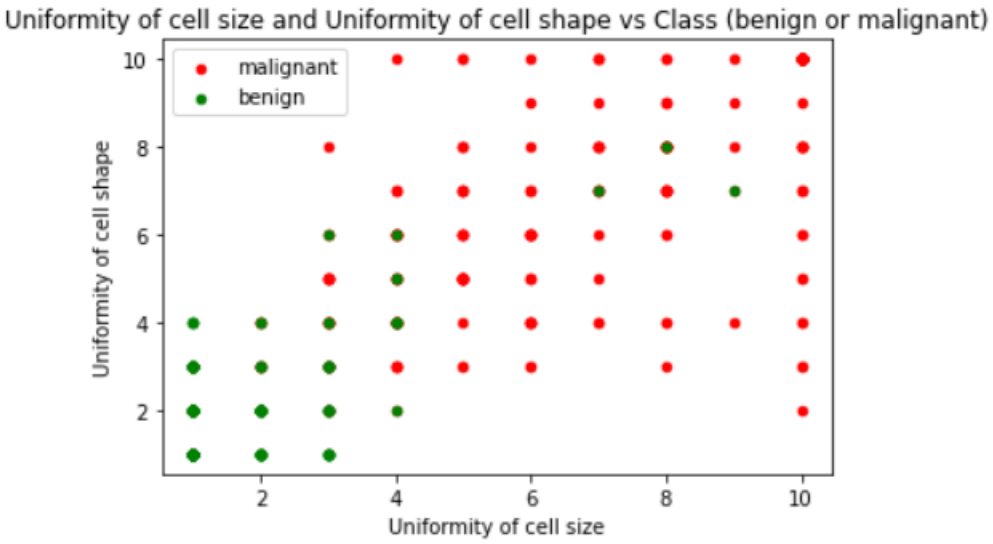
Analysis. This included plotting some diagrams.

1. Clump thickness and Uniformity of cell size vs Class (benign or malignant):



Although in the data set there are more benign examples than malignant ones, malignant tumors are more diverse than benign ones.

2. Uniformity of cell size and Uniformity of cell shape vs Class (benign or malignant) is shown on the next page.



This plot shows the same tendency: there is more diversity among malignant tumors than among benign tumors.

3. Pearson correlation coefficients between Class and all the features:

	Feature	Pearson correlation	P-value
0	Clump	0.714790	7.292504e-108
1	UnifSize	0.820801	8.922226e-168
2	UnifShape	0.821891	1.369425e-168
3	MargAdh	0.706294	2.979778e-104
4	SingEpiSize	0.690958	4.733540e-98
5	BareNuc	0.822696	3.401103e-169
6	BlandChrom	0.758228	1.267712e-128
7	NormNucl	0.718677	1.465645e-109
8	Mit	0.423448	4.304040e-31

The p-values are low for all Pearson correlation coefficients for correlation between Class and features. This means that there is correlation between the class of tumor and all the features. The highest correlation is between the class of tumor and its uniformity of cell size, uniformity of cell shape, and bare nuclei.

Data cleaning and feature engineering included:

- Removal of rows with missing values
- Converting all the features into numerical
- Data normalization

Also, for convenience of future analysis, I created a new column for class. In the original data set, benign examples have Class = 2, and malignant examples have Class = 4. But in the new column, I used categorical values - “benign” for benign examples and “malignant” for malignant examples.

Summary of training clustering models

In this project, I built and trained the following clustering models:

- K-means clustering
- Hierarchical agglomerative clustering (HAC) with ward linkage
- HAC with complete linkage
- HAC with average linkage
- HAC with single linkage

For all the models, I used the same training set which contained all the examples from the data set, but only features. I removed the labels (class) and also the patients’ identifiers. There was no need to divide the examples into training and testing sets because for evaluation I used the labels (class).

For evaluation of all the models, I used the same metrics – adjusted Rand index scores and Fowlkes Mallows scores.

1. K-Means

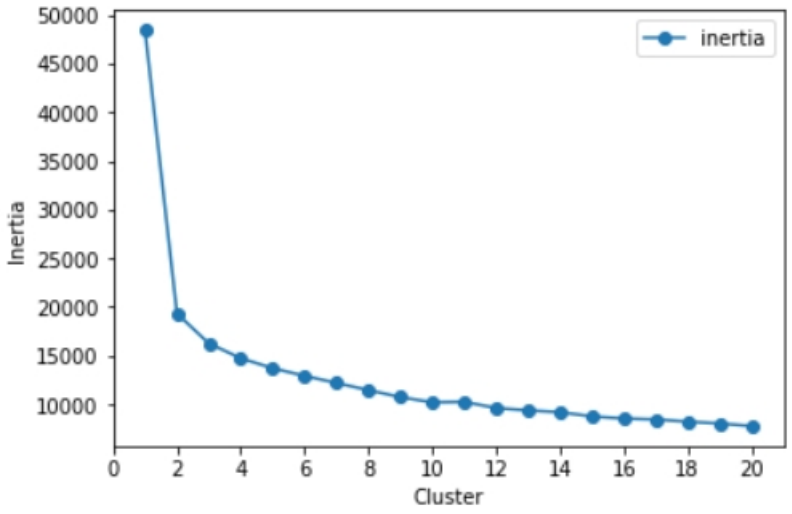
In [46]:
```
from sklearn.cluster import KMeans
km = KMeans(init = "k-means++", n_clusters = 2, n_init = 12)
km = km.fit(df[tr_columns])

df['kmeans'] = km.predict(df[tr_columns])
```

In [47]:
```
(df[['class','kmeans']]
 .groupby(['class','kmeans'])
 .size()
 .to_frame()
 .rename(columns={0:'number'}))
```

Out[47]:

		number
class	**kmeans**	
benign	**0**	9
	1	435
malignant	**0**	221
	1	18



This diagram shows that increasing the number of clusters would decrease the error of the model. However, the purpose of the clustering model in this project is to divide the examples into clusters that would match the classes (benign or malignant) as much as possible. Since the number of classes is 2, so the number of clusters also has to be 2.

2. Hierarchical Agglomerative Clustering (HAC)

2.1. HAC with ward linkage

In [20]:
```
from sklearn.cluster import AgglomerativeClustering
ag_1 = AgglomerativeClustering(n_clusters=2, linkage='ward', compute_full_tree=True)
ag_1 = ag_1.fit(df[tr_columns])
df['agglom-w'] = ag_1.fit_predict(df[tr_columns])
```

In [48]:
```
(df[['class','agglom-w']]
 .groupby(['class','agglom-w'])
 .size()
 .to_frame()
 .rename(columns={0:'number'}))
```

Out[48]:

		number
class	agglom-w	
benign	0	22
	1	422
malignant	0	238
	1	1

2.2. HAC with complete linkage

In [22]:
```
ag_2 = AgglomerativeClustering(n_clusters=2, linkage='complete', compute_full_tree=True)
ag_2 = ag_2.fit(df[tr_columns])
df['agglom-c'] = ag_2.fit_predict(df[tr_columns])
```

In [51]:
```
(df[['class','agglom-c']]
 .groupby(['class','agglom-c'])
 .size()
 .to_frame()
 .rename(columns={0:'number'}))
```

Out[51]:

		number
class	agglom-c	
benign	0	443
	1	1
malignant	0	128
	1	111

2.3. HAC with average linkage

In [24]:
```
ag_3 = AgglomerativeClustering(n_clusters=2, linkage='average', compute_full_tree=True)
ag_3 = ag_3.fit(df[tr_columns])
df['agglom-a'] = ag_3.fit_predict(df[tr_columns])
```

In [50]:
```
(df[['class','agglom-a']]
 .groupby(['class','agglom-a'])
 .size()
 .to_frame()
 .rename(columns={0:'number'}))
```

Out[50]:

		number
class	agglom-a	
benign	0	8
	1	436
malignant	0	208
	1	31

3.4. HAC with single linkage

In [26]:
```
ag_4 = AgglomerativeClustering(n_clusters=2, linkage='single', compute_full_tree=True)
ag_4 = ag_4.fit(df[tr_columns])
df['agglom-s'] = ag_4.fit_predict(df[tr_columns])
```

In [52]:
```
(df[['class','agglom-s']]
 .groupby(['class','agglom-s'])
 .size()
 .to_frame()
 .rename(columns={0:'number'}))
```

Out[52]:

		number
class	agglom-s	
benign	0	444
malignant	0	238
	1	1

Comparison of division into clusters by all the models:

class	kmeans	agglom-w	agglom-c	agglom-a	agglom-s	
benign	0	0	0	0	0	6
				1	0	2
			1	0	0	1
	1	0	0	1	0	13
		1	0	0	0	1
				1	0	421
malignant	0	0	0	0	0	97
				1	0	13
			1	0	0	110
					1	1
	1	0	0	1	0	17
		1	0	1	0	1

The summary of evaluation of all the models with metrics:

Model	Adjusted Rand	Fowlkes Mallows
K-Means	0.8465	0.9307
HAC (ward linkage)	0.8690	0.9393
HAC (complete linkage)	0.3607	0.7662
HAC (average linkage)	0.7817	0.9031
HAC (single linkage)	0.0025	0.7375

The HAC model with ward linkage gives the best performance. The K-Means model has performance just a little bit lower. Other models have lower performance.

Recommended final model

Among all the models used in this project, the HAC model with ward linkage has the highest scores on both on adjusted Rand index and on Fowlkes Mallows score. K-Means model has just a little bit lower performance. Other models (HAC with different kinds of linkage) have much lower scores and therefore cannot be recommended.

However, it is also important to look how these models actually divide the examples into classes. All the models used in this project create two clusters – cluster 0 and cluster 1. Most benign examples are in cluster 1 and most malignant examples are in cluster 0. Therefore, cluster 1 more or less matches the benign class and cluster 0 more or less matches the malignant class.

As can be seen from the tables below, HAC model with ward linkage (agglom-w) made less errors in division the examples into clusters than K-means model (kmeans). Moreover, K-means model misclassified 18 cases of benign tumors while HAC model with ward linkage misclassified only 1 case of benign tumor.

		number
class	kmeans	
benign	0	9
	1	435
malignant	0	221
	1	18

		number
class	agglom-w	
benign	0	22
	1	422
malignant	0	238
	1	1

For models that are used to diagnose cancer, the most important is that they should not misclassify benign tumors and therefore should not miss the cases of cancer.

This means that HAC model with ward linkage is the best model and should not be recommended.

Summary Key Findings and Insights

The goal of this project was to develop a clustering model that can be used to predict whether a tumor is benign or malignant.

Exploratory descriptive analysis of the data set used in this project, showed that there is correlation between the class (benign or malignant) and all the features. Therefore, all the features were used to train the models.

In this project, I built and trained 5 clustering models: one model with K-means clustering (for n_clusters = 2) and four HAC models with different linkages: ward, complete, average, and single.

I used two metrics for evaluating all the models – adjusted Rand index scores and Fowlkes Mallows scores. HAC model with ward linkage and K-means clustering model had the highest scores. Other models (that is, HAC models with other kinds of linkages had much lower scores.

The table of adjusted Rand index scores and Fowlkes Mallows scores for all the models looks like this:

Model	Adjusted Rand	Fowlkes Mallows
K-Means	0.8465	0.9307
HAC (ward linkage)	0.8690	0.9393
HAC (complete linkage)	0.3607	0.7662
HAC (average linkage)	0.7817	0.9031
HAC (single linkage)	0.0025	0.7375

HAC model with ward linkage not only had higher scores on the metrics than K-means, but also misclassified much less malignant tumors. Therefore, this is the best model.

Suggestions for next steps in analyzing this data

Although HAC model with ward linkage has quite good performance, it is probably possible to create even a better model tuning hyper-parameters and trying other kinds of models, for example, DBSCAN.

Also, the data set used in this project is not very large. It contains less than 700 examples. Model accuracy can also be improved by using more examples for training.

Another possible way to increase model performance is to use larger data sets with more features and more examples. If model is trained on more features, its performance will also increased.