Technical University of Denmark

02450 - Introduction to Machine Learning and Data Mining

E19

Project 1

Student:	**Student id**
Rasmus Tuxen	s173910
Irad Ohayon	s191499
Callum Blair	s193096

September 30, 2019

Contents

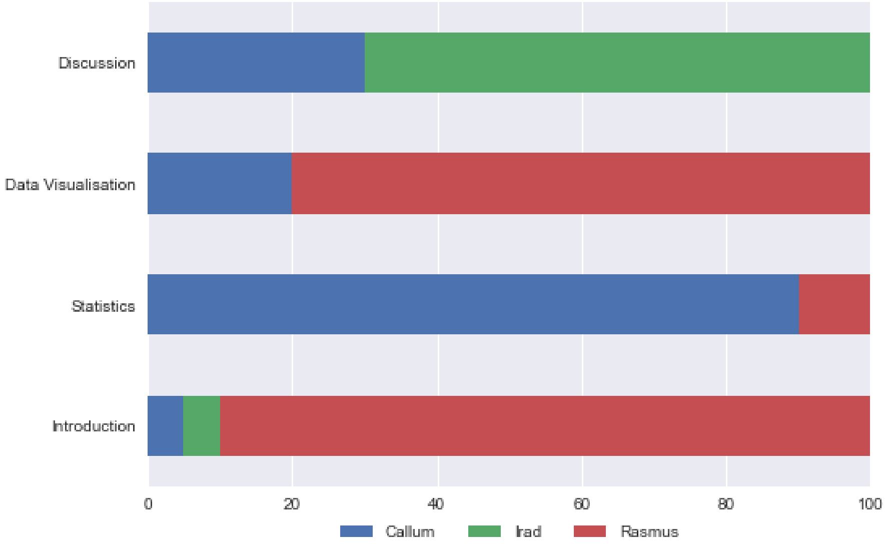


Figure 1: Split of responsibility per section

1 Introduction

This report is made for the course 02450: Introduction to Machine Learning and Data mining at DTU (Technical University of Denmark). The objective of this first assignment is to apply methods learned in the first section of the course, *"Data: Feature extraction, and visualization"* on a self-selected data set to get a understanding of the selected data prior to further analysis.

1.1 Data set

We have selected the Ecoli Data Set. The data set is retrieved from the UCI Machine Learning Repository. The title of the data set is *"Protein Localization Sites"*. It contains 336 instances and 9 attributes; one is a specific name for each instance, one is the class name for the different protein locations in the data set and the seven remaining are attributes used to classify the localization site of a protein on a score between 0 and 1.
The location of a protein is useful for determining its function and usefulness for research and industry purposes. Characterization of protein localization is accurate but slow and labor-intensive. However, methods involving the amino acid sequence and sequenced genomic data are newer and more efficient. Six of these methods for determining the location of the proteins features in this data set
The data set have previously been used[1] to predict protein localization sites in eukaryotic cells using machine learning techniques such as Decision Tree, Two-Layer Feed-Forward Neural Network and Perceptrons. The three techniques averaged a successful prediction 66.83% and an SD of 5.82% over 100 runs for the Decision Tree and the Perceptrons, and over 50 runs for the Two-Layered Neural Network with a threshold of 0.7.

The primary machine learning modeling aim for the data is classification given that the data set contains 1 class name attribute (protein location site) with 7 predictive features (scores from different methods). The goal of this classification would be to define a model for the given class of protein location given the predictive features to assign a class label to a new test instance of protein locations.
Linear regression would not be suitable for this data, due to the predictions being discrete classes, rather than a continuous variable. However, if the labels were to be omitted, the data could also be used as a clustering problem. Applying clustering algorithms to the data may allow for extraction of relations between the binding sites that were not known prior.
As the number of attributes in the data is not especially high, we will consider all of them in our analysis.

1.1.1 Attributes of the data

A complete list of all the attributes featured in the data set can be seen in table 1:

[1] Chen, Yetian : "Predicting the Cellular Localization Sites of Proteins Using Decision Tree and Neural Networks"

	Variable:	**Description**	Type of attribute
1	**Sequence Name**	Accession number for the SWISS-PROT database	Discrete/Nominal
2	**mcg**	McGeoch's method for signal sequence recognition	Continous/Interval
3	**gvh**	von Heijne's method for signal sequence recognition	Continous/Interval
4	**lip**	von Heijne's Signal Peptidase II consensus sequence score	Discrete/Nominal
5	**chg**	Presence of charge on N-terminus of predicted lipoproteins	Discrete/Nominal
6	**aac**	score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins	Continous/Interval
7	**alm1**	score of the ALOM membrane spanning region prediction program	Continous/Interval
8	**alm2**	score of ALOM program after excluding putative cleavable signal regions from the sequence	Continous/Interval
9	**type**	The class is the localization site	Discrete/Nominal

Table 1: Table of attributes in Ecoli data set

The **Sequence Name** contains a unique id for the SWISS-PROT database for each of the tested proteins. The attribute is therefore discrete and nominal (we can only apply equal/not equal). Since this is a unique id there is no gain of performing machine learning with this attribute. We will therefore not be using this attribute further in this project.
The attributes **mcg, gvh, aac, alm1** and **alm2** are all continuous values between 0 and 1. They are interval because distance between values can be measured (addition/subtraction). They are not ratio because zero on their scale does not mean absence of what is measured.
The attributes **lip** and **chg** are binary values and are thus discrete and nominal.
The measured proteins in the data set, **type**, are classified into 8 locations: cytoplasm (cp), inner membrane without signal sequence (im), perisplasm (pp), inner membrane with uncleavable signal sequence (imU), outer membrane (om), outer membrane lipoprotein (omL), inner membrane lipoprotein (imL), inner membrane with cleavable signal sequence (imS). The **type** attribute is thus discrete and nominal.

In the data set there are no missing values and it shows no signs of containing corrupted data. There is however an oddity in the fact that the binary attributes not are either [0,1], but [0.48,1] and [0.5,1] respectively. Looking at the attribute **chg** that measures a presence of charge, it having the value 0.5 instead of just 0 to mean no presence doesn't seem consequential and we therefore judge it do no be an issue.

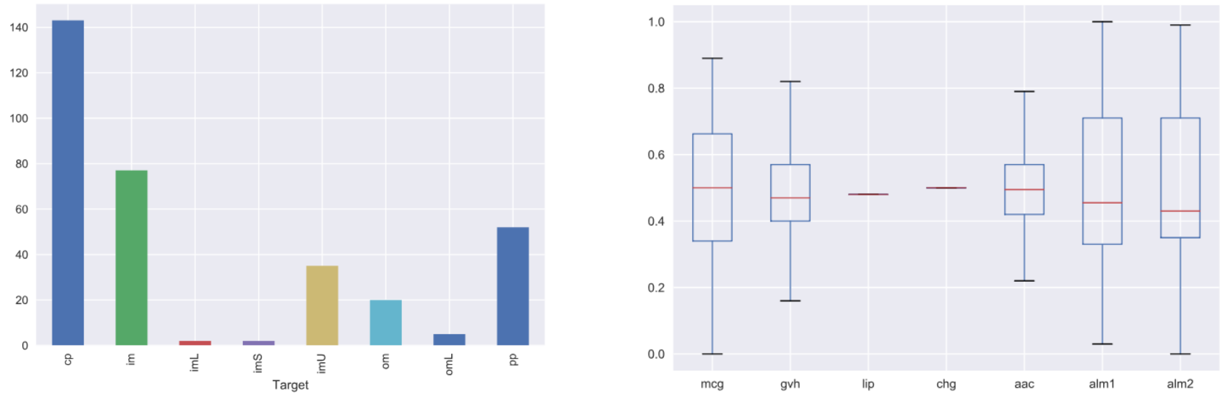
2 Statistical Analysis of Data

A first initial look at the summary statistics of the data set can be seen in table 2. It is a good way to get an initial feel of data and spot potential problems. As there are no missing values this summary is easily obtained.

	mcg	gvh	lip	chg	aac	alm1	alm2
count	336.000	336.000	336.000	336.000	336.000	336.000	336.000
mean	0.500	0.500	0.495	0.501	0.500	0.500	0.500
std	0.195	0.148	0.088	0.027	0.122	0.216	0.209
min	0.000	0.160	0.480	0.500	0.000	0.030	0.000
25%	0.340	0.400	0.480	0.500	0.420	0.330	0.350
50%	0.500	0.470	0.480	0.500	0.495	0.455	0.430
75%	0.662	0.570	0.480	0.500	0.570	0.710	0.710
max	0.890	1.000	1.000	1.000	0.880	1.000	0.990

Table 2: Statistical analysis of the Attributes

As we can clearly see from the table, the binary attributes **lip** and **chg** have incredibly low standard deviation. This owes to the fact that the vast majority of the entries for both of these attributes are the same value, which corresponds to the zero value (the other entries are 1, i.e. affirmative). This brings into question their usefulness as predictive attributes, as they are rarely affirmative.
All attributes have a mean of almost exactly 0.5, which indicates that the data has been adjusted to ensure they are all on the same scale. This explains why the binary values are not 0 and 1, as they have been altered to keep the means consistent between attributes.



(a)

(b)

Figure 2: Frequency of Target labels and box plots of Attributes

The plots shown in figure 2 show the frequency of the target labels in the data set and the box plots of the attributes. Unfortunately, the data is not well spread between the class labels, with some of them appearing only one or twice in the data. This will surely make predicting these labels incredibly difficult. However, predicting more broad classifications between the other categories should be achievable.

The box plots show that, while the mean of the data is all consistent, there is some variance in the location of the median value. This implies that the distribution is not a simple Gaussian shape for most of the attributes. This is made clear in figure 3 below.

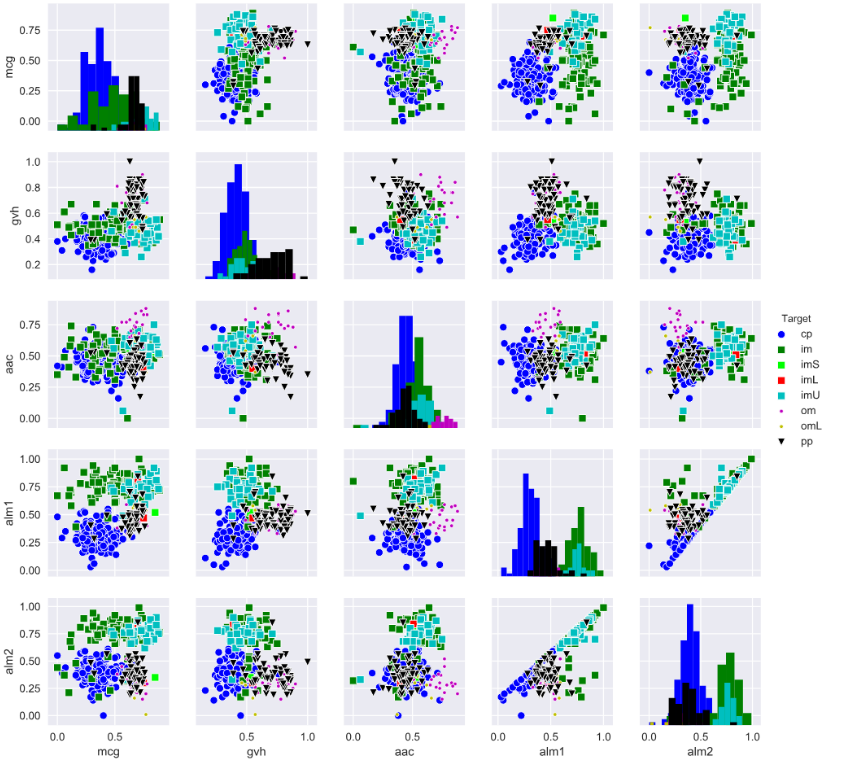


Figure 3: Grid plot with histogram on the diagonal

Here we see, on the diagonal, that the histograms of the distributions of the attributes show multiple peaks. Plotting the distribution by colour, depending on the label, shows that each of the peaks tends to correspond with a different class.
The remaining plots show the correlation between the attributes when plotted against each other. Once again the class labels have been included, which allows us to see if there is any separation of the classes due to the attributes. It should be noted that the binary attributes have been omitted from this graph as they offer no additional insight.
Many of the attributes do seem to offer some separation of the classes, as the overlapping clusters seem to demonstrate. We can also see that the the attributes **alm1** and **alm2** (perhaps unsurprisingly) have a strong correlation.

Additionally, to further visualise the correlation of the variables in the data set we've created a correlation heat map, see figure 4, where lighter colours translates to correlation and dark colours no correlation. Once again we see the strong correlation between **alm1** and **alm2**. Furthermore, there seems to be a decent correlation between **gvh** and **mcg**.

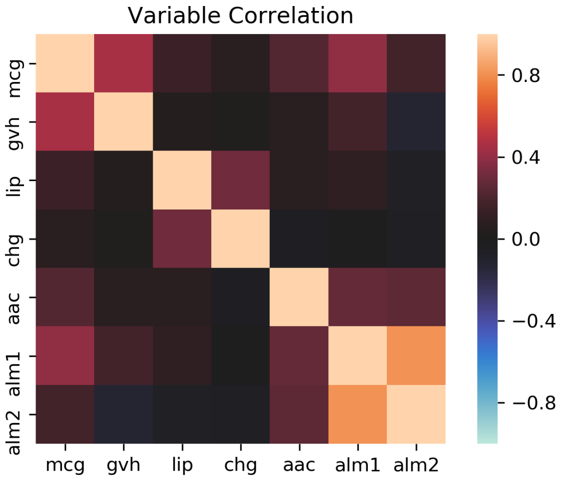


Figure 4: Heat map of correlation between attributes

The simple clustering that already appears in figure 3 indicates that it would be possible to predict the labels of the classes, especially after applying further analysis or some dimensionality reduction.

3 Data visualization

The principal component analysis (PCA) was used to find lower-dimensional representations of the attributes in the data-set, which initially contained 7 predictive attributes. To compute the PCA we center the data (see eq. 1) and since all attributes lie in the same interval between 0 and 1 there is no need to also normalize by dividing with its standard deviation.

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}, \qquad \mathbf{m} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{1}$$

Then to find the PCA directions ($\mathbf{V}$) and the variation in the data explained by the PCA's we compute the singular value decomposition (SVD) of the zero mean data $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
Now it is possible to obtain a subspace where the data-set in projected in the principle component direction to maximize variability within the data-set to hopefully find a lower-dimensional representation. A plot of the cumulative and individual variance explained by each principle component, i.e the singular values, found in the diagonal matrix $\mathbf{\Sigma}$, can bee seen in figure 5.

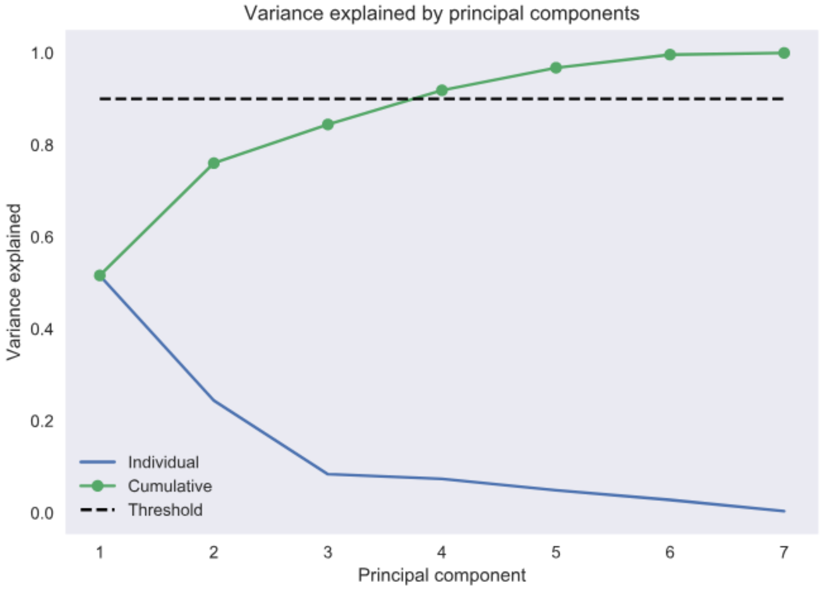


Figure 5: Variance explained of the cumulative and individual principal components

From figure 5 we observe, that more than 90% of the variation in the data is explained by the first 4 principal components and 95% for the first 5 principle components. This means that we are able to lower the dimensionality of the data set while still maintaining most of the variance in the data set possible.
If we choose the 3 first of the principal components and look at their coefficients for all attributes, see figure 6.

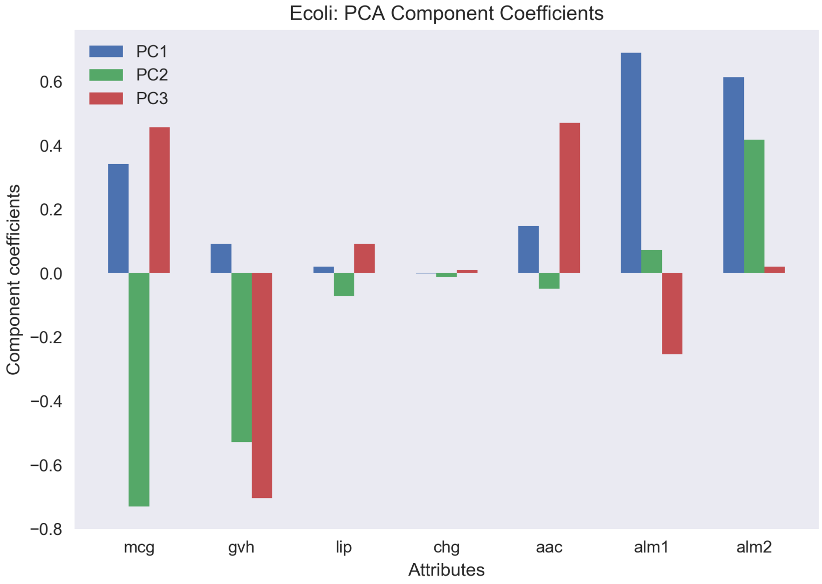


Figure 6: Principal directions for all attributes

We observe, that 1st principal component has a large coefficient for attributes **alm1** and **alm2** which both will have a positive projection onto PC1. In that way PC1 seems to favour protein location that receive a high score of the ALOM program. Observing PC2 we see that **mch** and

gvh will have a large negative projection while the **alm2** attribute will have a large positive projection. This indicates that PC2 seems to mostly describe the variance from the signal sequence methods (mch and gvh). While PC3 mostly shows the variance for **mcg**, **gvh** and **aac**.
It should also be noted that the binary values don't seem to be impacted much by projection of the first 3 principle components, but this doesn't mean that they are irrelevant but simply that their variance must be smaller compared to some of the other attributes.

3.1 2D PCA

We then look at the projected data on PC1 and PC2 (see figure 7) to visualize the relationship of the data set.

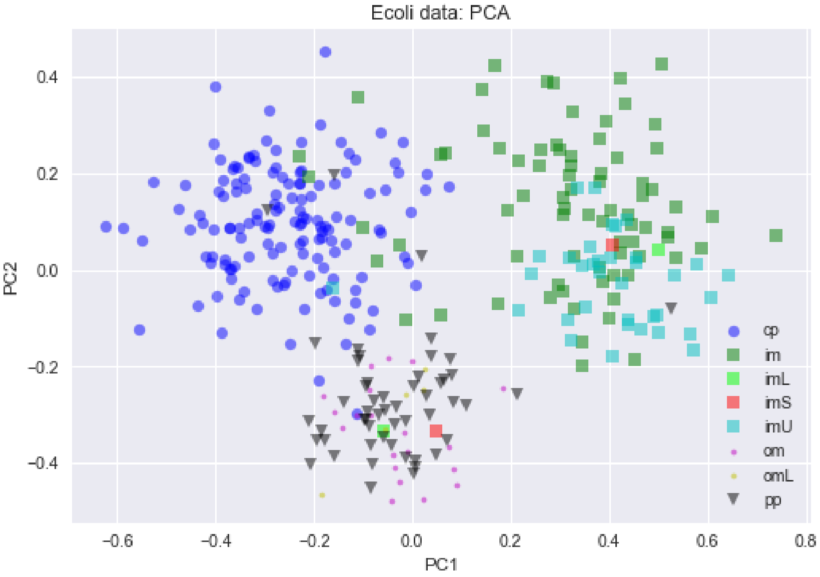
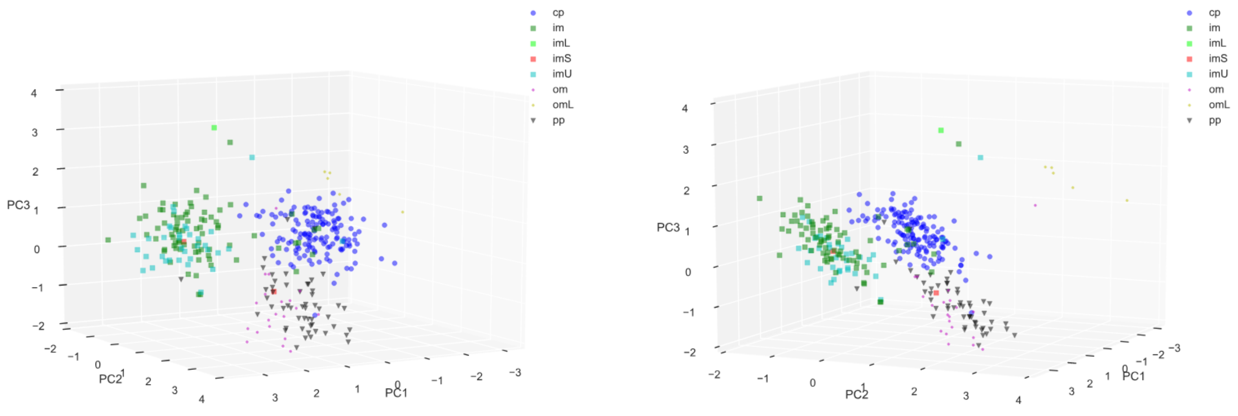


Figure 7: Data projected on two principle components

In the above scatter plot it can be seen that there are some grouping of similar protein location which is a good indication that classification of the protein locations from the data is possible. It is also interesting to see how similar locations like inner membrane locations (im,imL,imS,imU) and the outer membrane (om,omL) seems be grouped together. It is however also clear, that the two first principle components aren't enough to clearly classify all locations in the data set. This makes sense given that the two first principle components capture about 76% of the total variance in the data set. And that we would therefore need more principle components to keep a meaningful amount of the variance.

3.2 3D PCA

As we can see from figure 5, we should be able to visualise a significant portion more of the variance by including a third principle component. The results of plotting the principle components in three dimensions can be seen below in 8.



(a)

(b)

Figure 8: Plot showing 3 principle components

This plot has been zoomed in to the main clusters and it should be noted one point was excluded for the purpose of this visualisation. As expected, we see the same clusters as before, now with a depth direction. However, it can be seen that the extra dimension does not seem to aid too much in better separating the data. Especially the **im** and **imL** classes, which remain closely entwined.

4 Discussion

For our project we chose a data-set that we believe is a Classification problem but may also be used as a Clustering problem. The attributes of the data-set have been normalized to have a mean of 0.5, with the exception of rounding errors for the binary values, which allows for an easier analysis of the data-set. A statistical analysis of the attributes shows that the variance of most of the attributes implies of a distribution that is not normal and does not have the Gaussian shape, which makes the analysis of their correlation a more interesting challenge. After analyzing the Principal Components, we displayed the results in graphs that help show the relationship between the different attributes. Using the visualization of them, we managed to explain the functionality of the principal components in the data-set, as well as ways to apply them in our machine learning objective.

4.1 Suitability for Machine Learning

Having investigated the data, we can be confident that it is possible to carry out machine learning algorithms to classify the data. As the data is labelled, it has classes to act as targets to be used in supervised learning. Additionally, it would be possible to run clustering algorithms on the data, which can be visualised better by including the labels and testing whether they correspond with our given labels. It seems unlikely that we will be able to produce a model which can predict the classes to a high accuracy. This is due to the strong overlap which can be seen in the PCA between similar classes. As well as a lack of data for certain labels. It should, however, be possible to accurately predict whether the binding site is internal or external, as these show good separation in most cases.

Furthermore, it seems unlikely that the binary attributes in the data-set will be especially useful in the further analysis. Owing mainly to the fact that they are almost all negative. This leaves us with only 5 useful attributes to make the predictions on. Having so few attributes may restrict the predictive power of future models.

4.2 Comparison to t-SNE

While it is true that PCA is a powerful tool for dimensionality reduction, t-SNE is a more modern algorithm for dimensionality reduction. The theory of the method is outside of the scope of this report, However, for a comparison to the PCA we have done we carried out a t-SNE reduction with the tools from the sklearn library [2].

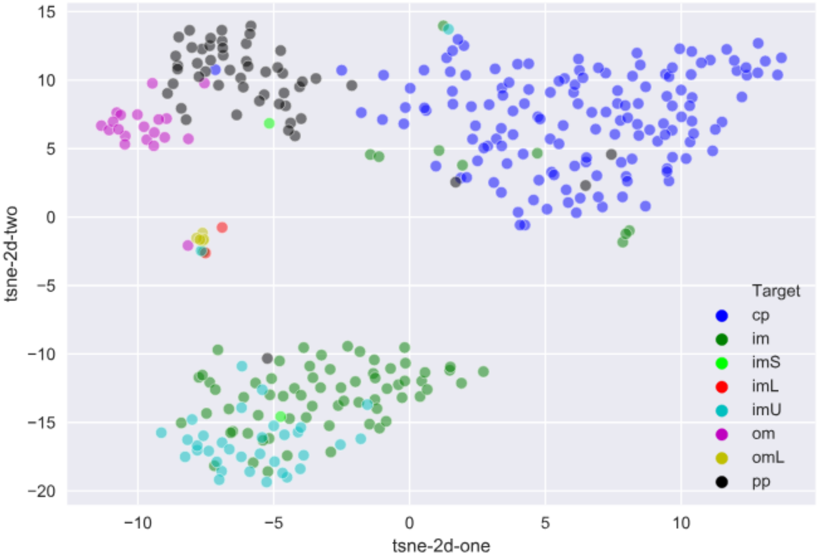


Figure 9: t-SNE with 2 components

As seen above, the method results in similar separation compared to PCA. While the clusters seem slightly tighter in spread, this method also has difficulty separating the classes **im** and **imU**. This visualisation lends to the conclusion that effective separation of classes is indeed possible to some extent, notwithstanding some of the trickier distinctions.

4.3 Use of other sklearn functions

Analysis of the data was carried out mainly manually by Rasmus, following the guidelines of the exercises. This makes us the bulk of what can be seen in the 2D analysis and plotting. In parallel, Callum carried out the same analysis using the PCA ability of the decomposition module of sklearn. The 2D plots have not been included, as they were redundant. It should be noted that the StandardScalar function used to scale the data also divides by the standard deviation, which was not done in the manual calculation. The effects of this were small on the grouping of the data, so we determined it was not of great importance. However, the 3D PCA plots have been made from the sklearn method, so the scaling of the principle components is slightly altered.

References

[1] Chen, Yetian : Predicting the Cellular Localization Sites of Proteins Using Decision Tree and Neural Networks
pdfs.semanticscholar.org/749e/51315f9a5fce4331994ffcab6a887aa1d5b6.pdf

[2] sklearn.manifold.TSNE
https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html