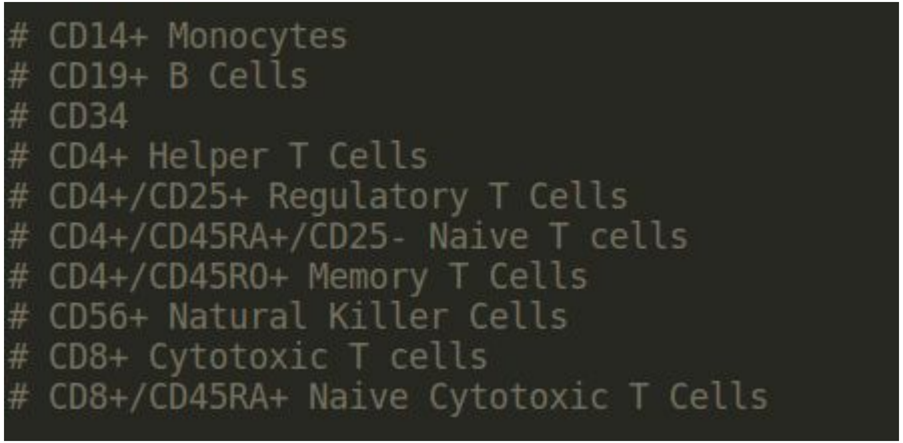
Machine learning exercises

Ed Morrissey

Morrissey Group: Quantitative biology of cell fate and tissue dynamics
Centre for Computational Biology - WIMM

Classify cells using single cell rna-seq

- PBMC data
- Published by 10x genomics
  - Massively parallel digital transcriptional profiling of single cells, Zheng et al. 2017
  - Data downloaded from
    - https://support.10xgenomics.com/single-cell-gene-expression/datasets/
- They published several data sets
- We will use
  - 10 bead purified populations of cells
  - Train different classifiers



```
# CD14+ Monocytes
# CD19+ B Cells
# CD34
# CD4+ Helper T Cells
# CD4+/CD25+ Regulatory T Cells
# CD4+/CD45RA+/CD25- Naive T cells
# CD4+/CD45RO+ Memory T Cells
# CD56+ Natural Killer Cells
# CD8+ Cytotoxic T cells
# CD8+/CD45RA+ Naive Cytotoxic T Cells
```

Read data in and plot

- Original data had 10,000 cells per cell type I’ve processed and filtered the data
  - 400 cells per file and ~6,000 genes that were filtered on expression level and variability
  - Data is normalised by dividing by the total counts per cell and multiplied by 10,000
- Should be 10 files, read them in using pandas
  - Concatenate into a single file
  - Create a labels vector by repeating the file names 400
    - Tip get the names with “data_sets = os.listdir("Exercises/Sep_data")”
- Plot the data
  - Use umap with 20 neighbours to get lower dim for plotting
  - Create a data frame with umap dims and labels as an extra column
  - Use seaborn to scatterplot and colour by cell type

Split data into train and test and run decision tree

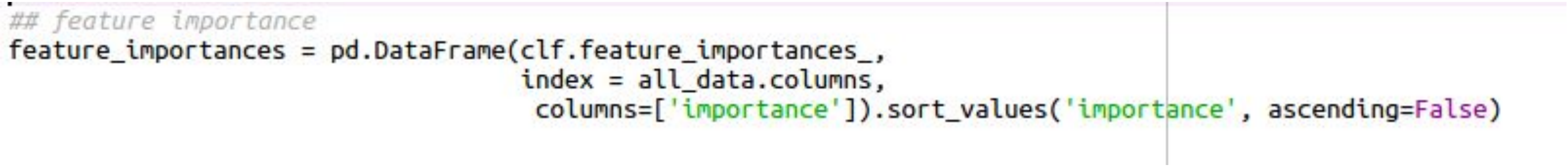
- Use “from sklearn.model_selection import train_test_split”
  - Split the data and the labels into 0.25 test data
- Train the DecisionTreeClassifier
- Use the score method to see how well it predicts
  - Do this for both the train and test data
  - Are the values different? What does it mean?

Random forest

- Train a random forest with the same data
- Calculate the score for training data and test data
    - Has it changed?
    - Better or worse?
- Change the number of trees used (n_estimators) to 250
    - Does it improve the score?

Random forest

- The random forest can calculate the importance for each feature it uses (genes) to classify
- Look at the top features (stored in the fit object “feature_importances_”)



```
## feature importance
feature_importances = pd.DataFrame(clf.feature_importances_,
                                   index = all_data.columns,
                                    columns=['importance']).sort_values('importance', ascending=False)
```

- Do the top genes make sense?
  - The feature scoring is known to have a few biases but still very useful

More classifiers

- Choose another two classifiers from the scikit list and repeat
- Are any of them better than the random forest?

Misclassification and plotting

- To study the misclassification calculate the confusion matrix
  - from sklearn.metrics import confusion_matrix
- Run on the prediction and true labels
- Convert to pandas df to add rows and columns names
  - pd.DataFrame(cm, index = data_sets, columns=data_sets)
  - Where does the algorithm do well?
  - Where does it do poorly
- Plot two umaps next to each other using subfigures
  - Colour code one with the true labels
  - Colour the other one with the predicted labels

Extra: genes and proteins

- Purified populations used CD proteins
- Latest data sets measure these proteins as well as genes
- If you classify the cells with the genes does this give the same CD populations?
- Load latest dataset with protein
  - https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_protein_v3?
  -

    Feature / cell matrix (raw) 84.37 MB bee95f96e10b14770bcdeb15147a
- Process the data divide by total reads per cell x 10,000 (separate the proteins)
- Classify cells with classifier
  - Histogram protein levels for each class