Exercise 1

Get familiar with the following functions and packages:

- `linear_model, tree, svm, ensemble, model_selection, metrics` modules of the `scikit-learn` package.
- https://scikit-learn.org/stable/modules/classes.html

Exercise 2

You need to use the `Ex2data1.csv` to fit multiple models which will be used to classify the target variable in column `label`.

- Use only two features `p1, p2`
  - Plot the points with different colours for the different classes.
  - Split the data into a training set and a testing set. In each of the following points, the models are to be fit to the training set, the accuracy is to be analysed on both the training set and testing set.
  - Train the cross-validated logistic regression. Plot the ROC curve, compute the area under the ROC curve for the model.
  - Train a classification tree on the same dataset. Investigate the accuracy when using the full tree. Compare it to the accuracy attained with the logistic regression. Note: given the particular structure of the data, training the full tree may take a while.
  - Plot the ROC curve for the full tree (both for the training and learning set.
- Use all three features `p1, p2, p3`
  - Train a classification tree on the dataset. Investigate the accuracy on both the full tree, and also on the pruned tree. Compare it to the accuracy and ROC curve. Is there any need for pruning? Explain the outcome.
  - Train a random forests on the same dataset. Compare it to the performance attained with the classification tree.

Exercise 3

The `Ex2data2.csv` is a dataset containing sales of child car seats at 400 different stores. You need to develop a few classification models to predict the stores with high sales.

- Create a new feature (column) `High` based on `Sales`. Hint: Sales which are greater than 6, can be classifed as `High`.
- Split the dataset into a training set and a test set.

- Fit a classification tree to the training set. Plot the tree, calculate the confusion matrix and interpret the results. What test accuracy do you obtain?
- Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test accuracy?
- Use random forests to analyze this data. What test accuracy do you obtain? Determine which features are most important.

Exercise 4

- Consider the plane equation
$$\langle w, x\rangle + b = 0.$$
Using the fact that w is normal to the plane, find λ such that $x = \lambda w$ is on the plane. Express the distance of the plane (that is the length of λw) from the origin in terms of $\|w\|$ and $|b|$.
  - Using the previous part, show that the distance of the planes:
$$\langle w, x\rangle + b = 0, \text{ and}$$
$$\langle w, x\rangle + b = 1$$
is $\frac{1}{\|w\|}$, hence the margin is $\frac{2}{\|w\|}$.
  - For any support vector v_i, the following equality holds:
$$u_i(\langle w, v_i\rangle) = 1 - \xi_i,$$
for some non-negative ξ_i. Express the distance of v_i from the plane
$$u_i(\langle w, v_i\rangle + b) = 1$$
in terms of ξ_i and $\|w\|$.

Exercise 5

- You need to use the `Ex2data3.csv` to fit SVM models which will be able to classified target variable in column V1.
  - Split the data-set into training set and testing set.
  - Run SVM using the following kernels: `linear`, `rbf`, `sigmoid`. Use the default value for `C`. Generate the contour plot of classification, evaluate the accuracy of the classification on the testing set. How would you make a choice between the kernels?
  - Pick one of the kernels, and run SVM using cross-validation. Use the different value for `C`. Generate the contour plot of classification, evaluate the accuracy of the classification on the testing set. How would you make a choice of different value for `C`?