**Goal:** The goal of this homework is to practically study the metrics of classifier accuracy.

**Assignment:**

The data set for this assignment can be found at the UCI Machine Learning Repository, at [http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#](http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+)

Download  “**Occupancy Detection Data Set**”. The class label in this dataset is 0/1, based on whether an office was occupied or not.

Develop a simple rule-based classifier, that will work similarly to the fish classifier discussed in lecture. Use one feature a time. Using the data in “datatraining.txt” and K-fold cross validation, find the optimal decision boundary that minimizes classification error. The K-fold cross validation should be implemented without  using  Matlab’s  built-in  functions such as “cvpartition”. This means all code must be yours!

Compare the classifiers in terms of error measures shown on slide 24 (“Measuring errors”). Compute ROC and AUC for each classifier. Determine the best classifier and corresponding error.

Now use data from “datatest.txt” and apply your best classification model. Assess classification errors. Provide conclusions about the generalization capabilities of the developed classification model.

* + **Graduate only assignment**

Using the same data set you used in Problem 1 and a “classificationLearner” in MATLAB, implement an algorithm to evaluate the importance of each feature in the training set.

Train a model which includes all features. Then, the importance of each feature may be evaluated by reducing the entire training set to only a single feature at the time and then constructing  a  predictor  and  calculating  the  area  under  the  ROC  curve  (using  cross-validation).

Those features with the largest AUC should be considered the most important. Provide the top several features  for  each  of  the  classification  models  above  and  comment  on  the  differences.  Finally, look  at  the  best  features  and  comment  on  whether  your  results  make  sense  in  terms  of  what  you originally thought would be important features.

Classification learner: <https://www.mathworks.com/help/stats/classification-learner-app.html>