Evaluation in Machine Learning

0. Work through `05-Evaluation`; `05-Sampling-Methods` and `05-ROC` notebooks.
1. The *confusion matrix* below shows the evaluation results for a binary classifier when applied to a test set of 768 examples, which are annotated with the class labels: (Pass, Fail).

Predicted Class

Pass	Fail		
407	93	Pass	Actual Class
108	160	Fail	

From this table calculate:

a) The *precision* score for both of the classes.

b) The *recall* score for both of the classes.

c) The *F1-measure* score for both of the classes.

d) The *overall classification accuracy* for the full test set.

e) The *average class accuracy* measure for the full test set

2. The table below shows the true class labels for a test set of 12 emails, which are labelled as "spam" or "non-spam". The table also reports the predictions made by three different binary classifiers for those emails.

Example	True Class Label	KNN Prediction	J48 Prediction	SVM Prediction
1	spam	spam	spam	spam
2	non-spam	non-spam	spam	non-spam
3	spam	non-spam	non-spam	spam
4	non-spam	non-spam	non-spam	non-spam
5	spam	spam	spam	spam
6	non-spam	non-spam	non-spam	non-spam
7	non-spam	spam	spam	non-spam
8	non-spam	non-spam	spam	spam
9	spam	spam	non-spam	spam
10	spam	spam	non-spam	non-spam
11	spam	non-spam	non-spam	spam
12	spam	spam	spam	spam

a) Calculate the *overall accurac*y for each of the classifiers on this data. Based on your calculations, which classifier the most accurate?

b) Calculate the *precision* of each classifier relative to the "spam" class. Based on your calculations, which classifier achieves the highest precision for this class?

c) If you had to recommend a classifier for this dataset, which would you recommend and why?

3. The table below shows the number of correct and incorrect predictions made by an image classifier during a 10-fold cross validation experiment, where the goal was to classify 5,000 images into one of three categories: {cats, dogs, people} (i.e. each test set contains 500 images).

Fold	**Class: Cats**		**Class: Dogs**		**Class: People**	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	82	68	82	68	164	36
2	81	69	102	48	176	24
3	99	51	97	53	160	40
4	81	69	102	48	148	52
5	94	56	99	51	148	52
6	97	53	91	59	162	38
7	81	69	94	56	148	52
8	76	74	79	71	181	19
9	76	74	97	53	160	40
10	96	54	79	71	179	21

a) What is the *overall accuracy* of the classifier based on the cross-validation results?

b) What conclusion might be draw about the different classes in the data, based on the results above?

c) Would *leave-one-out cross validation* be an appropriate evaluation strategy on this dataset?

4. The notebook `05 ROC Exercise` contains code for loading the diabetes dataset `(diabetes.csv).`

   a) Using the code in `05 ROC` as a template produce ROC curves for kNN, SVM and Naive Bayes classifiers on the diabetes data.

   b) Repeat this exercise using synthetic data generated using the code below. What insights do these ROC curves provide?

```
from sklearn.datasets import make_classification
X, y = make_classification(n_samples=1000, n_features=4, n_classes=2
                           class_sep=0.75, random_state=1)
```