CONDUCTING EXPERIMENTS

Shingchern D. You

Conducting experiments

- Training/Testing or Training/validation/testing
- **Cross validation** (10-folds or 5-folds)
  - Partition training set into 10 subsets $(A_1, \cdots, A_{10})$ with equal samples (cardinality)

  For i = 1.. 10

  Use A_i as test set and the rest subsets as training set

  Compute and report average accuracy

Accuracy in two classes

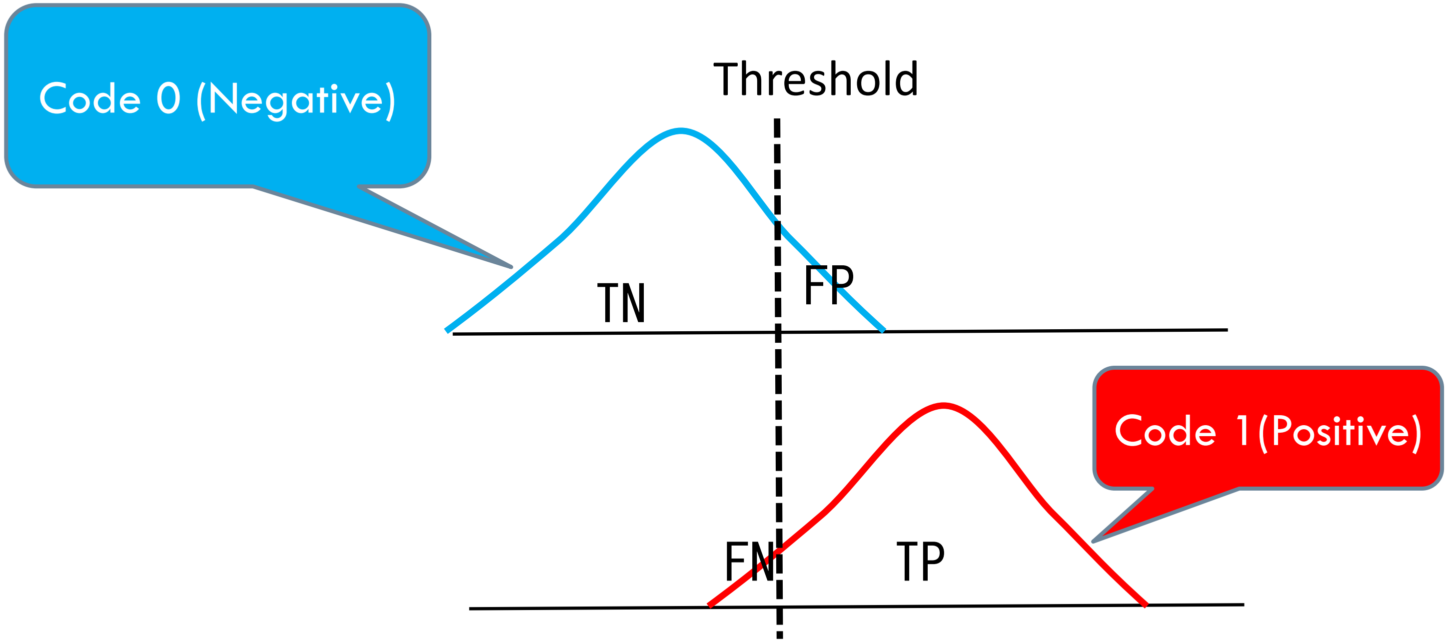
- Simplest case
  - Treat false positive & false negative equally weighted
  - Report accuracy
- Want to distinguish false positive & false negative
  - Errors not equally weighted
  - For medical reports (usually class imbalance)
  - More insights in error analysis
- Ref: (https://en.wikipedia.org/wiki/Precision_and_recall)

Accuracy in two classes

- P (condition positive): actual positive cases in the data
- N (condition negative): actual negative cases in the data
- TP (true positive): predicted positive & real positive
- TN (true negative): predicted negative & real negative
- FP (false positive, false alarm, type I error): number of negative cases predicted as positive
- FN (false negative, miss, type II error): number of positive cases predicted as negative

Accuracy in two classes

- Consider the BPSK problem again
- “0” is in (-2,1), “1” is in (-1,2)



Sensitivity vs Specificity (medical)

- **Sensitivity**, recall, hit rate, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

- **Specificity**, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

- *Fall-out or false positive rate (FPR)*

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

- *Miss rate or false negative rate (FNR)*

$$FNR = \frac{FP}{P} = \frac{FP}{TP + FN} = 1 - TPR$$

Precision vs recall

- **Precision** $= \frac{TP}{TP+FP}$
- **Recall** $= \frac{TP}{TP+FN}$ (same as sensitivity)
- Accuracy $= \frac{TP+TN}{P+N}$
- Some papers also use F_1-measure
- $F_1 = 2\,\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$
- What is the range of F_1

Numerical example in COVID-19

- All patients: positive 1% & negative 99%
- Test kit with sensitivity 30% & specificity 95%
- TP = P × TPR = 0.3 %
- TN = N × TNR = 94.05 %
- FP = 99% - 94.05% = 4.95 %
- FN = 1% - 0.3% = 0.7 %
- Precision $= \frac{0.3}{0.3+4.95} = 5.71\%$, Recall = 30%
- $F_1 = 2\frac{0.0571\times 0.3}{0.0571+0.3} = 0.096$

Numerical example

- Tossing a coin to determine positive or negative
- FP=99 %/2 = 49.5%
- FN=0.5%
- Precision $= \frac{0.5}{0.5+49.5} = 1\ \%$
- Recall = 50 %
- $F_1 = 2\frac{0.01\times0.5}{0.01+0.5} = 0.020$
- Therefore, test kit is slightly better in F_1-measure

Binary classification with threshold

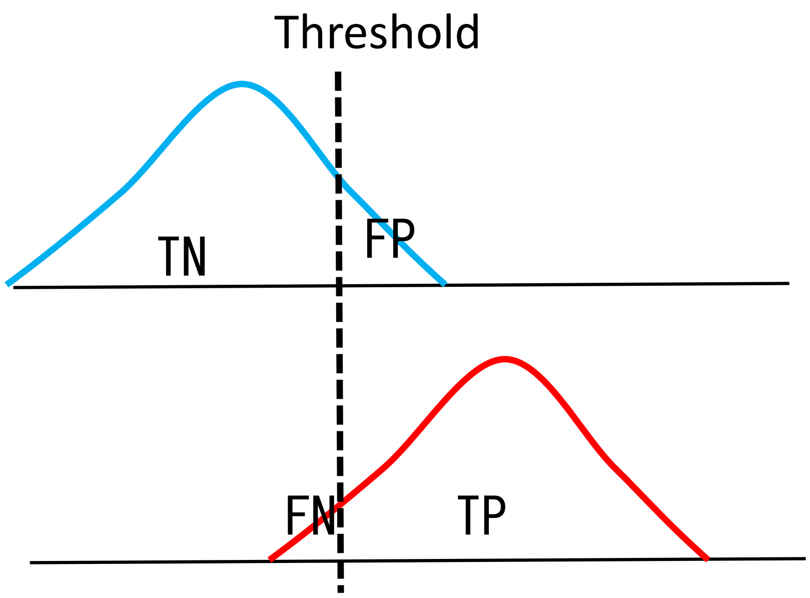
- Classifier produces values in [0,1] (continuous) instead of binary
  - If classifier output > threshold: class 1
  - Else class 2
- Detect out-of-database samples
- How to compare accuracy between two classifiers
  - Unfair comparison if threshold not optimized
  - Want to use curves for fair comparison

ROC curve for binary classification

- Receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a **binary classifier** system as its discrimination **threshold is varied** –Wiki
- Plotting the **true positive rate** (TPR, in Y axis) against the **false positive rate** (FPR, in X axis) at various threshold settings
- AUC: area under curve (usually ROC AUC)

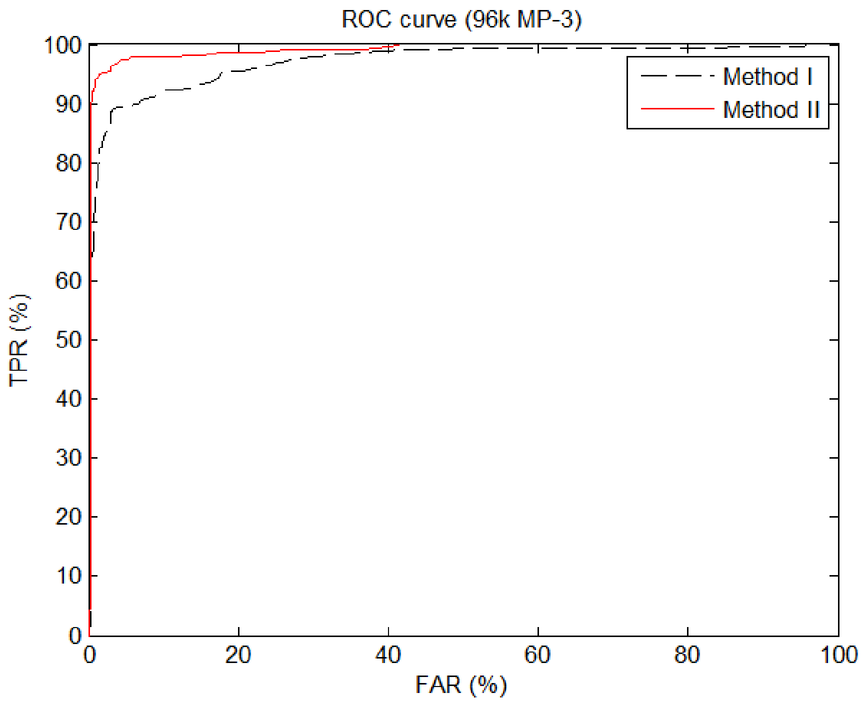
ROC curve for binary classification

- Consider the BPSK problem again
- Moving threshold toward left increases TPR, but also increases FPR (FAR)



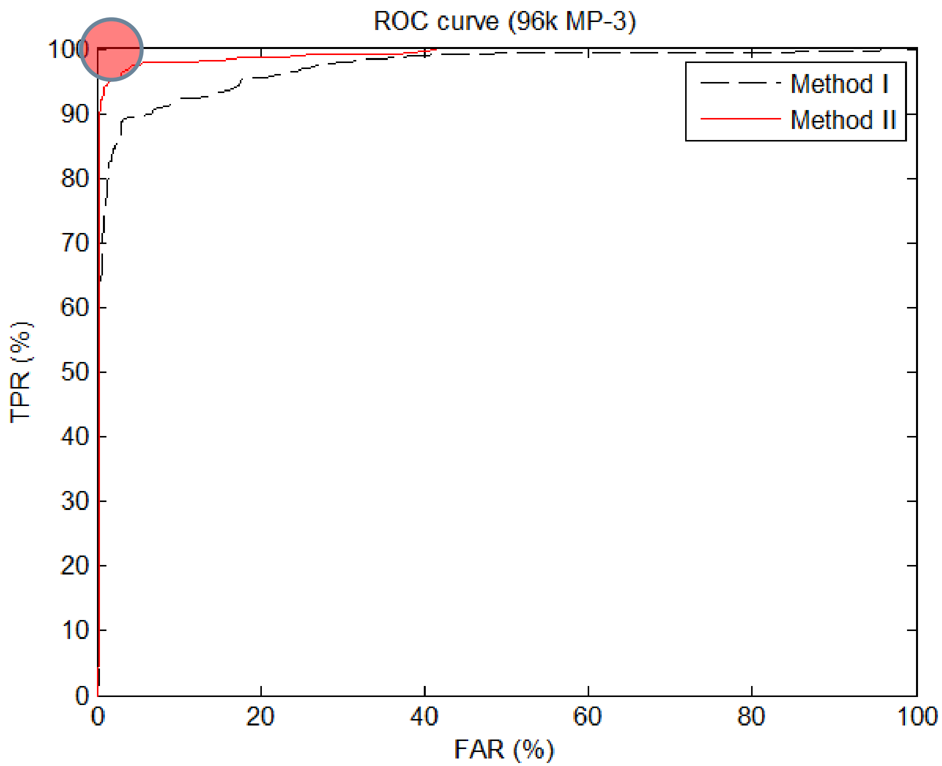
ROC curve for binary classification

- As threshold moves toward left (previous picture), TPR ↑, but FAR aslo ↑
- Which method is better in plot below



ROC curve for binary classification

- The left-upper corner is the best case (why?)
- A curve closer to this corner is better (method II is better)

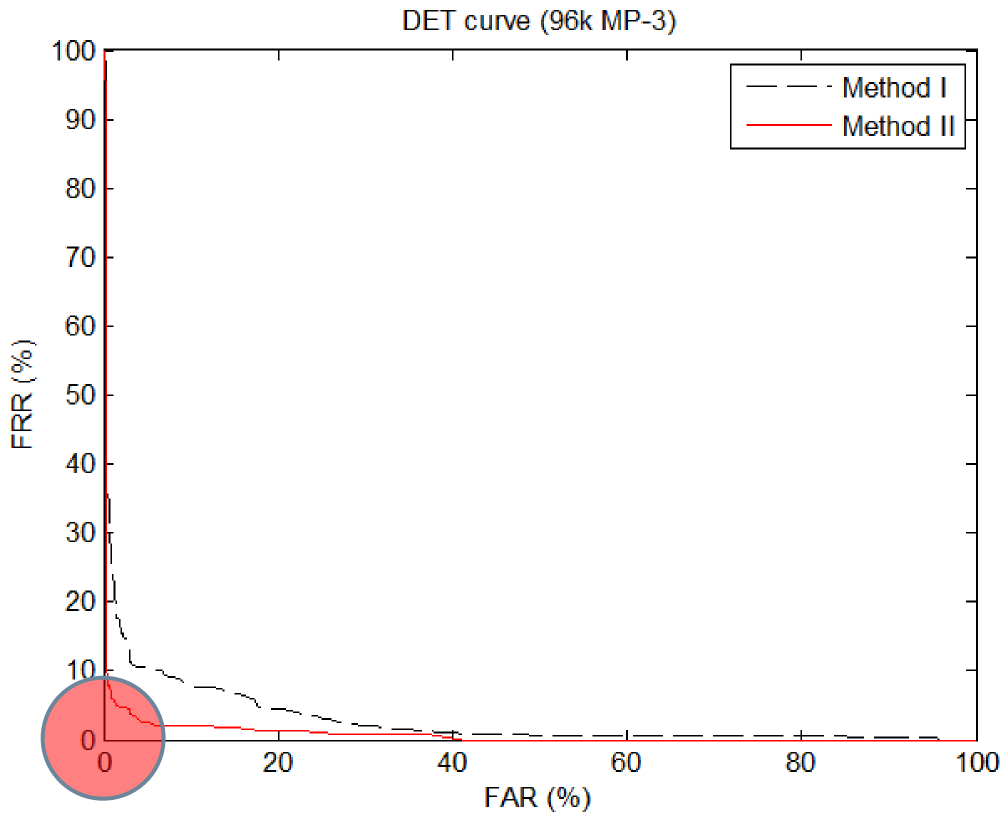


DET curve for binary classification

- Ref: A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET curve in assessment of detection task performance,”I n Proceedings of the Eurospeech, vol.4, pp.1899– 1903, Rhodes, Greece, September1997
- DET is also widely used, like ROC
- A detection error tradeoff (DET) graph is to plot **false rejection rate (Y axis) vs. false acceptance rate (X axis)**
- DET curve usually uses **log** scales in both axes (to make the curves more linear)
  - Shortcoming of using log: origin (0,0) undefined

DET curve without log

- Left-lower corner is best (thus, method II is better)

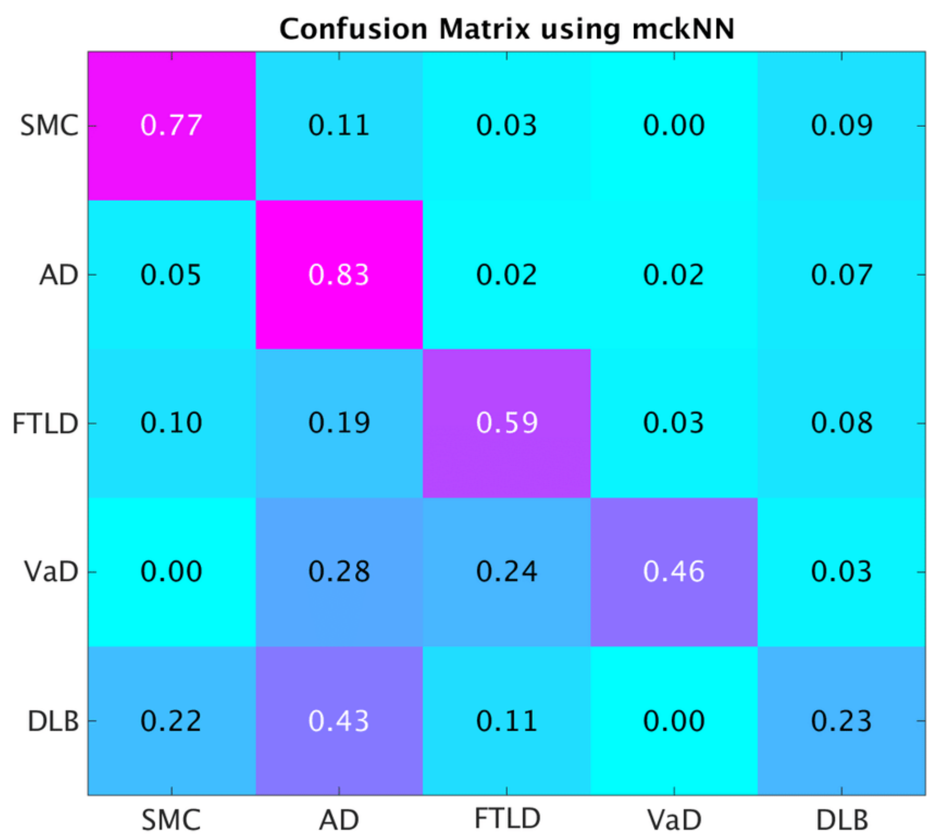


Confusion matrix

- Use TPR, TNR, FPR, & FNR for binary classification
- Use confusion matrix for multiclass classification
- Put "actual" class in vertical and "predicted" class in horizontal (or vice versa)
- Fill in percentage of $\left(\frac{b_j}{a_i}\right)$ in each cell
  - S_i: set of test samples in class i
  - a_i: $|S_i|$ (i.e., total # of elements in the set)
  - b_j: elements of S_i predicted as class j

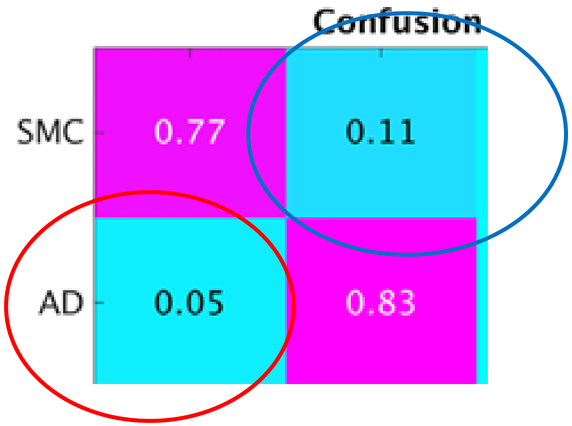
Confusion matrix example

- https://www.researchgate.net/figure/Confusion-matrix-of-the-classification-results-using-different-classifiers_fig3_317547458



Confusion matrix example

- How to read this matrix
- For actual SMC, 77% of samples are correctly predicted as SMC
- Thus, diagonal values are more important
- Matrix may not be symmetric (like this example)



Confusion matrix example

- Guess labels in which axis represents “actual” class
  - Sum over all predicted percentages is 100%
  - Y axis (so add numbers in horizontal direction to 100%)
- In this example, which class Z is hard to classify
  - DLB
- If predicted as DLB, sample is likely from class DLB
- If sample in class DLB, it has 43% change predicted as AD, i.e., P(predict as AD | DLB) = 0.43
  - Compute P(DLB|predict as AD)