ROC Analysis of Diagnostic Tests

BSTT 538 Class Presentation, UIC

Shujun Xu, MS Candidate in Biostatistics
Epidemiology and Biostatistics, School of Public Health
University of Illinois at Chicago

December 6, 2019

Table of Contents

Origin of the Term

The term ***Receiver Operating Characteristic*** has its roots in World War II. ROC curves were originally developed by the British as part of the "Chain Home" radar system. ROC analysis was used to analyze radar data to differentiate between enemy aircraft and signal noise (e.g. flocks of geese). As the sensitivity of the ***receiver*** increased, so did the number of false positive (in other words, specificity wend down).

Timeline of Application

- Originated in the early 1950's with electronic *Signal Detecting Theory (SDT)*
- Introduced to psychology in the early 1950's
- Adapted into diagnostic radiology and radionuclide imaging in early 1960's
- Has been used in medicine, biometrics, forecasting of natural hazards, meteorology, model performance assessment, and other areas for many decades and is increasingly used in machine learning and data mining research.

Signal Detection Theory (SDT)

SDT is a means to measure the *ability to differentiate* between information-bearing patterns (called *stimulus* in living organisms, *signal* in machines) and random patterns that distract from the information (called *noise*, consisting of background stimuli and random activity of the detection machine and of the nervous system of the operator).

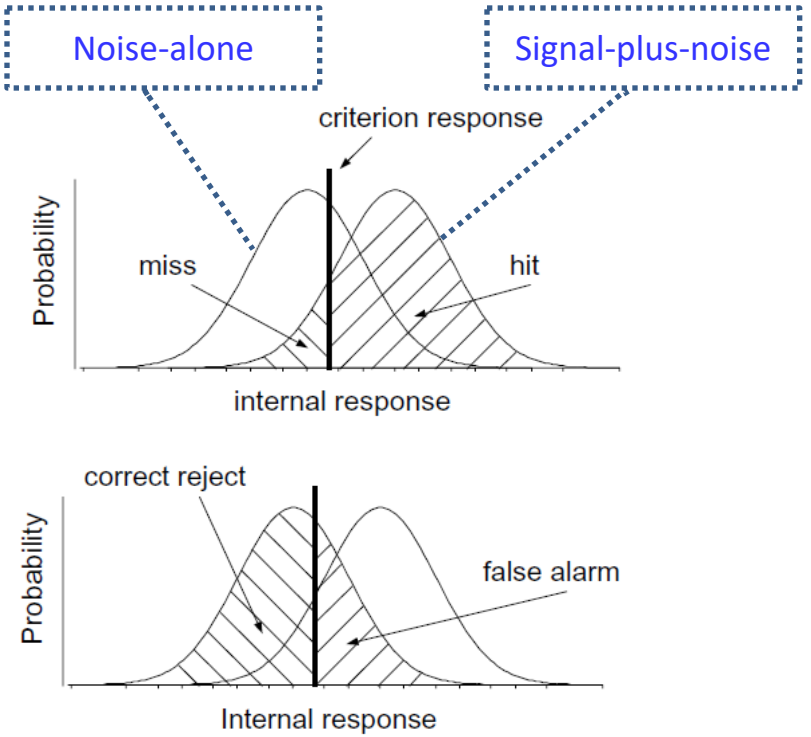
Basic concepts include : *Hits, False Alarms, Criterion, ROC curves , and d'.*

Signal Detection Theory (SDT) - Illustration

Medical Scenario : A radiologist examining a CT scan

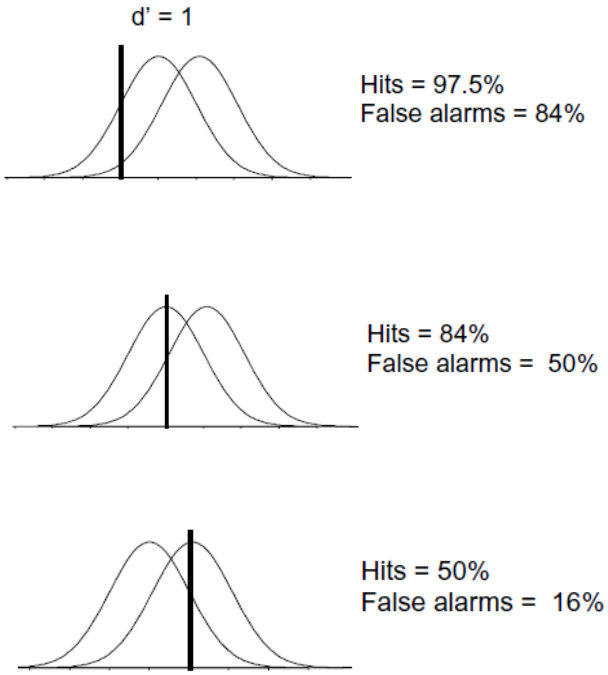
Two main components to
the decision-making process :
Information acquisition and *criterion*

True Tumor Status		Doctor's Opinion	
		Yes	No
Present		Hit	Miss
		(TP)	(FN)
Absent		False Alarm	Correct Rejection
		(FP)	(TN)

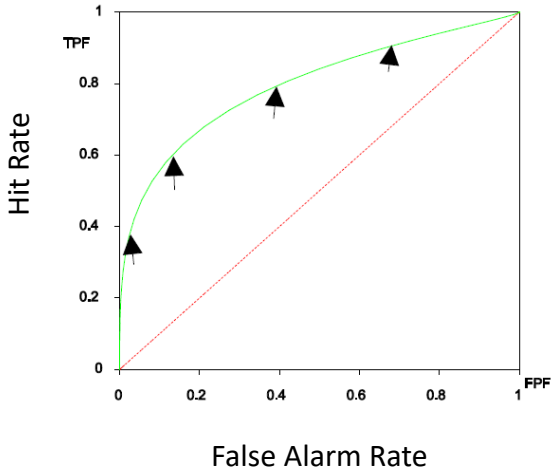


Signal Detection Theory (SDT) – ROC Curve

- The Role of the Criterion

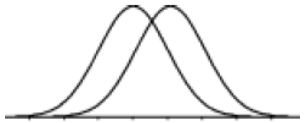


We can describe the full range of the doctor's opinions in a single curve, called an *ROC curve*.

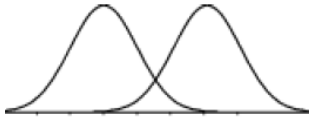


Signal Detection Theory (SDT) – Discriminability Index (*d'*)

- The Role of Information

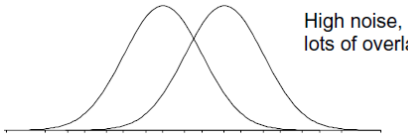


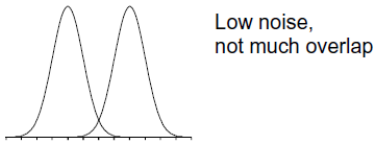
d' = 1 (lots of overlap)



d' = 3 (not much overlap)

- Varying Noise

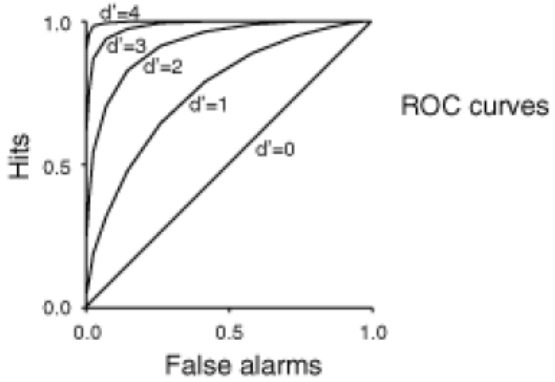




d' = separation / spread

Its value does not depend upon the criterion.

Discriminability is made easier either by increasing the separation (stronger signal) or by decreasing the spread (less noise) . In either case, there is less overlap between the curves.



Measures of diagnostic accuracy – simple statistics

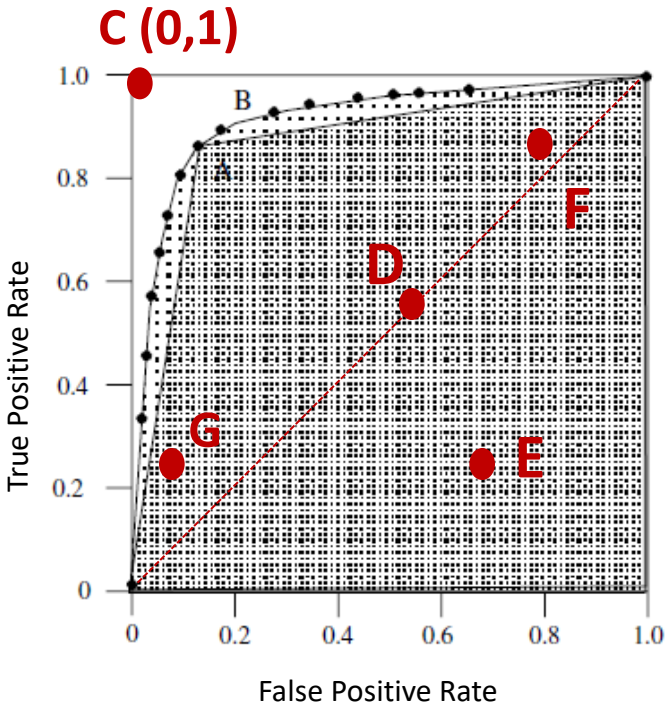
True Disease Status	Test Results	
	Yes	No
Present	TP	FN
Absent	FP	TN

Conventionally a two-by-two confusion matrix (also called *contingency table*) can be constructed in the case of a *binary predictor*.

Sensitivity = Ture Positive Rate = TP/(TP+FN) = 1-FNR
Specificity = Ture Negative Rate = TN/(TN+FP) = 1-FPR
Positive Predictive Value (PPV) = TP/(TP+FP)
Negative Predictive Value(NPV) = TN(TN+FN)
Likelihood Ratio = Sensitivity / (1-Specificity)

(Note: LR is independent of prevalence of the disease, LR=1 indicates that the test result is equally likely in patients with and without the disease)

Measures of diagnostic accuracy – ROC Curve



Evaluating the discriminating power of the diagnostic test can be achieved by a ROC curve in the case of predictors measured on a *continuous or ordinal scale*. A ROC graph depicts relative trade-offs between benefits (TP) and costs (FP).

- A "separator" (or decision) variable
- Discrete classifiers and scoring classifiers
- ROC Space (diagonal line=chance level)
- Slope at any point (=LR)
- AUC (Area Under the Curve)

AUC and Hypothesis Testing

The AUC can be interpreted as the probability that a randomly chosen diseased subject is rated or ranked as more likely to be diseased than a randomly chosen non-diseased subject. The other interpretation is the average value of sensitivity for all the possible values of specificity. "In a summary sense, *the greater the area under the ROC curve, the better the predictions*."

- Identical to the value of another measure of predictive power, the concordance index (C-statistic); Calculation based on nonparametric Mann-Whitney U statistics; Equivalent to the Wilcoxon test of ranks (Henley and McNeil, 1982)
- Closely related to the Gini coefficient (Breiman et al., 1984) ;

  Gini +1 = 2 x AUC (Hand and Till, 2001)

AUC and Hypothesis Testing

- In general, a value of **0.5** for AUC indicates that the ROC curve will fall on the diagonal and hence suggests that the diagnostic test has no discriminatory ability; An AUC at **0.7 to 0.8** is considered acceptable, **0.8 to 0.9** is considered excellent, and **more than 0.9** is considered outstanding.

$$H_0 : AUC = 0.5 ;$$
$$H_1 : AUC \neq 0.5 ;$$
$$Z = \frac{\widehat{AUC} - 0.5}{SE\ (\widehat{AUC})}$$

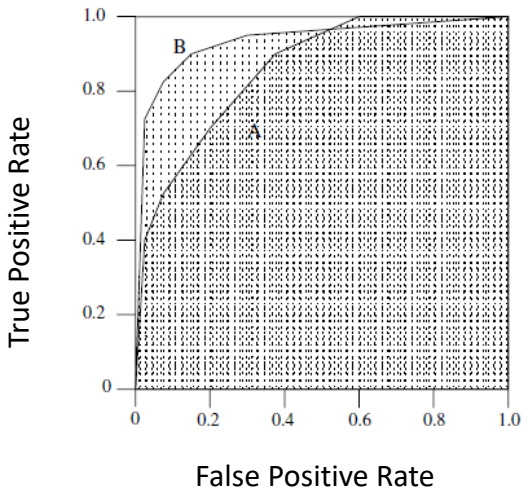
- Testing the accuracy in comparative study of two diagnostic tests: Correlated ROC curves (same subjects) and Uncorrelated ROC (subjects from different groups)

$$H_0: AUC_1 = AUC_2$$
$$H_1: AUC_1 \neq AUC_2$$
$$Z = \frac{\widehat{AUC_1} - \widehat{AUC_2}}{SE(\widehat{AUC_1} - \widehat{AUC_2})}$$

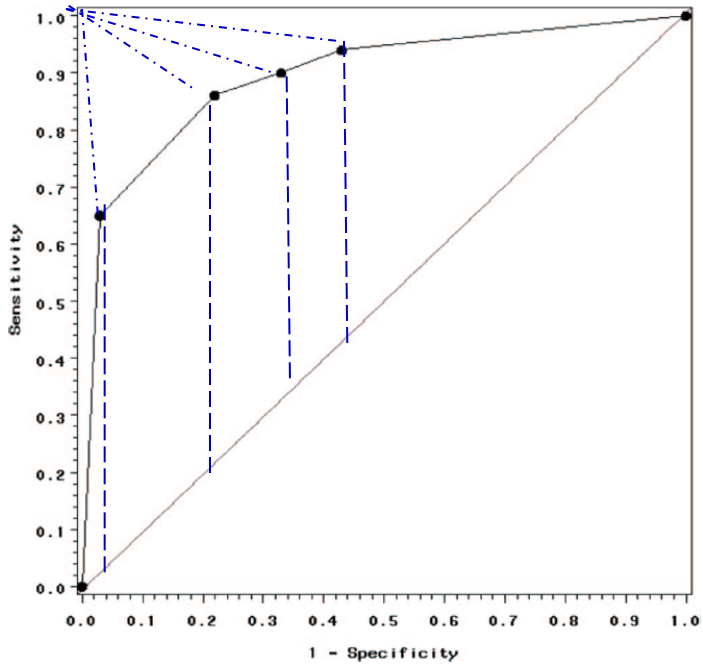
Comparison of Two or More Diagnostic Systems

Three important commonly used indices :

- *AUC :* desirable to compare the entire ROC curve rather than at a particular point
- *Partial Area Index* : a partial area under the curve corresponding to a clinical relevant FPR is recommended as an index of choice (the right side of the unit square is of no clinical relevance)
- *TPR for a given FPR :* especially in the case where two ROC curves cross and the area under the curves may be equal.



Optimal Cut-off value



In determining optimal cut off values, at least three methods have been proposed:

1. Minimize $d^2 = (1\text{-TPR})^2 + \text{FPR}^2$

2. Maximize *Youden* Index =TPR – FPR

3. Incorporates the financial costs for correct and false diagnosis and the costs of further work up for diagnosis.

Effect of Class Skew

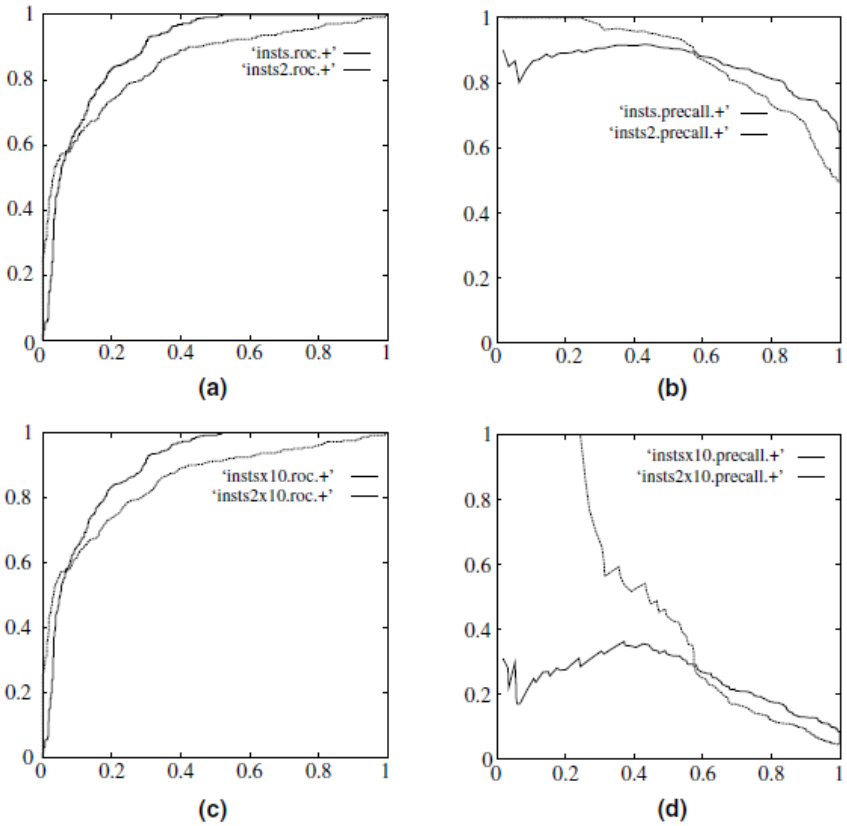
ROC curves have an attractive property: they are *insensitive* to changes in class distribution.

Recall: Let *p* =prevalence of disease

Lower p -> lower PPV , higher NPV

$$\text{PPV} = \frac{p * Sen}{p * Sen + (1-p) * Spe}$$

$$\text{NPV} = \frac{(1-p) * Spe}{p * (1-Sen) + (1-p) * Spe}$$



ROC Curves

PPV – sensitivity curves

Advantages of ROC Analysis (summary)

ROC curve analysis has several advantages :

- In contrast to single measures of sensitivity and specificity, the diagnostic accuracy is not affected by decision criterion and it is also independent of prevalence of disease since it is based on sensitivity and specificity.
- Several diagnostic tasks on the same subjects can be compared simultaneously in a ROC space.
- One can easily obtain the sensitivity at specific FPR by visualizing the curve.
- The optimal cut- off value can be determined using ROC curve analysis.

SAS Example (DeLong et al., 1988)

Testing whether the area under the ROC curve differs from 0.5 (chance)

```
data roc;
   input alb tp totscore popind @@;
   totscore = 10 - totscore;
   datalines;
3.0 5.8 10 0   3.2 6.3  5 1   3.9 6.8  3 1   2.8 4.8  6 0
3.2 5.8  3 1   0.9 4.0  5 0   2.5 5.7  8 0   1.6 5.6  5 1
3.8 5.7  5 1   3.7 6.7  6 1   3.2 5.4  4 1   3.8 6.6  6 1
4.1 6.6  5 1   3.6 5.7  5 1   4.3 7.0  4 1   3.6 6.7  4 0
2.3 4.4  6 1   4.2 7.6  4 0   4.0 6.6  6 0   3.5 5.8  6 1
3.8 6.8  7 1   3.0 4.7  8 0   4.5 7.4  5 1   3.7 7.4  5 1
3.1 6.6  6 1   4.1 8.2  6 1   4.3 7.0  5 1   4.3 6.5  4 1
3.2 5.1  5 1   2.6 4.7  6 1   3.3 6.8  6 0   1.7 4.0  7 0
3.7 6.1  5 1   3.3 6.3  7 1   4.2 7.7  6 1   3.5 6.2  5 1
2.9 5.7  9 0   2.1 4.8  7 1   2.8 6.2  8 0   4.0 7.0  7 1
3.3 5.7  6 1   3.7 6.9  5 1   3.6 6.6  5 1
;
ods graphics on;
ods html;
proc logistic data=roc;
   model popind(event='0') = alb;
   roc; roccontrast;
   run;
```

ROC Association Statistics

ROC Model	Mann-Whitney Area	Mann-Whitney Standard Error	Mann-Whitney 95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Model	0.7366	0.0927	0.5549	0.9182	0.4731	0.4809	0.1949
ROC1	0.5000	0	0.5000	0.5000	0	.	0

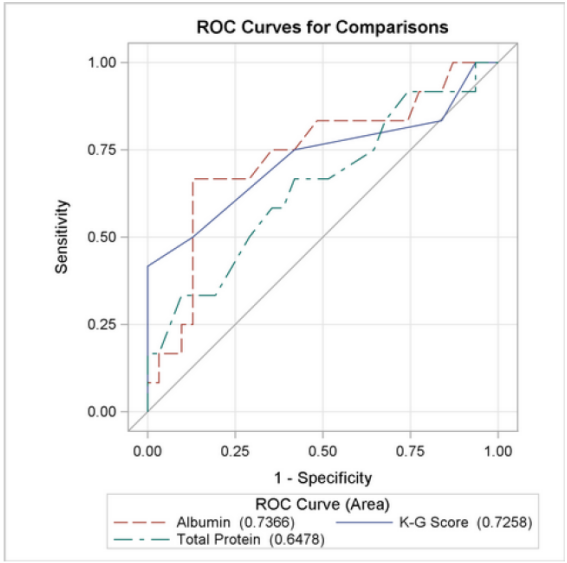
ROC Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	6.5129	0.0107

SAS Example (DeLong et al., 1988)

Comparing two or more *correlated (dependent)* ROC Curves

```
ods graphics on;
proc logistic data=roc plots=roc(id=prob);
   model popind(event='0') = alb tp totscore / nofit;
   roc 'Albumin' alb;
   roc 'K-G Score' totscore;
   roc 'Total Protein' tp;
   roccontrast reference('K-G Score') / estimate e;
run;
ods graphics off;
```



ROC Association Statistics

ROC Model	Area	Mann-Whitney Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Albumin	0.7366	0.0927	0.5549	0.9182	0.4731	0.4809	0.1949
K-G Score	0.7258	0.1028	0.5243	0.9273	0.4516	0.5217	0.1860
Total Protein	0.6478	0.1000	0.4518	0.8439	0.2957	0.3107	0.1218

ROC Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
Reference = K-G Score	2	2.5340	0.2817

Other Issues – In the Absence of a Gold Standard

- In medicine and statistics, a gold standard test is usually the diagnostic test or benchmark that is the best available under reasonable conditions. Other times, a gold standard is the most accurate test possible without restrictions.
- In order to obtain an unbiased estimator for the test accuracy, the true disease status for each patient (present or absent, and independent of the patient's test result) needs to be determined. The procedure that establishes the patient's true disease status is referred to as a gold standard.
- Hall and Zhou (2003) proposed a non-parametric estimator for the ROC curves of continuous-scale tests under the conditional independence assumption when the number of tests is more than two.

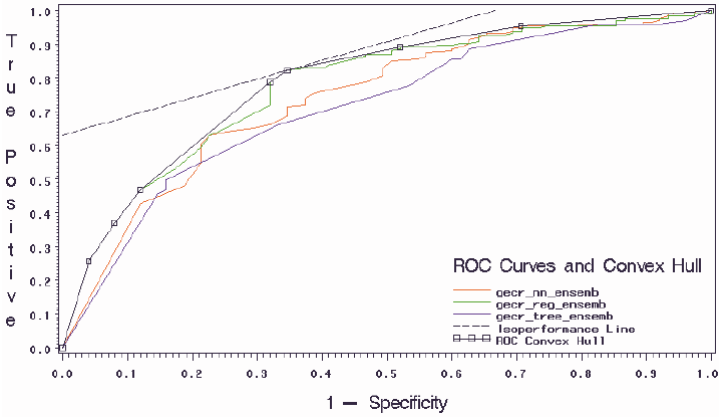
Further Reading: Non-Parametric Estimation of ROC Curves in the absence of a Gold Standard (Zhou XH, Castelluccio P, Zhou C., 2004)

Other Issues – The ROC convex hull (ROCCH)

In search for the optimal classifier that is robust with respect to skewed or imprecise class distributions and disparate misclassification costs, Provost and Fawcett (1998, 2001) show that a set of operating conditions may be transformed easily into a so-called *iso-performance line* in ROC space.

Two points in ROC space,
(FP1,TP1) and (FP2,TP2),
have the same performance if

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{c(\boldsymbol{Y},\boldsymbol{n})p(\boldsymbol{n})}{c(\boldsymbol{N},\boldsymbol{p})p(\boldsymbol{p})} = m$$



All classifiers corresponding to points on a line of slope *m* have the same expected cost. More generally, a classifier is potentially optimal if and only if it lies on the convex hull of the set of points in ROC space.

Other Issues – Rating Scale Version

- Rating scales (also called assessment scale) are used to elicit data about quantitative entities. Often, predictability of rating scales (also called ‘‘assessment scales’’) could be improved.
- Rating scales often use values: ‘‘1 to 10’’ and some rating scales may have over 100 items (questions) to rate.
- Other popular terms for rating scales are: survey and questionnaire.
- Rating itself is very popular on the Internet for ‘‘Customer Reviews’’ where often uses five stars (e.g., by Amazon.com) instead of ordinal numbers.

Further Reading: How to reduce the number of rating scale items without predictability loss? (Koczkodaj et al., 2017)

Other Issues – Creating Scoring Classifiers

- Many discrete classifier models may easily be converted to scoring classifiers by ‘‘looking inside’’ them at the instance statistics they keep.
- Even if a classifier only produces a class label, an aggregation of them may be used to generate a score. MetaCost (Domingos, 1999) employs bagging to generate an ensemble of discrete classifiers, each of which produces a vote. The set of votes could be used to generate a score.
- Some combination of scoring and voting can be employed. For example, rules can provide basic probability estimates, which may then be used in weighted voting (Fawcett, 2001).

Research Question: Implications of Various “Voting” standards ?

Other Issues – Multi-reader ROC Analysis

- Assume that we are considering diagnostic imaging studies where there are multiple readers (e.g., radiologists) who assign to each case (i.e., patient) a disease severity or disease likelihood rating based on the corresponding image or set of images, acquired using one or more imaging modalities. For studies where there is *Reader Variability*, the researcher typically will prefer to have conclusions apply to both the reader (can be treated as random sample) and case populations.
- Traditionally, multi-reader receiver operating characteristic (ROC) studies have used a "paired-case, paired-reader" design. The Dorfman-Berbaum-Metz (DBM) method has been one of the most popular methods for analyzing multi-reader ROC studies since it was proposed in 1992.

Further Reading: Recent Developments in the Dorfman-Berbaum-Metz Procedure for Multireader ROC Study Analysis (Stephen L. Hillis, Kevin S. Berbaum and Charles E. Metz, 2008)

REFERENCES

- Hajian-Tilaki P. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med.* 2013; 4(2): 627-635.
- Heeger D. Signal Detection Theory. New York University website. https://www.cns.nyu.edu/~david/handouts/sdt/sdt.html. 2003-2007. Accessed December 1, 2019.
- Mandrekar J N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J Thorac Oncol.* 2010; 5:1315–1316.
- Mandrekar J N. Simple Statistical Measures for Diagnostic Accuracy Assessment. *J Thorac Oncol.* 2010; 5: 763–764.
- Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27:861-874. doi:10.1016/j.patrec.2005.10.010.

REFERENCES

- Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology.* 1982; 143: 29-36.
- Koczkodaj WW, Kakiashvili T, Szymańska A, et al. How to reduce the number of rating scale items without predictability loss? *Scientometrics.* 2017; 111: 581–593. DOI 10.1007/s11192-017-2283-4.
- Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. *Stat Med.* 2012; 31(23): 2577–2587. doi:10.1002/sim.5328.
- Zhou XH, Castelluccio P, Zhou C. Non-Parametric Estimation of ROC Curves in the Absence of a Gold Standard. *UW Biostatistics Working Paper Series.*2004. Working Paper 231. http://biostats.bepress.com/uwbiostat/paper231. Accessed December 1, 2019.

REFERENCES

- Stephanie. Receiver Operating Characteristic (ROC) Curve: Definition, Example. Statistics How To website. https://www.statisticshowto.datasciencecentral.com/receiver-operating-characteristic-roc-curve/. August 27, 2016. Accessed December 1, 2019.
- Wikipedia. Detection theory. Wikipedia website. https://en.wikipedia.org/wiki/Detection_theory#cite_note-3. Accessed December 1, 2019.
- Web Interface for Statistics Education. Signal Detection: Overview. WISE website. http://wise.cgu.edu/wise-tutorials/tutorial-signal-detection-theory/signal-detection-overview-2/. Accessed December 1, 2019.