Metadata of the chapter that will be visualized in SpringerLink

Book Title	Biomedical and Other Applications of Soft Computing	
Series Title		
Chapter Title	Accuracy Measures and the Convexity of ROC Curves for Binary Classification Problems	
Copyright Year	2023	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	

Corresponding Author	Family Name	**Phuong**
	Particle	
	Given Name	**Le Bich**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Hanoi University of Mining and Geology
	Address	Bac Tu Liem, Vietnam
	Email	lebichphuong@humg.edu.vn
Author	Family Name	**Zung**
	Particle	
	Given Name	**Nguyen Tien**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Université de Toulouse and Torus Actions SAS
	Address	Toulouse, France
	Email	tienzung@math.univ-toulouse.fr
		ntzung@torus-actions.fr

Abstract	In this work, we offer a theoretical explanation of the convexity of the ROC (receiver operating characteristic) curves for rational binary classifiers, and show some natural important inequalities relating different measures of accuracy.

Accuracy Measures and the Convexity of ROC Curves for Binary Classification Problems

Le Bich Phuong and Nguyen Tien Zung

Abstract In this work, we offer a theoretical explanation of the convexity of the ROC (receiver operating characteristic) curves for rational binary classifiers, and show some natural important inequalities relating different measures of accuracy.

1 Introduction

In this work, we are interested in the accuracy measures of binary classifiers in the context of artificial intelligence. More specifically, we will be interested in the ROC (receiver operating characteristic) curves and their corresponding AUC (area under the curve), the maximal balanced accuracy and the maximal cost-weighted accuracy, and inequalities relating these accuracy measures to each other.

The main results presented in this paper may be summarized as follows:

- (Section 2) ROC curves of optimal machines are convex.
- (Section 3) Natural inequalities relating different measures of accuracy. In particular, if the machine is rational then the AUC is greater than the balanced accuracy.

Convexity of the ROC curve is not something new, and many research papers and monographs already discussed this convexity property in an empirical way, see, e.g., [1–6] and references therein. However, the only place we we found a rigorous theorem on convexity of the ROC curve is [2, Theorem 3], where the authors showed that the convexity of the ROC curve is equivalent to another natural condition, namely "the conditional event probability (or the likelyhood ratio) is nondecerasing". (The higher the "sigmoid value" is, the higher the probability of the event being true is). In this paper, we study *optimal* machines (see Definition 1), and show that if a machine

L. B. Phuong (✉)
Hanoi University of Mining and Geology, Bac Tu Liem, Vietnam
e-mail: lebichphuong@humg.edu.vn

N. T. Zung
Université de Toulouse and Torus Actions SAS, Toulouse, France
e-mail: tienzung@math.univ-toulouse.fr; ntzung@torus-actions.fr

N. Hoang Phuong and V. Kreinovich (eds.), *Biomedical and Other Applications of Soft Computing*, Studies in Computational Intelligence 1045,
https://doi.org/10.1007/978-3-031-08580-2_15

is optimal then its ROC curve is automatically convex. Our proof is based on an explicit construction of optimal machines. We then observe that for such machines the conditional event probability is automatically nondecreasing, which implies (as in [2, Theorem 3]) that the ROC curve is convex. To make our paper self-contained, we also write down explicitly a proof for this last implication, which takes only a few lines.

Since, by machine learning, most reasonable constructed machines tend to become "nearly optimal", our result gives an explanation why most ROC curves met in practice are "nearly convex".

The inequalities that we present in Sect. 3 are elementary, but they are important, and we have not seen them written down explicitly anywhere in the literature, and that's why we want to present them in this paper.

Our motivation for studying the ROC curves and the "most rational classifiers" comes from the following voting question: given several different binary classifiers for the same problem, what is the best arithmetical voting method for ensembling their results together to obtain highest possible accuracy. This will be the topic of a separate research paper.

2 Convexity of the ROC Curves

Let us fix some notations:

Ω denotes the space of all inputs, which comes together with some natural probability measure P.

A given binary decision problem on Ω is denoted by a function

$$Y : \Omega \to \{0, 1\},$$

that we want to replicate by a "machine"

$$M : \Omega \to [0, 1]$$

which takes values in the interval $[0, 1]$: We can fix a threshold $\sigma \in [0, 1]$ and declare that $Y_{predict}(x) = 1$ if $M(x) \geq \sigma$ and $Y_{predict}(x) = 0$ otherwise, and hope that $Y_{predict}$ is a good approximation of Y.

The prediction function M is constructed as the composition of two maps,

$$M = \Sigma \circ \phi,$$

where

$$\phi : \Omega \to \Phi$$

is a projection from Ω to a certain "features space" Φ (which can also be called the "information space"—it contains all the information that we can use for predicting Y), and

$$\Sigma : \Phi \to [0, 1]$$

is a "sigmoid function": for each input $x \in \Omega$, we first extract its features, $\varphi = \phi(x)$, and then calculate its sigmoid value, $\sigma = M(x) = \Sigma(\varphi)$. The probability measure on Φ is the push-forward of the probability measure from Ω.

The reason why Σ is called a "sigmoid function" is that sigmoid functions (such as the standard function $\sigma(z) = \exp(z)/(\exp(z) + \exp(-z))$ and similarly shaped functions that map $\mathbb{R}$ to $]0, 1[$) are often used as the last layer in a neural network for binary classification problems. Here we use the words "sigmoid function" in an abstract way, without referring to any specific function, and Φ can be multi-dimensional, doesn't have to be $\mathbb{R}$.

Remark that, in the AI literature, the value σ is very often called the probability, but it is not true in general. In fact, we can define the probability function $p : [0, 1] \to [0, 1]$,

$$p(\sigma) := P\left(Y(x) = 1 \mid \Sigma(\phi(x)) = \sigma\right),$$

which is *not* the identity function in general.

For each threshold value $\sigma \in [0, 1]$, one defines the corresponding *sensitivity* (= true positive) rate $TP(\sigma)$ and *specificity* (= true negative = 1 minus false positive) rate $TN(\sigma)$ by the following formulas:

$$TP(\sigma) = \frac{P(M(x) \geq \sigma | Y(x) = 1)}{P(Y(x) = 1)},$$

$$TN(\sigma) = 1 - FP(\sigma) = \frac{P(M(x) < \sigma | Y(x) = 0)}{P(Y(x) = 0)}.$$

The curve $ROC : [0, 1] \to [0, 1] \times [0, 1]$ given by the formula

$$ROC(\sigma) = (FP(\sigma), TP(\sigma))$$

is called the ROC curve in the literature, and is very widely used in many fields, see, e.g., [1–6] and references therein. See Fig. 1 for some examples of ROC curves (a test of AI for the detection of different classes of skin lesions, courtesy of Torus Actions SAS).

The ROC curves goes "backward" from the point $ROC(0) = (1, 1)$ to the point $ROC(1) = (0, 0)$ in the unit square, and the higher the curve the more accurate the machine. The so called AUC (area under the curve) is the area of the region under the ROC curve in the unit square, and is a popular measure for the accuracy of the machine.

Remark that, if we change the sigmoid function Σ by composing it with another function, $\Sigma' = \theta \circ \Sigma$, where $\theta : [0, 1] \to [0, 1]$ is a strictly increasing bijective function, then Σ and Σ' give the same ROC curve up to a reparametrization by θ. In other words, we can change a σ value to any other value by composing it with

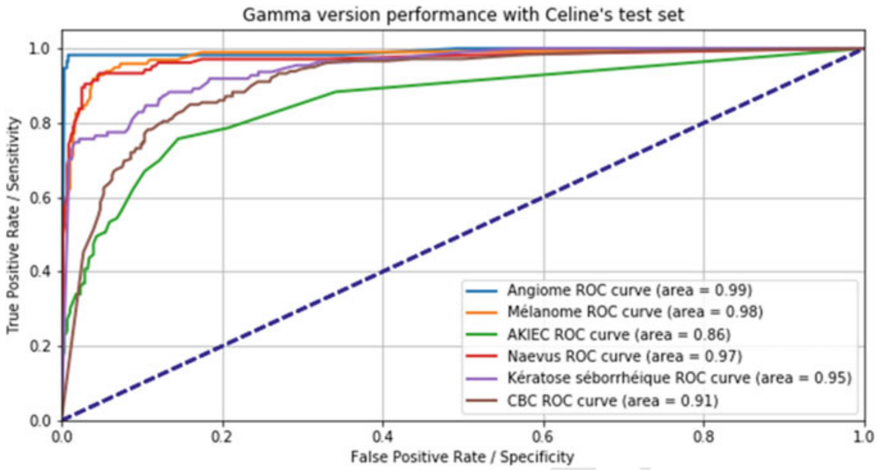

Fig. 1 Examples of ROC curves from a test on skin cancer images

a function, without changing the accuracy of the system, and this is one more reason why the σ values should not be called "probabilities" in general.

All the ROC curves in Fig. 1 are convex or almost convex (i.e. the regions under them almost coincide with their convex hulls). As a matter of fact, it has been observed since long time ago that "reasonable" ROC curves are nearly convex, and moreover, non-convexity is a kind of defect that could be corrected by randominizing certain things, see, e.g., [2]. Philosophically, machine learning aims at creating rational predictors (that are better than just random functions), and that's why ROC curves are mostly convex, but in practice our predictors are not completely rational, and so there are ROC curves that are not completely convex (another reason for non-convexity is lack of data).

Definition 1 We will say that machine $M = \Sigma \circ \phi$ (or its corresponding sigmoid function Σ), for a given binary classifier $Y : \Omega \to \{0, 1\}$ and features map ϕ, is *optimal* or *most rational*, if its ROC curves lies above the ROC curves of all the other sigma functions for the same problem (ϕ is fixed).

Proposition 1 *Among all the (measurable) sigmoid functions* $\Sigma : \Phi \to [0, 1]$ *(for a given binary classifier* $Y : \Omega \to [0, 1]$ *and a given features projection map* $\phi : \Omega \to \Phi$*) there is an optimal sigmoid function (called the "probability sigmoid function"), and its ROC curve is convex.*

As a consequence, if a machine learning algorithm is efficient enough so that it creates a "nearly optimal" machine, then the ROC curve of that machine is convex or nearly convex.

In order to prove the above proposition, we will construct the optimal machine "explicitly". For each $\varphi \in \Phi$, denote by $p(\phi)$ the conditional probability of $Y = 1$ given that the features value is φ:

$$p(\varphi) = P(Y(x) = 1 \mid \phi(x) = \varphi).$$

Then this probability function $\varphi \mapsto p(\varphi)$ is actually the best sigmoid function, i.e., if we put $\Sigma(\varphi) := p(\varphi)$, then this sigmoid function has the best ROC curve.

Indeed, denote by $\Sigma' : \Phi \to [0, 1]$ any other sigmoid function. Fix an arbitrary false positive level $\alpha \in]0, 1[$. Denote by $\sigma \in]0, 1[$ (resp., $\sigma' \in]0, 1[$) the threshold value for the probability sigmoid function $p(\varphi)$ (resp., the function $\Sigma'(\varphi)$ to get this level of false positive. Then we have the following formula,

$$\alpha = \frac{\int_{\{\varphi \in \Phi \mid p(\varphi) \geq \sigma\}} (1 - p(\varphi)) d\varphi}{\int_{\Phi} (1 - p(\varphi)) d\varphi} = \frac{\int_{\{\varphi \in \Phi \mid \Sigma'(\varphi) \geq \sigma'\}} (1 - p(\varphi)) d\varphi}{\int_{\Phi} (1 - p(\varphi)) d\varphi},$$

(where $\int_{\Phi} (1 - p(\varphi)) d\varphi = P(Y = 0)$ is the probability measure of the negative set $\{x \in \Omega \mid Y(x) = 0\}$), which implies that

$$\int_{\{\varphi \in \Phi \mid p(\varphi) \geq \sigma\}} (1 - p(\varphi)) d\varphi = \int_{\{\varphi \in \Phi \mid \Sigma'(\varphi) \geq \sigma'\}} (1 - p(\varphi)) d\varphi.$$

To simplify the notations, put $A = \{\varphi \in \Phi \mid p(\varphi) \geq \sigma\}$ and $B = \{\varphi \in \Phi \mid \Sigma'(\varphi) \geq \sigma'\}$. Then we have $\displaystyle\int_A (1 - p(\varphi)) d\varphi = \int_B (1 - p(\varphi)) d\varphi$, which implies that

$$\int_{A \setminus B} (1 - p(\varphi)) d\varphi = \int_{B \setminus A} (1 - p(\varphi)) d\varphi.$$

Since $(1 - p(\varphi)) \leq 1 - \sigma$ on $A \setminus B$ while $(1 - p(\varphi)) > 1 - \sigma$ on $B \setminus A$, we must have that $P(A \setminus B) \geq P(B \setminus A)$, which implies that

$$\int_{A \setminus B} p(\varphi) d\varphi \geq \sigma P(A \setminus B) \geq \sigma P(B \setminus A) \geq \int_{B \setminus A} p(\varphi) d\varphi,$$

which implies that $\int_A p(\varphi) d\varphi \geq \int_B p(\varphi) d\varphi$. This last inequality means exactly that the true positive level of p at the false positive level α is greater or equal to the true positive level of Σ' at the same false positive level. In other words, the ROC curve of the probability sigmoid function p lies above the ROC curve of Σ' everywhere, i.e., p is the optimal sigmoid function. The two ROC curves coincide if and only if, in the above formulas, B coincides with A (up to a set of measure zero) for every false positive level α, and it basically means that Σ' is obtained from p by composing it with a monotonous function. In other words, up to a reparametrization of the sigmoid values, the probability sigmoid function is the only optimal sigmoid function.

The convexity of the ROC curve of the optimal sigmoid function p follows directly from its construction, which assures that the conditional event probability is nondecreasing (the higher the sigmoid value σ, the higher the conditional conditional probability value, which is obvious because this value is equal to σ in our construction).

See Theorem 3 of [2]. Indeed, denote by

$$\alpha(\sigma) = \frac{\int_{\{\varphi\in\Phi\,|\,p(\varphi)\geq\sigma\}}(1-p(\varphi))d\varphi}{\int_{\Phi}(1-p(\varphi))d\varphi} \quad \text{and} \quad \beta(\sigma) = \frac{\int_{\{\varphi\in\Phi\,|\,p(\varphi)\geq\sigma\}}p(\varphi)d\varphi}{\int_{\Phi}p(\varphi)d\varphi}$$

the false negative and false positive levels at threshold σ for the probability sigmoid function p, then we have

$$\frac{d\beta}{d\alpha} = \frac{\int_{\Phi}(1-p(\varphi))d\varphi}{\int_{\Phi}p(\varphi)d\varphi}\cdot\frac{\sigma}{1-\sigma},$$

which is an increasing function in σ, but a decreasing function in α, because α itself is a decreasing function in σ. Hence β is a concave function in α, which means that the ROC curve is convex.

Remark 1 (*Under the above notations*) A machine may be called *consistent* if it satisfies the following natural monotonicity condition: the higher the sigmoid value, the higher the probability of $Y = 1$. In other words, a machine is consistent if $p(\sigma)$ is a strictly monotonous increasing function in σ, with $p(0) = 0$ and $p(1) = 1$. We have shown that optimal machines are automatically consistent (and hence their ROC curves are automatically convex).

Remark 2 Given an original probability distribution P on the data space Ω, which is imbalanced in the sense that $P(Y = 0) \neq P(Y = 1)$ (imbalance between negative and positive cases), we may change it to a new, balanced, probability distribution $\hat{P}$, defined by the following formula:

$$\hat{P}(A) = \frac{1}{2}\left[\frac{P(A\cap\{Y=0\})}{P(Y=0)} + \frac{P(A\cap\{Y=1\})}{P(Y=1)}\right].$$

It is easy to verify that the ROC curve (for a given machine $M = \Sigma\circ\phi : \Omega \to [0, 1]$) with respect to $\hat{P}$ coincides exactly with the ROC curve with respect to P. Indeed, for any $\sigma \in [0, 1]$, the true positive level at σ is $TP(\sigma) = \dfrac{P(\Sigma > \sigma, Y = 1)}{P(Y = 1)} = 2\hat{P}(\Sigma > \sigma, Y = 1) = \dfrac{\hat{P}(\Sigma > \sigma, Y = 1)}{\hat{P}(Y = 1)} = \widehat{TP}(\sigma)$, and similarly for $TN(\sigma)$. Thus in the study of accuracy, without loss of generality, one may suppose that the probability distribution is balanced with respect to Y in the sense that $P(Y = 0) = P(Y = 1) = 0.5$.

3 Inequalities Relating Different Measures of Accuracy

Besides AUC, another popular measure of accuracy is the so called *balanced accuracy* defined as follows (for a given machine, i.e., a given sigmoid function):

$$BA(\sigma) = \frac{TP(\sigma) + TN(\sigma)}{2}$$

and (the maximal possible BA when one varies the threshold)

$$MBA = \max_{\sigma \in [0,1]} BA(\sigma)$$

Remark that $BA(0) = BA(1) = 0.5$, so we always have $MBA \geq 0.5$. If $MBA = 0.5$ then the sigmoid function is "useless" (worse than random). Both AUC and MBA are of course smaller or equal to 1, and they are equal to 1 if the machine is "perfect". The following proposition gives relations between these two measures of accuracy, which implies that if one of them tends to 1 then the other also tends to 1.

Proposition 2 *For any given sigmoid function we have*

$$1 - 2(1 - MBA)^2 \geq AUC \geq 2MBA - 1$$

If, moreover, the sigmoid function is rational in the sense that its ROC curve is convex, then we have

$$AUC \geq MBA$$

Instead of the balanced accuracy, one can also use *cost-based weighted accuracy* (see, e.g., [3]): given weight $w \in]0, 1[$ (determined by a cost function), we put

$$WA(\sigma) = w.TP(\sigma) + (1 - w).TN(\sigma) = w.TP(\sigma) - (1 - w).FP(\sigma) + (1 - w),$$
$$MWA = \max_{\sigma \in [0,1]} WA(\sigma).$$

and we have the following similar inequalities relating MWA to AUC:

Proposition 3 *For any given weight $w \in]0, 1[$ and any given sigmoid function, we have*

$$1 - \frac{(1 - MWA)^2}{2w(1 - w)} \geq AUC$$

If, moreover, the sigmoid function is rational in the sense that its ROC curve is convex, then we have

$$AUC \geq 1 - \frac{(1 - MWA)}{2\min(w, 1 - w)}$$

When $w = 0.5$ then MWA coincides with MBA and Proposition 3 basically becomes Proposition 2. Remark also that both AUC and MBA are *balanced* measures of accuracy, in the sense that true positive and true negative ratios are equally important in their formulas, while MWA ($w \neq 0.5$) is not balanced. This fact is important for optimal arithmetical voting methods to be discussed in the next section.

All of the above inequalities are direct consequences of the following elementary proposition:

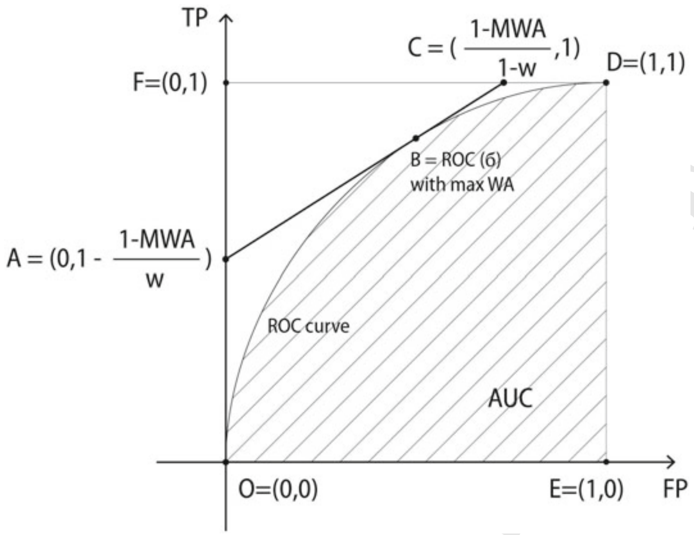

Fig. 2 The ROC curve and the tangent line at a maximal weighted average point

Proposition 4 *Give a sigmoid function* $\Sigma : \Omega \to [0, 1]$ *and a weight* $w \in]0, 1[$*, a number* $\sigma \in [0, 1]$ *is a threshold where* Σ *attains highest weighted accuracy if and only if the straight line through the point* $ROC(\sigma)$*, consisting of the points* $(FP(\sigma) + wt, TP(\sigma) + (1 - w)t)$*,* $t \in \mathbb{R}$*, lies above the ROC curve.*

The above simple propositions show that, the three different measures of accuracy AUC, MBA and MWA are correlate well with each other, and if one of them tends to 1 (maximal accuracy) then the other two also tend to 1.

Let us indicate here the proof of the above propositions. See Fig. 2 for the illustration.

The lines $\ell = \{(FP(\sigma) + wt, TP(\sigma) + (1 - w)t), t \in \mathbb{R}\}$ of slope $(w, 1 - w)$ in Proposition 4 are simply "lines of constant weighted accuracy". If the point $(FP(\sigma), TP(\sigma))$ of the ROC curve give maximal weighted accuracy, then no point of the ROC curve can lie above its corresponding line of slope $(w, 1 - w)$, because lying above means higher weighted accuracy.

The line $\ell = \{(FP(\sigma) + wt, TP(\sigma) + (1 - w)t), t \in \mathbb{R}\}$, where σ gives the maximal weighted accuracy for the sigmoid function Σ, cuts the boundary of the unit square at two points $A = (0, 1 - \frac{1 - MWA}{w})$ and $C = (\frac{1 - MWA}{1 - w}, 1)$. The triangle $\triangle ACF$, where $F = (0, 1)$, is disjoint from the region under the ROC curve, which implies that $AUC + \text{area}(\triangle ACF) \leq 1$. Since $\text{area}(\triangle ACF) = \frac{FA \cdot FC}{2} = \frac{(1 - MWA)^2}{2w(1 - w)}$, we get the inequality

$$AUC \leq 1 - \frac{(1-MWA)^2}{2w(1-w)}.$$

On the other hand, the region under the ROC curve contains the rectangle whose vertices are $(FP(\sigma), 0)$, $(FP(\sigma), TP(\sigma)$, $(1, TP(\sigma))$, $(1, 0)$. The surface area of this rectangle is $TP(\sigma) \cdot (1 - FP(\sigma)) = TP(\sigma) \cdot TN(\sigma) = TP(\sigma) + TN(\sigma) - 1 + (1 - TP(\sigma))(1 - TP(\sigma)) \geq TP(\sigma) + TN(\sigma) - 1 = 2BA(\sigma) - 1$ (for every σ). Hence we get the inequality

$$AUC \geq 2MBA - 1.$$

If the ROC curve is convex, then the region below it contains the quadrilateral $OBDE$, where $O = (0, 0)$, $B = (FP(\sigma), TP(\sigma))$, $D = (1, 1)$, $E = (1, 0)$ (for any σ). The surface area of this quadrilateral is exactly equal to $BA(\sigma)$, i.e., to MBA, hence we get the inequality $AUC \geq BA(\sigma)$ for any σ, i.e., we have

$$AUC \geq BA(\sigma).$$

Finally, the inequality $AUC \geq 1 - \dfrac{(1-MWA)}{2\min(w, 1-w)}$ in the case when the ROC curve is convex and the weight w is arbitrary is a direct consequence of the inequalities

$$AUC \geq area(OBDE) \geq \min(area(OADE), area(OCDE))$$

and the equalities

$$area(OADE) = 1 - \frac{(1-MWA)}{2w}, \quad area(OCDE) = 1 - \frac{(1-MWA)}{2(1-w)}.$$

(See Figure 2).

References

1. Tom Fawcett, An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
2. T. Gneiting, Peter Vogel Receiver Operating Characteristic (ROC) Curves (2018). https://arxiv.org/abs/1809.04808
3. J. Hernandez-Orallo, P.A. Flach, C. Ferri, A unified view of performance metrics: translating threshold choice into expected classification loss. J. Mach. Learn. Res. **13**, 2813–2869 (2012)
4. L.L. Pesce, Ph.D. 1, C.E. Metz, Ph.D. 1, K.S. Berbaum, Ph.D. 2 On the convexity of ROC curves estimated from radiological test results, Acad Radiol. **17**(8), 960–968.e4 (2010). https://doi.org/10.1016/j.acra.2010.04.001
5. J.A. Swets, R.M. Dawes, J. Monahan, Psychological science can improve diagnostic decisions, Psychological Science in the Public Interest, VOL. 1, NO. 1 (2000)
6. Wikipedia page on ROC: https://en.wikipedia.org/wiki/Receiver_operating_characteristic
7. Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms* 1st edn. (Chapman Hall/CRC, 2012)