The Property of ROC Curves

© Kai Feng, Han Hong, Ke Tang and Jingyuan Wang

Beihang University
Stanford University
Tsinghua University
Beihang University

January 3, 2020

- Much attention has been drawn to the potential of machine learning (ML) in assisting human decision making.
- Binary classification decision making is a foundational building block of related works.
- Accuracy is insufficient to evaluate the quality of binary classifiers.

Example : **Prostate Cancer**

In the U.S., about 1.4 percent of men aged 44 to 64 have prostate cancer. A simple prediction of all patients as low risk would result in an accuracy higher than 95%.
This high accuracy diagnosis strategy is not intended because it does not distinguish between high risk and low risk groups.

- The ROC (Receiver operating characteristics) curve is an alternative to accuracy and plays a key role in the binary classification problem.

		True Condition	
		Condition Positive	Condition Negative
Predictied Condition	Predictied Positive	True Postive	False Positive Type I Error
	Predictied Negative	False Negative Type II Error	True Negative

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Suppose there are 100 diseased and 100 healthy people, a doctor diagnoses 20 of the healthy as diseased, and 10 diseased are missed. The TPR/FPR of the doctor is $(0.9, 0.2)$.

- Let X_i be a set of features.
  Let $Y_i \in \{0,1\}$ be the outcome (label), $\hat{Y}_i \in \{0,1\}$ the prediction.
  Let $\hat{p}(X_i) \in [0,1]$ a sample estimate of the probability of the label taking 1 conditional on the features.
- Recall the definitions (in sample):

$$\text{TPR} = \frac{\text{Outcome True, Predicted Positive}}{\text{Outcome True}} = \frac{\sum_{i=1}^{n} Y_i \hat{Y}_i}{\sum_{i=1}^{n} Y_i},$$

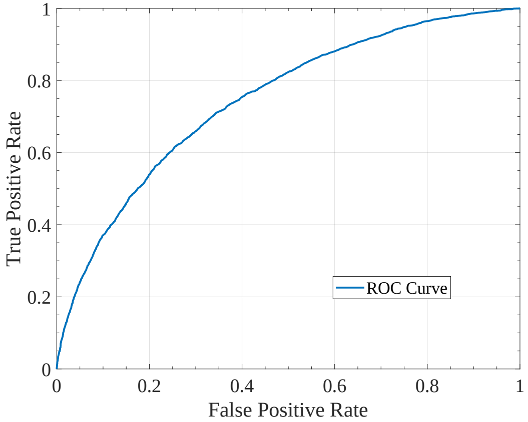
$$\text{FPR} = \frac{\text{Outcome False, Predicted Positive}}{\text{Outcome False}} = \frac{\sum_{i=1}^{n} (1 - Y_i)\hat{Y}_i}{\sum_{i=1}^{n} (1 - Y_i)}.$$

- A ROC curve is the collection of the set of all TPR/FPR pairs corresponding to decision rules $\hat{Y}_i = \mathbb{1}(p(X_i) > c), c \in [0,1]$.
  Let $\hat{\alpha}(c) = \text{FPR}(c)$, $\hat{\beta}(c) = \text{TPR}(c)$.

$$\text{ROC} := \hat{\alpha}(c) \mapsto \hat{\beta}\left(\hat{\alpha}^{-1}(\alpha)\right).$$

  It present the tradeoff between TPR and FPR.

Sample of ROC curve.



We provide a statistical formulation of the ROC curve, we demonstrate:

The relation between ROC curve with *loss (utility) function* and *decision rule*.

Confidence level for an estimated ROC to account for its sampling uncertainty.

The influence of AUC (area under curve) and its implication for *model selection*.

Neyman Pearson Lemma and Decision Rules

- Binary decision making is inherently related to hypothesis testing. For a general classification rule $\hat{Y}_i = \mathbb{1}(X_i \in R)$, denote the population analogs of TPR/FPR as PTPR and PFPR

$$\text{TPR} \xrightarrow{\mathbb{P}} \text{PTPR} \equiv \frac{\mathbb{E}[Y_i \mathbb{1}(X_i \in R)]}{p},$$
$$\text{FPR} \xrightarrow{\mathbb{P}} \text{PFPR} \equiv \frac{\mathbb{E}[(1 - Y_i)\mathbb{1}(X_i \in R)]}{1 - p}.$$

  where $p = \mathbb{E}[Y_i]$ is the overall population portion of positive labels.
- Then by Bayes law

$$\text{PTPR} = \int \mathbb{1}(X \in R) f(X|Y = 1)\, \mathrm{d}X, \quad \text{PFPR} = \int \mathbb{1}(X \in R) f(X|Y = 0)\, \mathrm{d}X.$$

- PTPR is the power of the test; PFPR is the size of the test.

- The classical Neyman Pearson Lemma states that the collection of likelihood ratio tests

$$R_{NP}(d) = \left\{ x : \frac{f(X|Y=1)}{f(X|Y=0)} > d \right\},$$

where $d \in (0, \infty)$ varies, are *most powerful tests* that maximize power for whatever size it achieves.

- By the Bayes law, write

$$R_{NP}(d) = \left\{ x : \frac{p(x)}{1-p(x)} > d\frac{p}{1-p} \right\} = \left\{ x : p(x) > c = \frac{dp}{1-p+dp} \right\},$$

where $p(X_i) = \mathbb{P}(Y_i = 1|X_i)$ is the true probability function.

- Consequently, the ROC corresponding to the decision rules

$$\hat{y} = \mathbb{1}(p(x) > c) \quad c \in [0, 1]$$

has the *Neyman-Pearson optimality* that it lies weakly above the ROC of any alternative collection of decision rules.

- With arbitrary cost functions, Bayesian optimal PTPR/PFPR pair can lie below the optimal ROC curve or even below the 45 degree line.
- Consider the Loss (function) "matrix"

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	0	$C_{0R}(x)$
$Y = 1$	$C_{1A}(x)$	0

- The minimizing rejection region R is then

$$\bar{R} = \left\{ x : p(x) > c(x) = \frac{c_{0R}(x)}{c_{0R}(x) + c_{1A}(x)} \right\}.$$

Statistical Inference of ROC Curves

- We derived asymptotic pointwise confidence bands for an estimated ROC to account for its sampling uncertainty.
- Consider parametric models of $p\left(X_i, \theta\right)$ under i.i.d sampling assumptions, write TPR/FPR as

$$\hat{\beta}\left(c\right) = \frac{1/n \sum_{i=1}^{n} y_i \mathbb{1}\left(p\left(x_i, \hat{\theta}\right) > c\right)}{\hat{p}},$$

$$\hat{\alpha}\left(c\right) = \frac{1/n \sum_{i=1}^{n} \left(1 - y_i\right) \mathbb{1}\left(p\left(x_i, \hat{\theta}\right) > c\right)}{1 - \hat{p}},$$

where $\hat{p} = 1/n \sum_{i=1}^{n} y_i$.

- The PTPR and PFPR are written as

$$\beta\left(c\right) = \frac{1}{p} \mathbb{E}\left[p\left(X\right) \mathbb{1}\left(p\left(X, \theta_0\right) > c\right)\right],$$

$$\alpha\left(c\right) = \frac{1}{1-p} \mathbb{E}\left[\left(1 - p\left(X\right)\right) \mathbb{1}\left(p\left(X, \theta_0\right) > c\right)\right].$$

- Let $\hat{\beta}_\alpha = \hat{\beta}\left(\hat{\alpha}^{-1}\left(\alpha\right)\right)$ and $\beta_\alpha = \beta\left(\alpha^{-1}\left(\alpha\right)\right)$.

- To construct an asymptotic confidence inteval for β_α,

$$\lim_{n\to\infty}\inf \mathbb{P}\left(\hat{\beta}_\alpha - \hat{d} \leq \beta_\alpha \leq \hat{\beta}_\alpha + \hat{d}\right) \geq 1-\eta, \quad (1)$$

we derive the asymptotic distribution of $\hat{\beta}_\alpha - \beta_\alpha$:

Theorem

Assuming $p(x,\theta)$ satisfies a typical stochastic equicontinuity condition and there is a consistent estimate of $\hat{\theta}$ with an asymptotic linear influence function representation

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\kappa_i + o_{\mathbb{P}}(1), \quad \text{where} \quad \kappa_i = \kappa(y_i, x_i) \quad (2)$$

Then, the asymptotic distribution of $\hat{\beta}_\alpha - \beta_\alpha$ is of the form:

$$\sqrt{n}\left(\hat{\beta}_\alpha-\beta_\alpha\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_{\mathbb{P}}(1), \quad \text{where} \quad \psi_i = \psi(y_i, x_i, \alpha). \quad (3)$$

- It follows that

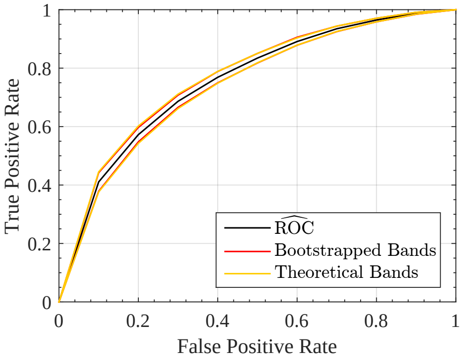
$$\sqrt{n}\left(\hat{\beta}_\alpha-\beta_\alpha\right) \xrightarrow{d} N\left(0,\sigma^2\right), \quad \text{where} \quad \sigma^2 = Var(\psi_i).$$

We estimated the asymptotic distribution by sample analogs and by bootstrapping.

The data generating process is specified to be a logit model,

$$p(X) = \exp(X'\beta) / (1 + \exp(X'\beta))$$

where $X = (X_1, X_2)$, $\beta = (1, -0.5)$, $X_1 \sim N(2,1)$, $X_2 \sim N(0,1)$, $B \sim$ Uniform$(0,1)$ and $Y = \mathbb{1}(p(X_1, X_2) > B)$, X_1 and X_2 are independent.



AUC and Model Comparison and Selection

- The sample AUC corresponding to θ is given by

$$\mathsf{SAUC}(\theta) = \frac{1}{n^2 \hat{p}(1-\hat{p})} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\left(p\left(x_i, \hat{\theta}\right) > p\left(x_j, \hat{\theta}\right)\right) y_i (1 - y_j).$$

- This takes the form of a U-process and converges to a population AUC, defined as

$$\begin{aligned} &\mathrm{PAUC}(\theta) \\ &= \frac{1}{p(1-p)} \iint \mathbb{1}(p(x,\theta) > p(w,\theta))\, p(x)(1-p(w))\, f(x) f(w)\, \mathrm{d}x \mathrm{d}w. \end{aligned}$$

- This integral would be maximized if the indicator is turned on whenever $p(x) > p(w)$.
- Under correct specification, this can obviously be achieved when $\theta = \theta_0$, where $p(x, \theta_0) = p(x) > p(w) = p(w, \theta_0)$.
  Therefore, by standard M-estimator arguments (Newey and McFadden, 1994) the maximum AUC estimator is consistent under correct specification and suitable sample regularity conditions.

- We prove by further use the the U-process stochastic equicontinuity results in (Sherman, 1993).

Theorem

Let

$$\eta\left(z_i, z_j, \theta\right) = \left(\mathbb{1}\left(p(x_i, \theta) > p\left(x_j, \theta\right)\right) - A\right) y_i \left(1 - y_j\right),$$

and $Q\left(\theta\right) = \mathbb{E}\left[\eta\left(z_i, z_j, \theta\right)\right]$, then

$$\sqrt{n}\left(\hat{A} - A\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i + o_{\mathbb{P}}\left(1\right), \quad \sqrt{n}\left(\hat{A} - A\right) \xrightarrow{d} N\left(0, Var\left(\xi_i\right)\right), \tag{4}$$

the asymptotic covariance can be calculate as

$$\xi_i = \frac{1}{p\left(1 - p\right)} \left[\eta_1\left(z_i, \theta^*\right) + \eta_2\left(z_i, \theta^*\right) + \frac{\partial}{\partial \theta} Q\left(\theta^*\right) \kappa_i\right],$$

in which

$$\eta_1\left(z_i, \theta\right) = \mathbb{E}_{z_j}\left[\eta\left(z_i, z_j, \theta\right)\right], \quad \eta_2\left(z_j, \theta\right) = \mathbb{E}_{z_i}\left[\eta\left(z_i, z_j, \theta\right)\right].$$

- The results derived above provide the basis for constructing model tests.
- It is possible that a different criterion function, such as cross entropy, is used to estimate parameters before the use of the AUC criterion for model selection.
- Consider two competing models with parameters θ and ϑ, and corresponding sample AUCs $\hat{A}_1\left(\hat{\theta}\right)$ and $\hat{A}_2\left(\hat{\vartheta}\right)$, then it follows from (4) that

$$\hat{A}_1\left(\hat{\theta}\right) - \hat{A}_2\left(\hat{\vartheta}\right) = \left(A_1\left(\theta^*\right) - A_2\left(\vartheta^*\right)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i^1 - \xi_i^2\right) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

- A test of the null hypothesis of $A_1\left(\theta^*\right) = A_2\left(\vartheta^*\right)$ between two models relies on asymptotic distribution of $\xi_i^1 - \xi_i^2$.

- Next table reports an AUC-based model selection exercise between two misspecfied models.
- The model (M1) is a logit model with $p(X_1) = \frac{\exp(\theta_1 X_1)}{1+\exp(\theta_1 X_1)}$; the model (M2) is a logit model with $p(X_2) = \frac{\exp(\theta_2 X_2)}{1+\exp(\theta_2 X_2)}$.

Table: Model Selection

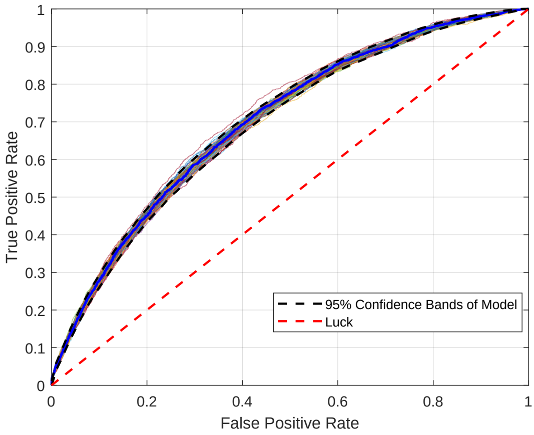
	Bootstrap	Theoretical
A1 (mean)	0.7341	0.7314
A2 (mean)	0.6214	0.6191
A1-A2 (mean)	0.1127	0.1124
A1-A2 (std)	0.0102	0.0103

- We obtain a significant z score: $z = \frac{\hat{A}_1(\hat{\theta})-\hat{A}_2(\hat{\vartheta})}{std(\hat{A}_1(\hat{\theta})-\hat{A}_2(\hat{\vartheta}))}$, which rejects the null hypothesis that M1 is equivalent to M2.

Application

- A data set derived from Haidian District Maternal and Child Health Hospital in Beijing, comprehensively records birth process in the hospital from 2001 to 2010.
- Altogether 545 features are available for each observation, including blood test, urine test and pregnogram examination results.
- The data used in the current analysis includes 108911 records, a total of 15.5% of our sample had hyperglycemia in pregnancy.
- We used a logistic regression with L_1 regularization for prediction and used an 8:2 training and test partition.
- Only data collected up to the 20th week are used for prediction.

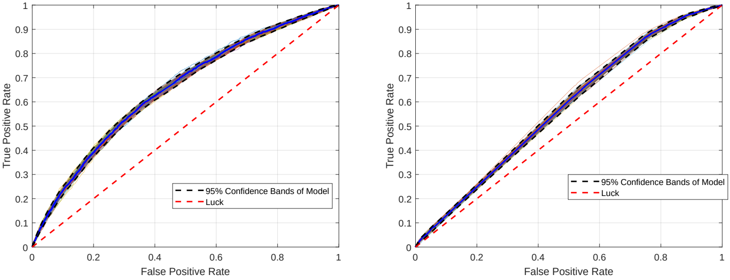
If we use all the features, the AUC of the model is 0.6988 ± 0.0092.



One may interested in whether certain types of checks are better for prediction.

The AUC of pregnogram examination features is 0.6506 ± 0.0098.
The AUC of blood test features is 0.5738 ± 0.0107.
We can further get $std(\text{AUC}_P - \text{AUC}_B) = 0.0080$,
the z score: $z = 9.60$, which implies that the pregnogram model is better.

Figure: Capabilities of pregnogram and blood test features to predict hyperglycemia in pregnancy



"Decision Making with Machine Learning and ROC Curves"
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3382962

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Sherman, R. P. (1993): "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61(1), 123–137.