Equivalência entre a Área sob a Curva Kolmogorov-Smirnov e o Índice de Gini na Avaliação de Desempenho de Decisões Binárias

Paulo J. L Adeodato, Sílvio B. Melo

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 15.064 – 91.501-970 – Recife – PE – Brazil

{pjla,sbm}@cin.ufpe.br

Abstract. *This paper proposes and proves the important equivalence between the Gini index and the area under the Kolmogorov-Smirnov (KS) distribution curve. The proof's rationale is similar to that used in the proof of equivalence between AUC_ROC and AUC_KS. But different from that, this one uses a transformation that preserves the 1-to-1 correspondence between the ideal classifier on the KS and Lorenz curve domains. As metrics, this paper proves that the Gini index ratio to that of the ideal classifier is equivalent to the area under the KS curve ratio to that of its ideal classifier. That is Gini_Index_Ratio = AUC_KS_Ratio. This complements the proven equivalence between KS and ROC area metrics extending it to the Gini index.*

Resumo. *Este artigo propõe e prova a importante equivalência entre o índice de Gini e a área sob a curva da distribuição Kolmogorov-Smirnov (KS). A lógica da prova é semelhante à utilizada na prova de equivalência entre a AUC_ROC e a AUC_KS. Mas, diferente daquela, esta usa uma transformação que preserva a relação 1-para-1 entre o classificador ideal nos domínios das curvas KS e de Lorenz. Como métricas, este artigo prova que a razão do índice de Gini pelo do classificador ideal é equivalente à razão da área sob a KS pela área do classificador ideal. Isso é Gini_Index_Ratio = AUC_KS_Ratio. Isso complementa a equivalência entre as métricas de área KS e ROC, estendendo-a para o índice de Gini.*

1. Introdução

Em negócios, uma das estratégias de decisão mais comuns e relevantes para a seleção dos indivíduos ou objetos elegíveis para uma ação é classificá-los de acordo com uma pontuação de classificação (escore), escolhendo aqueles acima de um limite pré-definido [Provost e Fawcett, 2013]. Essa abordagem é usada em aplicações como seleção de pessoal, detecção de fraude e alocação de recursos em políticas públicas. Esta pontuação é calculada ou por uma ponderação de um conjunto de variáveis com base em parâmetros definidos por humanos ou pela aplicação de uma função aprendida por um algoritmo de classificação a partir de dados com rótulos binários como respostas desejadas, de acordo com critérios específicos de otimização. Esse mapeamento em um escalar (escore) é fundamental para dar ao decisor humano o poder de controlar a decisão pela simples definição de um limiar sobre o escalar. Isso se aplica a respostas, tanto por ponderação humana quanto aprendidas, otimizando o limiar não só para o desempenho técnico do classificador, como também para os indicadores de desempenho do negócio (*Key Performance Indicators* - KPIs) [Provost e Fawcett, 2013].

Como em aplicações em geral, ainda sem uso específico definido para o classificador, o seu ponto de operação (limiar de decisão) não pode ser pré-especificado. Em tal cenário, as métricas de avaliação de desempenho devem medir características gerais do classificador, sem qualquer suposição sobre esse ponto, considerando todo o intervalo de pontuação. Isso torna a matriz de confusão, as taxas de erro, a máxima diferença vertical de Kolmogorov-Smirnov e outras métricas de pontos específicos de operação inadequadas para avaliação de classificadores que dão controle ao decisor.

Essa é uma provável razão por que a área sob a curva ROC (AUC_ROC) [Provost e Fawcett, 2001] tornou-se tão popular entre os cientistas na avaliação de qualidade em classificação binária. A curva de Lorenz e o índice de Gini [Bellù e Liberatti, 2006] constituem uma ferramenta estatística consolidada que tem sido utilizada em economia para a avaliação de desigualdades entre as populações humanas, focando principalmente na distribuição de renda ou riqueza.

Recentemente, a área sob a curva de Kolmogorov-Smirnov (AUC_KS) foi provada equivalente à da curva ROC (AUC_ROC) [Adeodato e Melo, 2016], a menos de um termo subtrativo. Mais especificamente, AUC_KS = AUC_ROC - 0,5. Esse é um resultado importante, não só por trazer o conhecimento estatístico de uma distribuição consolidada no teste de hipóteses [Conover, 1999], mas também por tornar os cálculos mais simples para a avaliação de desempenho dos classificadores. Este artigo prova que a área normalizada sob a curva da distribuição Kolmogorov-Smirnov (AUC_KS_Ratio) e o índice de Gini normalizado (Gini_Index_Ratio) são equivalentes.

Este artigo está organizado em mais três seções. A Seção 2 apresenta o KS e os conceitos Lorenz/Gini e curvas para a avaliação de desempenho em problemas binários. A Seção 3 apresenta a prova formal da equivalência das áreas. A Seção 4 conclui o artigo discutindo alguns impactos que tais fundamentos estatísticos podem trazer à interpretação da classificação binária nas suas aplicações e futuros avanços teóricos.

2. Métricas de Desempenho em Classificação Binária

Para um classificador binário de uso geral ser usado em um sistema de suporte à decisão, sua resposta deve ser contínua para dar ao decisor humano o poder de controlar os impactos da decisão. Assim, classificadores que produzem decisões "duras", apresentando a classe predita como resposta estão excluídos desta pesquisa. Os classificadores de interesse aqui mapeiam o espaço de entrada multidimensional em um escalar sobre o qual o decisor define o limiar de decisão criando as duas classes com base nos KPIs do negócio. Assim, o desempenho do classificador é avaliado pela comparação entre a classe prevista e a classe real para cada padrão da amostra de teste considerando todos os potenciais limiares de decisão.

Consequentemente, as métricas de avaliação de desempenho também devem contemplar pelo menos uma faixa deste escalar na região de interesse para a tomada de decisão [Provost e Fawcett, 2013]. Métricas baseadas em limiares específicos, como taxas de erro, não são adequados para avaliar esses classificadores flexíveis. Este artigo foca em métricas baseadas em área, tais como AUC_ROC, AUC_KS e o índice de Gini, que medem o desempenho integrando o impacto do classificador ao longo do intervalo de pontuação. Essas são todas relacionadas e este artigo está focado na equivalência entre as áreas sob a curva de distribuição estatística de Kolmogorov-Smirnov e o índice de Gini medido como o dobro da área entre a curva de Lorenz e a diagonal.

2.1. Distribuição Kolmogorov-Smirnov

A distribuição de Kolmogorov-Smirnov foi originalmente concebida como um teste de hipótese para medir a aderência de uma distribuição aos dados [Conover, 1999]. Em problemas de classificação binária, ela tem sido usada como medida de dissimilaridade para avaliar o poder discriminante do classificador medindo a distância que a sua pontuação produz entre as funções de distribuição acumulada (FDA) das duas classes de dados [Krzanowski e Hand, 2009]. A métrica comum para ambos os fins é a diferença vertical máxima entre as FDAs (Max_KS), que é invariante para faixa e escala de escore, tornando-a adequada para comparações de classificadores, no ponto de máximo.

Uma métrica baseada na área sob a curva Kolmogorov-Smirnov (AUC_KS) já havia sido usada em aplicação prática [Adeodato *et al.*, 2008] e acaba de ser formalmente proposta e provada equivalente à AUC_ROC [Adeodato e Melo, 2016]. Mais precisamente, AUC_ROC = 0,5 + AUC_KS. A AUC_KS é invariante a faixa e escala de escore, permitindo a sua aplicação na avaliação de desempenho de classificadores binários sobre toda a faixa de pontuação, independente do ponto de operação. Os exemplos são ordenados pelos seus escores e divididos em quantis, tornando assim a área sob a curva de Kolmogorov-Smirnov (AUC_KS) uma métrica robusta para a avaliação de desempenho. Além disso, a AUC_KS é não-paramétrica, tornando-a mais simples para calcular o desempenho médio em procedimentos de validação cruzada do tipo *k-fold*. A Figura 1 (esquerda) mostra a curva KS e suas principais características.

2.2. Curva de Lorenz e Índice de Gini

A curva de Lorenz [Bellù e Liberatti, 2006] é uma representação estatística da função de distribuição acumulada (FDA) dos exemplos da classe alvo (em um problema binário) exibida graficamente contra a fração de toda a amostra (todos os exemplos de todas as classes). Ela define uma métrica de área, o índice de Gini, que é duas vezes a área entre a FDA da classe alvo e a diagonal aleatória, que corresponde aos exemplos que são ordenados aleatoriamente a partir da amostra de teste. A Figura 1 (direita) mostra a curva de Lorenz com os delimitadores da área de Gini. O classificador ideal é representado pelas linhas tracejadas e define um triângulo com a diagonal cuja área é (1 - fração da classe alvo)/2, cujo índice de Gini é duas vezes esta área. Concebido por Corrado Gini em 1912 [Ceriani e Verme, 2012], o índice de Gini é uma métrica de dispersão estatística, tradicionalmente usada em ciências sociais para medir as desigualdades nos padrões de vida da população humana.

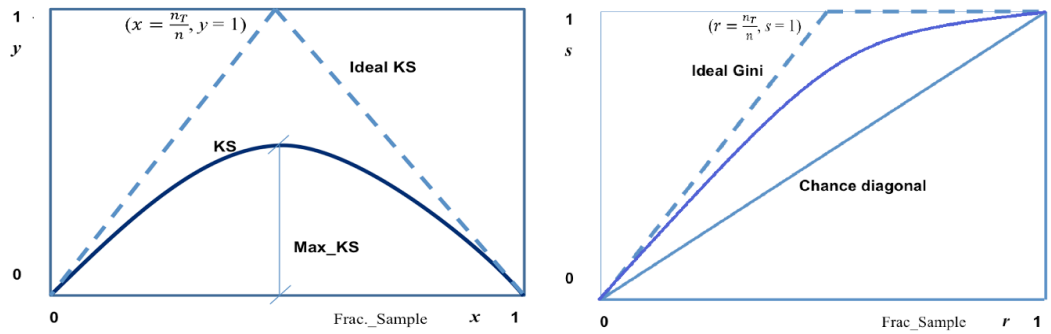

Figura 1. Curvas de Kolmogorov-Smirnov (esquerda) e de Lorenz (direita).

3. Prova de Equivalência das Áreas sob as Curvas

3.1. Motivação

Este artigo foi inspirado pela relação geométrica entre essas curvas e as do classificador ideal e pela prova de equivalência de Adeodato e Melo (2016) entre as curvas ROC e KS. A Figura 2 ilustra a semelhança geométrica entre as áreas, considerando a rotação e expansão do eixo-*x* da curva KS. A equivalência conceitual entre a abscissa da curva KS (eixo-*x*) e a diagonal da curva de Lorenz foi a base do raciocínio, juntamente com o caso limite do classificador ideal que separa perfeitamente os exemplos da classe alvo daqueles da classe complementar. A Figura 1 mostra que embora as formas de ambas as curvas ideais dependam da fração dos exemplos da classe alvo na amostra, a AUC_KS ideal independe dessa fração. Apesar da diferença dos classificadores ideais, há um mapeamento de 1-para-1 entre as suas curvas, com todos os *n*-1 possíveis classificadores ótimos para um conjunto de dados de tamanho *n*, contendo de 1 a *n*-1 exemplos da classe alvo sendo mapeados nos seus pontos correspondentes.

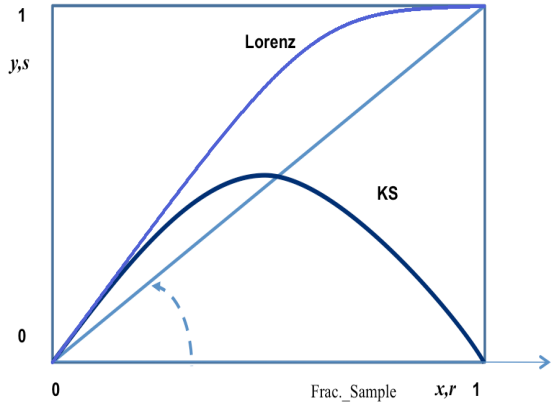

Figura 2. Transformação gráfica: Kolmogorov-Smirnov ➔ Lorenz.

3.2. Prova de Equivalência

Primeiro definamos as quantidades e frequências relativas necessárias para a construção das curvas KS e de Lorenz, em cada potencial ponto *i* de decisão: n é o tamanho da amostra; n_T é o número de exemplos da classe alvo; $n_{\bar{T}}$ é o número de exemplos da classe complementar; n_{T_i} é o número de exemplos da classe alvo até o *i*-ésimo exemplo; $n_{\bar{T}_i}$ é o número de exemplos da classe complementar até o *i*-ésimo exemplo; x, y são os eixos coordenados da curva KS; e r, s são os eixos coordenados da curva de Lorenz. No método de construção de gráficos em planilhas de cálculo, tanto para a curva KS quanto para a de Lorenz, depois de ordenar os exemplos rotulados de acordo com o escore de propensão, estes indicadores de desempenho são uma sequência de operações simples realizada sobre as frequências relativas da classe alvo, da classe complementar e da população.

A curva de Lorenz, por definição, é o conjunto de pontos representado pelos pares $(r_i, s_i) = \left(\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T}\right)$, enquanto a curva KS é o conjunto de pontos representado por $(x_i, y_i) = \left(\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T} - \frac{n_{\bar{T}_i}}{n_{\bar{T}}}\right)$, para $i = 1, \ldots, n$. Observe que a abscissa da curva de Lorenz é exatamente a mesma da curva KS, enquanto a sua ordenada é a mesma da curva ROC

correspondente [Provost e Fawcett, 2001]. A suposição mais simples para a equivalência das curvas é que uma seja uma aplicação linear da outra. Podem-se tomar os classificadores ideais correspondentes em ambas as métricas para determinar a matriz que produz este mapeamento linear: no caso da representação de Lorenz, a curva do classificador ideal é formada pelos segmentos de reta a partir de (r=0, s=0) até ($r = \frac{n_T}{n}$, s = 1), e de ($r = \frac{n_T}{n}$, s = 1) até (r=1, s=1), seguindo a ordem de escores descendentes, ilustrado em linhas tracejadas na Figura 1 (direita). A curva correspondente na representação KS é formada pelos segmentos de reta de (x=0, y=0) a ($x = \frac{n_T}{n}$, y = 1), e de ($x = \frac{n_T}{n}$, y = 1) a (x=1, y=0), ilustrados em linhas tracejadas na Figura 1 (esquerda). Assim, o vetor que é imagem de ($\frac{n_T}{n}$, 1) é ele próprio, enquanto a imagem de (x=1, y=0) é o vetor (r=1, s=1), fazendo com que a AUC_KS fique cisalhada e, portanto, não equivalente à área de Gini. Isto fica claro ao se considerar a área de Gini do classificador ideal, dada em função da probabilidade a priori da classe alvo: Área de Gini= $(1 - \frac{n_T}{n})/2$, que é menor que 0,5. No entanto, se se considera a razão de cada área pela área do classificador ideal, a equivalência se verifica (AUC_KS_Ratio = Gini_Index_Ratio). Isto é o que se precisa provar.

A partir das correspondências de vetores mencionadas acima, e com o uso de conceitos de Álgebra Linear, pode-se mostrar que a transformação resultante é dada por: $T(x_i, y_i) = (x_i, x_i + (1 - \frac{n_T}{n})y_i)$. Esta transformação é uma mudança de escala anisotrópica com fatores 1 e $1 - \frac{n_T}{n}$ seguida de um cisalhamento na direção da ordenadas, com fator 1. Pode-se notar que esta aplicação não preserva área, já que o seu determinante não é igual a 1 (o valor é $1 - \frac{n_T}{n}$). Todas as áreas de figuras planas são escaladas por este valor, portanto, quando se faz uma razão de áreas, este valor aparece tanto no denominador quanto no numerador, cancelando-se, daí por que a razão de áreas é preservada. Substituindo-se as expressões de KS do i-ésimo ponto (x_i ,y_i) na matriz de T segue que

$$T(\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T} - \frac{n_{\bar{T}_i}}{n_{\bar{T}}}) = (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n} + \left(1 - \frac{n_T}{n}\right)\left(\frac{n_{T_i}}{n_T} - \frac{n_{\bar{T}_i}}{n_{\bar{T}}}\right)) = (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n} + \frac{n_{T_i}}{n_T} - \frac{n_{T_i}}{n} - \frac{n_{\bar{T}_i}}{n_{\bar{T}}} + \frac{n_T . n_{\bar{T}_i}}{n . n_{\bar{T}}}) = (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{\bar{T}_i}}{n} + \frac{n_{T_i}}{n_T} - \frac{n_{\bar{T}_i}}{n_{\bar{T}}} + \frac{n_T . n_{\bar{T}_i}}{n . n_{\bar{T}}})$$

$$= (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T} + \frac{n_{\bar{T}} . n_{\bar{T}_i}}{n . n_{\bar{T}}} - \frac{n . n_{\bar{T}_i}}{n . n_{\bar{T}}} + \frac{n_T . n_{\bar{T}_i}}{n . n_{\bar{T}}}) = (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T} + \frac{(n_T + n_{\bar{T}}) . n_{\bar{T}_i} - n . n_{\bar{T}_i}}{n . n_{\bar{T}}})$$

$$= (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T} + \frac{n . n_{\bar{T}_i} - n . n_{\bar{T}_i}}{n . n_{\bar{T}}}) = (\frac{n_{T_i}}{n} + \frac{n_{\bar{T}_i}}{n}, \frac{n_{T_i}}{n_T}) = (r_i, s_i)$$

que representa o i-ésimo ponto (r_i ,s_i) na curva de Lorenz. Agora, com a equivalência entre as áreas normalizadas provada, a AUC_KS_Ratio pode ser usada para avaliação de desigualdades entre populações enquanto o índice de Gini_Index_Ratio pode ser usado para avaliar classificadores binários.

4. Conclusões

Este artigo demonstrou a equivalência entre o índice de Gini e a área sob a curva da distribuição Kolmogorov-Smirnov (AUC_KS), ambas normalizadas pelas áreas dos classificadores ideais. Isto é, Gini_Index_Ratio = AUC_KS_Ratio. Este resultado é muito importante porque associa o índice de Gini às áreas AUC_ROC e AUC_KS já provadas equivalentes. Isso dá ao cientista de dados fundamentos para a avaliação de desempenho de classificação binária com diferentes perspectivas e ferramentas.

Além disso, a AUC_KS_Ratio torna-se métrica de desempenho interessante para classificação binária: (i) é métrica baseada na área de toda a faixa de escore, (ii) é não paramétrica, dando o controle da decisão por limiar ao decisor, (iii) é o dobro da AUC_KS, (iv) a AUC_KS é fácil de calcular pelo método dos trapézios, (v) a AUC_KS_Ratio varia de 0 (caso aleatório) a 1 (caso ideal), (vi) a AUC_KS possibilita o cálculo direto de curvas médias, e (vii) as curvas de erro em validação cruzada *k-fold* podem ser projetadas no domínio ROC através da transformação linear entre KS e ROC, mais simples e precisa do que a média vertical ou de limiar [Fawcett, 2006].

Este trabalho ainda integra o conhecimento estatístico sobre a curva de Lorenz de 1912 [Ceriani e Verme, 2012] com o da distribuição KS disponível desde 1933 [Kolmogorov, 1933] e análise ROC utilizada em telecomunicações [Peterson *et al*., 1954], bem antes de sua estreia em inteligência artificial. Este trabalho abre a perspectiva de integrar todas essas três abordagens de estatística e de teoria da informação. A prova de equivalência iniciada por Krzanowski e Hand (2009) entre as curvas ROC e KS para métricas de ponto único, recentemente foi ampliada por Adeodato e Melo (2016) para métricas de área e agora foi estendida para a curva de Lorenz/índice de Gini trazendo nova perspectiva ao domínio para interpretar decisões binárias.

Referências Bibliográficas

Adeodato, P. J. L. e Melo, S. B. (2016) "On the equivalence between Kolmogorov-Smirnov and ROC curve metrics for binary classification". Cornell University Library ARXIV, 2016arXiv160600496A, https://arxiv.org/abs/1606.00496.

Adeodato, P. J. L. *et al*. (2008) "The Power of Sampling and Stacking for the PAKDD-2007 Cross-Selling Problem". Int. Jour. Data War. Mining, 4, pp. 22–31.

Bellù, L. G. e Liberati, P. (2006) "Inequality Analysis – The Gini Index". Food and Agriculture Organization, United Nations.

Ceriani, L. e Verme, P. (2012) "The origins of the Gini index: extracts from Variabilità e Mutabilità (2012) by Corrado Gini". J. Econ. Inequal. 10:421–443.

Conover, W. J. (1999) "Practical Nonparametric Statistics", (3rd ed.), John Wiley & Sons, New York, NY.

Fawcett, T. (2006) "An introduction to ROC analysis". Patt. Rec. Lett. 27, pp.861–874.

Kolmogorov, A. N. (1933) "Sulla determinazione empirica di una legge di distribuzione". Giornale dell'Istituto Italiano degli Attuari, 4, pp. 83–91.

Krzanowski, W. J. e Hand, D. J. (2009) "ROC Curves For Continuous Data", Chapman and Hall/CRC.

Peterson,W.W., Birdsall, T. G. e Fox, W. C. (1954) "The theory of signal detectability". In: Proc. of the IRE Professional Group on Information Theory 4, pp.171–212.

Provost, F. e Fawcett, T. (2001) "Robust Classification for Imprecise Environments". Machine Learning Journal, 42 (3), (Mar. 2001), pp. 203–231.

Provost, F. e Fawcett, T. (2013) "Data Science for business". O´Reilly Media Inc., Sebastopol, CA.