# Curvas ROC para avaliação de classificadores

R. C. Prati, G. E. A. P. A. Batista e M. C. Monard

***Resumo* — Gráficos ROC foram recentemente introduzidos como uma poderosa ferramenta para a avaliação de algoritmos de aprendizado. Apesar de gráficos ROC serem conceitualmente simples, existem algumas interpretações errôneas a seu respeito. Neste artigo, é feita uma introdução à análise ROC dentro do escopo de aprendizado de máquina e mineração de dados, ressaltando as vantagens de sua utilização bem como apontando os erros mais comuns quanto à sua interpretação e utilização.**

***Palavras-chave* — Curvas ROC (*ROC graphs*), Aprendizado de Máquina (*Machine Learning*), Mineração de Dados (*Data Mining*), Avaliação de Modelos (*Model Evaluation*).**

## I. Introdução

ANÁLISE ROC — do inglês *Receiver Operating Characteristic* — é um método gráfico para avaliação, organização e seleção de sistemas de diagnóstico e/ou predição. Gráficos ROC foram originalmente utilizados em detecção de sinais, para se avaliar a qualidade de transmissão de um sinal em um canal com ruído [1]. Gráficos ROC também são muito utilizados em psicologia para se avaliar a capacidade de indivíduos distinguirem entre estímulo e não estímulo [2]; em medicina, para analisar a qualidade de um determinado teste clínico [3, 4]; em economia (onde é conhecida como gráfico de Lorenz), para a avaliação de desigualdade de renda [5]; e em previsão do tempo, para se avaliar a qualidade das predições de eventos raros [6].

Recentemente, a análise ROC foi introduzida em Aprendizagem de Máquina — AM — e Mineração de Dados —MD — como uma ferramenta útil e poderosa para a avaliação de modelos de classificação [7, 8]. Ela é particularmente útil em domínios nos quais existe uma grande desproporção entre as classes ou quando se deve levar em consideração diferentes custos/benefícios para os diferentes erros/acertos de classificação. Análise ROC também tem sido utilizada para a construção [9] e refinamento de modelos [10]. Entretanto, apesar de gráficos ROC serem conceitualmente simples, existem várias interpretações errôneas a seu respeito.

O principal objetivo deste artigo é apresentar a análise ROC dentro do escopo de AM e MD em uma linguagem clara e objetiva, ressaltando as vantagens de sua utilização na avaliação de modelos de classificação. Também são apresentados alguns pontos que muitas vezes são mal interpretados dentro do contexto de AM e MD.

Este trabalho está organizado da seguinte maneira: na Seção II, são revisados alguns conceitos sobre probabilidade conjunta e condicional para a avaliação de modelos de classificação. Na Seção III, é discutida a avaliação de modelos, ressaltando a diferença entre classificação e ordenação. Na Seção IV, é apresentado o gráfico ROC propriamente dito, destacando a sua utilização em AM e MD. Finalmente, na Seção V são apresentadas as considerações finais.

## II. Probabilidade Conjunta e Condicional

Para induzir um classificador, um algoritmo de aprendizado supervisionado utiliza uma amostra de casos para os quais se conhece a classificação verdadeira. Cada caso é descrito por um conjunto de atributos. Para se distinguir casos entre as possíveis classificações, cada caso é rotulado com um atributo especial, denominado classe, cujos valores se referem à classificação verdadeira dos casos. Casos rotulados são chamados de exemplos, e a amostra utilizada pelo algoritmo de aprendizado para induzir o modelo de classificação é chamada de conjunto de exemplos de treinamento.

A seguir, restringiremos nossa discussão a problemas de classificação binária, *i.e.*, que tenham somente duas classes. Sem perda de generalidade, denominaremos as classes como **positiva** e **negativa**. Uma maneira natural de apresentar as estatísticas para a avaliação de um modelo de classificação é por meio de uma tabulação cruzada entre a classe prevista pelo modelo e a classe real dos exemplos. Essa tabulação é conhecida como tabela de contingência (também chamada de matriz de confusão). Na Tabela I(a), é mostrada uma matriz de contingência com freqüências absolutas (contagem). Quando um exemplo positivo é classificado como positivo, ele é denominado verdadeiro positivo. Quando um exemplo negativo é classificado como positivo, ele é denominado falso positivo. Nomenclatura similar é utilizada no caso dos exemplos classificados como negativos. Nessa tabela, $TP$, $FP$, $FN$ e $TN$ correspondem, respectivamente, às quantidades de verdadeiro/falso positivo/negativo. $PP$ e $PN$ correspondem ao número de exemplos preditos como positivos/negativos e $POS$ e $NEG$ ao número real de exemplos positivos/negativos na amostra. $N$ é o número de elementos da amostra.

Se dividirmos cada entrada na matriz mostrada na Tabela I(a) pelo tamanho da amostra, cada entrada dessa matriz representará uma estimativa da probabilidade conjunta da classe real do exemplo e da predição dada pelo modelo. Para se obter uma estimativa mais confiável, em amostras grandes é recomendável a utilização de um conjunto independente de exemplos de teste. Caso o tamanho da amostra seja pequena, geralmente utilizam-se métodos de reamostragem, tal como validação cruzada. Essa nova matriz é mostrada na Tabela I(b), na qual $X$ representa a variável aleatória **classe real do**

Os autores agradecem as agências de fomento brasileiras Capes, CNPq, FAPESP e FPTI/Br.

R.C. Prati é pós-doutorando no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP São Carlos). G. E. A. P. A. Batista e M. C. Monard são docentes no mesmo instituto. E-mails {prati, gbatista,mcmonard}@icmc.usp.br

**exemplo = positiva** e $Y$ representa a variável **classe predita do exemplo = positiva**. $\overline{X}$ e $\overline{Y}$ representam a negação de $X$ e $Y$.

TABELA I
MATRIZ DE CONTINGÊNCIA PARA MODELOS DE CLASSIFICAÇÃO

| | predito | | |
|---|---|---|---|
| real | $TP$ | $FN$ | $POS$ |
| | $FP$ | $TN$ | $NEG$ |
| | $PP$ | $PN$ | $N$ |

(a) Freqüência absoluta

| | $Y$ | $\overline{Y}$ | |
|---|---|---|---|
| $X$ | $p(X,Y)$ | $p(X,\overline{Y})$ | $p(X)$ |
| $\overline{X}$ | $p(\overline{X},Y)$ | $p(\overline{X},\overline{Y})$ | $p(\overline{X})$ |
| | $p(Y)$ | $p(\overline{Y})$ | 1 |

(b) Probabilidade conjunta

Como a matriz mostrada na Tabela I(a) é apenas uma re-escala da matriz mostrada na Tabela I(b), elas são equivalentes. Além disso, toda a informação necessária para avaliar o modelo está contida nessas matrizes. No entanto, uma análise mais refinada pode ser feita pela decomposição das probabilidades conjuntas em probabilidades condicionais e marginais. Probabilidades condicionais podem ser obtidas a partir das leis básicas de probabilidade:

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X),$$

na qual $P(X|Y)$ é a probabilidade condicional de $X$ ser verdade, dado que $Y$ é verdade. É importante ressaltar que $P(X,Y)=P(Y,X)$, mas $P(X|Y)\neq P(Y|X)$. Dado um conjunto de exemplos, todas essas probabilidades podem ser estimadas como proporções. Mais especificamente,

$$P(X|Y)=\frac{p(X,Y)}{p(Y)}=\frac{TP}{PP}, \text{ e } P(Y|X)=\frac{p(X,Y)}{p(X)}=\frac{TP}{POS}.$$

Essas probabilidades condicionais também podem ser calculadas para as quatro entradas da matriz e podem ser representadas como matrizes.

A probabilidade $P(X|Y)$ é importante para o usuário do modelo, uma vez que ela dá a probabilidade de que a classe seja positiva, dado que a previsão feita pelo modelo é positiva. Essa probabilidade também é conhecida como *confiança*. Entretanto, em termos de avaliação do modelo, $P(Y|X)$ é muito mais útil. Uma das vantagens da fatoração de $P(X,Y)$ em $P(Y|X)$ e $P(X)$ é que $P(Y|X)$ é condicional ao valor de $X$, *i.e*, é condicional à proporção de exemplos entre as classes. Essa probabilidade condicional é freqüentemente conhecida como *crença* ou *verossimilhança*, uma vez que ela especifica a probabilidade de que uma predição particular é feita dada a ocorrência de uma observação específica. Essa probabilidade indica o quanto um modelo é capaz de discriminar os casos entre as possíveis classes.

Além disso, as probabilidades marginais de $X$ são as únicas que não envolvem, de maneira alguma, as previsões do modelo. Por esse motivo, a distribuição de $X$ é geralmente assumida como uma característica do domínio e não dependente do modelo. Em outras palavras, a distribuição de $P(X)$ é assumida como fixa (não varia com o tempo) e pode ser estimada a partir do conjunto de treinamento. O valor de $P(X)$ é geralmente conhecido como prevalência da classe $X$ (classe positiva). Feita essa suposição, somente dois valores são necessários para descrever a matriz de contingência, pois $P(Y|X)=1-P(\overline{Y}|X)$ e $P(Y|\overline{X})=1-P(\overline{Y}|\overline{X})$. Essas duas probabilidades são independentes da proporção de exemplos *a priori* entre as classes (independente da prevalência de $X$). Como será abordado na Seção III, essa é uma propriedade importante na avaliação de modelos em domínios com classes desbalanceadas e/ou com diferentes custos de classificação.

## III. AVALIAÇÃO DE MODELOS

A avaliação de um modelo de classificação é baseada na análise da matriz de contingência (ou de suas derivações). Uma das maneiras mais comuns de avaliar modelos é a derivação de medidas que, de alguma maneira, tentam medir a "qualidade" do modelo. Essas medidas geralmente podem ser obtidas a partir da matriz de contingência, reduzindo suas quatro células principais a um único índice numérico de qualidade.

Reduzir a matriz de contingência a uma única medida tem algumas vantagens aparentes. A principal delas é que é mais fácil escolher o "melhor" em termos de um único valor. Entretanto, é comum encontrar casos em que uma dada medida é apropriada para um problema, mas ela é irrelevante para outros. Também, é comum encontrar situações em que a avaliação é um problema de múltiplas faces, nas quais é possível definir várias medidas, sendo perfeitamente possível que um modelo seja melhor que outro para algumas dessas medidas, mas pior com relação a outras. Nesses casos, utilizar uma única medida pode dar a falsa impressão de que o desempenho pode ser avaliado utilizando-se apenas essa medida.

Tomemos como exemplo a taxa de erro de classificação, uma das medidas mais comumente utilizadas em aprendizado de máquina. A taxa de erro pode ser calculada a partir da matriz de contingência da seguinte maneira:

$$Erro=P(\overline{Y},X)+P(Y,\overline{X})=\frac{FP+FN}{N}$$

Existem várias situações em que a taxa de erro de classificação não é apropriada para a avaliação de modelos de classificação. Uma situação comum se dá quando o número de exemplos em cada uma das classes é muito desbalanceado [11]. Por exemplo, suponha que em um dado domínio o número de exemplos de uma das classes seja 99%. Nesse caso, é comum se obter baixas taxas de erro, pois um modelo que sempre retorna a classe majoritária terá uma taxa de erro de apenas 1%. No entanto, esse modelo que sempre classifica um novo exemplo na classe majoritária não irá acertar nenhuma classificação de exemplos da classe minoritária.

Além disso, a taxa de erro assume custos iguais para os erros em ambas as classes. Em muitos domínios é comum haver diferentes custos de classificação para as diferentes

classes. Em medicina, por exemplo, o custo de classificar incorretamente um paciente doente como sadio para uma dada doença grave é muito maior do que classificar um paciente sadio como doente pois, no primeiro caso, a falha no diagnóstico pode levar à morte do paciente. Conhecendo-se os custos de classificação, esse problema pode ser remediado pela substituição da taxa de erro pelo custo médio esperado [12]. Entretanto, esses custos geralmente não são conhecidos, ou até mesmo podem variar, dependendo de fatores externos.

Um outro problema é que tanto a taxa de erro quanto o custo médio esperado são dependentes da distribuição de exemplos entre as classes. Na maioria das aplicações de AM e MD assume-se que os exemplos disponíveis para o treinamento constituem uma amostra representativa, isto é, independente e identicamente distribuída da população de casos. Nesse caso entende-se que a proporção de exemplos amostrados para cada uma das classes é equivalente à proporção da população de interesse [13]. Em outras palavras, assume-se que não existe nenhum vício na amostra e que a proporção de exemplos para cada uma das classes aproxima bem a proporção de casos na população da qual os exemplos foram amostrados. Essa suposição, mesmo que plausível, não é, necessariamente, correta. Suponha que queiramos construir um modelo para reconhecer se uma dada seqüência de aminoácidos corresponde ou não a um gene humano. Nesse caso, a amostra não pode ter a proporção real de exemplos, pois não se conhece o número total de genes que formam o genoma humano. Esse número já foi estimado em 100.000, 70.000 e 50.000, entre outros. A estimativa atual é entre 30.000 e 40.000. Desse modo, nem a taxa de erro nem o custo médio esperado têm o significado a eles atribuídos, pois variações na proporção de exemplos (diferente estimativas do número de genes) irão alterar os valores das medidas de desempenho, mesmo que o desempenho global do modelo não mude.

Essas deficiências não são exclusivas da taxa de erro. Qualquer medida que tenha como objetivo reduzir a avaliação de um modelo de classificação a um único valor terá, em maior ou menor grau, uma perda de informação. Geralmente, a não ser que se tenham domínios com critérios para avaliação claramente definidos e estáticos, a avaliação de um modelo utilizando uma única medida pode levar a conclusões errôneas. Em outras palavras, não existe uma única medida boa, a não ser que seja possível definir, para aquele domínio, o significado de bom.

Ainda que consideremos apenas o caso em que tanto observações quanto previsões são binárias, muitos sistemas de aprendizado fornecem um valor contínuo para as previsões, mesmo no caso em que as classes são discretas. Por exemplo, redes neurais geralmente produzem valores entre zero e um. Para se obter uma classificação, geralmente coloca-se um limiar na variável de predição. Todas as predições acima desse limiar são atribuídas a uma das classes, e as predições abaixo desse limiar são atribuídas à outra. Dessa maneira, para um dado limiar, é possível se obter uma previsão binária pela discretização da previsão contínua. Essa discretização contribui com um outro problema para a avaliação do modelo de classificação: a escolha do limiar é arbitrária e geralmente baseada na taxa de erro. Assim, cada possível limiar produz uma matriz de contingência diferente, e, por conseqüência, diferentes valores para as medidas.

Em termos de avaliação, ao invés de estabelecer um limiar de classificação e avaliar o modelo de classificação derivado desse limiar, é mais interessante avaliar como o modelo ordena os exemplos. Um bom modelo deve fazê-lo de tal maneira que, observando-se a variável de predição, exemplos de classes semelhantes sejam agrupados em faixas contínuas de valores. É importante ressaltar que, como é possível derivar uma classificação colocando-se um limiar na variável contínua, avaliar como os exemplos são ordenados é mais vantajoso, pois é independente do limiar e engloba a avaliação da classificação.

Também é importante ressaltar que, assim como na taxa de erro (e outras medidas), é possível derivar estatísticas a respeito da qualidade da ordenação. Uma dessas estatísticas é o teste de ordenação de Wilconxon. Como será visto na Seção IV, essa estatística tem uma correlação com a análise ROC. Entretanto, da mesma maneira que as medidas para se avaliar o modelo de classificação, avaliar a ordenação dos exemplos com uma única medida também tem as suas desvantagens.

## IV. O GRÁFICO ROC

Uma alternativa à avaliação utilizando medidas é o uso de gráficos e/ou diagramas. Gráficos permitem uma melhor visualização da multidimensionalidade do problema de avaliação. O gráfico ROC é baseado na probabilidade de detecção, ou taxa de verdadeiros positivos ( $tpr = P(Y|X)$ ), e na probabilidade de falsos alarmes, ou taxa de falsos positivos ( $fpr = P(Y|\overline{X})$ ). Para se construir o gráfico ROC plota-se $fpr$ no eixo dos ordenadas – eixo $x$ – e $tpr$ no eixo das abscissas – eixo $y$ .

Um modelo de classificação é representado por um ponto no espaço ROC. Para se obter o ponto no espaço ROC correspondente a um modelo de classificação, calcula-se a taxa de verdadeiros e falsos positivos ( $tpr$ e $fpr$ ) desse modelo a partir da sua matriz de contingência (vide Seção II e Tabela I).

Alguns pontos no espaço ROC merecem destaque. O ponto $(0,0)$ representa a estratégia de nunca classificar um exemplo como positivo. Modelos que correspondem a esse ponto não apresentam nenhum falso positivo, mas também não conseguem classificar nenhum verdadeiro positivo. A estratégia inversa, de sempre classificar um novo exemplo como positivo, é representada pelo ponto $(100\%,100\%)$. O ponto $(0,100\%)$ representa o modelo perfeito, *i.e.*, todos os exemplos positivos e negativos são corretamente classificados. O ponto $(100\%,0)$ representa o modelo que sempre faz predições erradas. Modelos próximos ao canto inferior esquerdo podem ser considerados "conservativos": eles fazem uma classificação positiva somente se têm grande segurança na classificação. Como conseqüência, eles cometem poucos erros falsos positivos, mas freqüentemente têm baixas taxas de verdadeiros positivos. Modelos próximos ao canto superior direito podem ser considerados “liberais”: eles predizem a classe positiva com maior freqüência, de tal maneira que classificam a maioria dos exemplos positivos

corretamente, mas, geralmente, com altas taxas de falsos positivos.

A linha diagonal ascendente $(0,0)$ - $(100\%,100\%)$ representa um modelo de comportamento estocástico: cada ponto $(p,p)$ pode ser obtido pela previsão da classe positiva com probabilidade $p$ e da classe negativa com probabilidade $100\%-p$. Pontos pertencentes ao triângulo superior esquerdo a essa diagonal representam modelos que desempenham melhor que o aleatório e pontos pertencentes ao triângulo inferior direito representam modelos piores que o aleatório. A diagonal descendente $(0,100\%)$ - $(100\%,0)$ representa modelos de classificação que desempenham igualmente em ambas as classes $(tpr=1-fpr=tnr)$. À esquerda dessa linha, estão os modelos que desempenham melhor para a classe negativa em detrimento da positiva e, à direita, os modelos que desempenham melhor para a classe positiva. O espaço ROC está representado esquematicamente na Fig. 1.

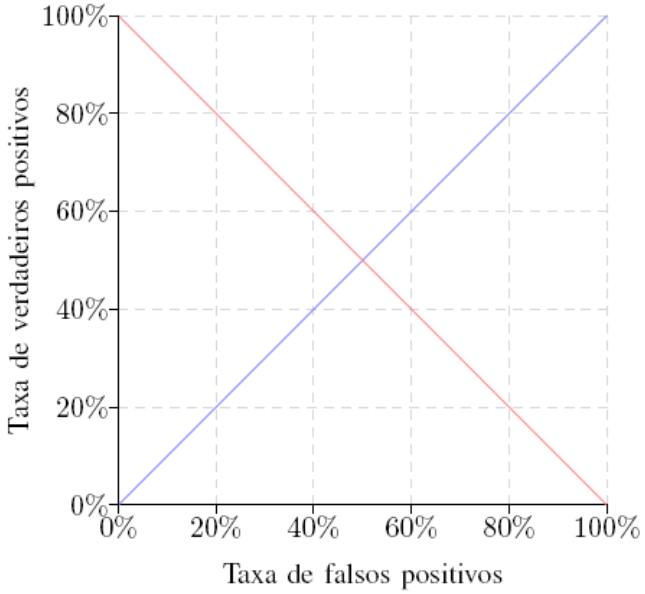
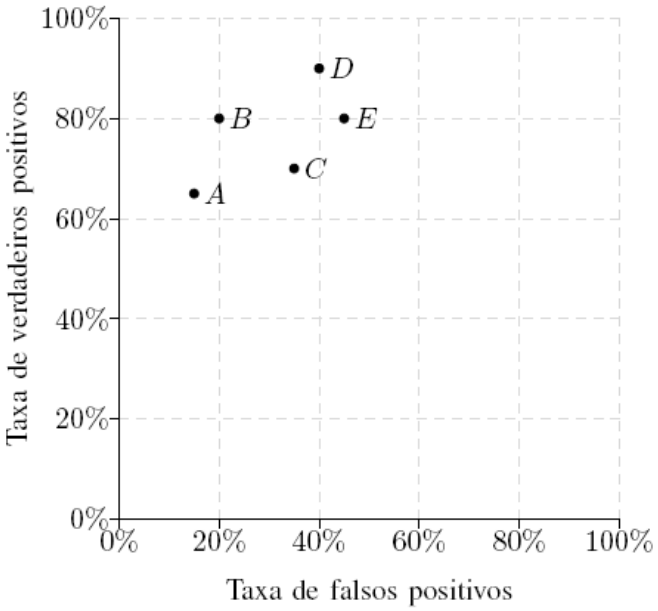
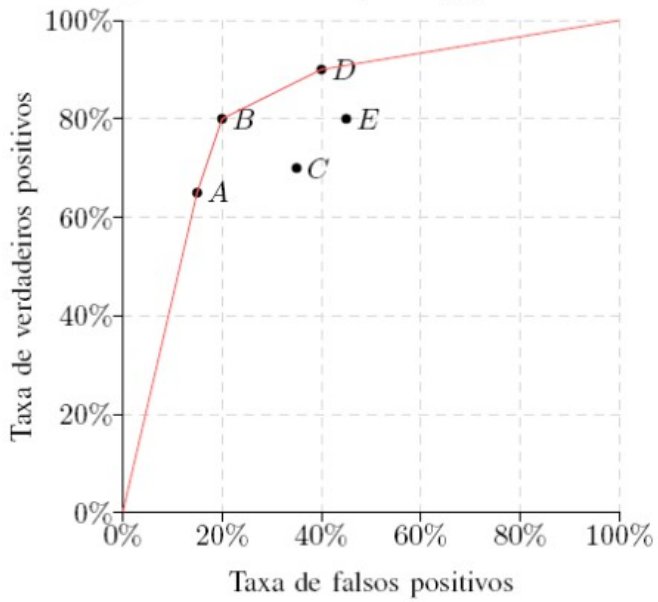



Fig. 1. O espaço ROC

Na Figura 2(a) é mostrado um gráfico ROC com 5 pontos arbitrários representando 5 modelos de classificação diferentes ($A$, $B$, $C$, $D$, e $E$), para um conjunto de dados também arbitrário. Uma rápida inspeção visual permite algumas constatações. Neste caso hipotético, $A$ é o mais conservativo e $D$ é o mais liberal. Além disso, é fácil de perceber que um ponto no espaço ROC é melhor que outro se e somente se ele está acima e à esquerda do outro ponto, *i.e.*, tem uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos [14].




(a) Modelos de classificação no espaço ROC




(b) Região convexa

Fig. 2. Modelos de classificação no espaço ROC

Além disso, é possível mostrar que os modelos que se encontram no envelope externo convexo (*convex hull*) que mais se aproxima ao ponto $(0,100\%)$, como mostrado na Figura 2(b), são os modelos que podem ser considerados ótimos, dada uma certa condição operacional. O termo condição operacional engloba fatores como proporção de exemplos *a priori* entre as classes e/ou custos/benefícios de classificação. Os outros modelos que não fazem parte do envelope convexo podem ser descartados [15, 16]. Isso se deve ao fato de que cada condição operacional define a inclinação de uma linha no espaço ROC, chamada de linha de *isodesempenho*. Ela recebe esse nome porque todos os pontos que fazem parte dessa linha têm uma característica em comum:

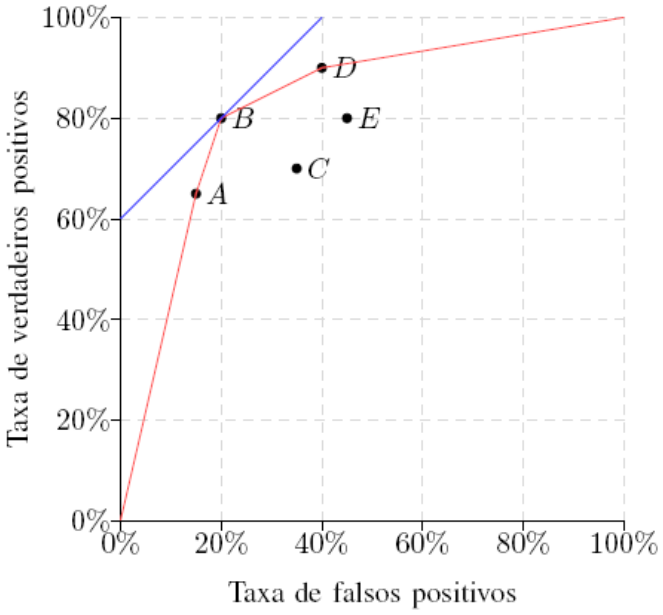
a taxa de erro (ou custo médio esperado) é a mesma. A inclinação dessa linha está relacionada a quanto um determinado erro é relativamente mais importante que o outro. O modelo ótimo para uma dada condição operacional deve estar em uma linha com essa inclinação. Além disso, o modelo ótimo deve estar o mais próximo possível do ponto $(0,100\%)$. Essas duas propriedades implicam que os modelos ótimos estejam no envelope convexo --- uma prova detalhada pode ser encontrada em [16].

Na Fig. 3 é mostrado o mesmo gráfico ROC da Fig. 2, com a adição de diferentes linhas de isodesempenho. A inclinação da linha de isodesempenho mostrada na Figura 3(a) é 1. Isso é equivalente a dizer que se essa linha representar a condição operacional verdadeira, a proporção de exemplos entre as classes (ou o custo de classificar erradamente um exemplo positivo ou negativo) é a mesma. Nessas condições, o modelo $B$ irá apresentar a menor taxa de erro (ou o menor custo de classificação). Já para o gráfico mostrado na Figura 3(b), a inclinação da curva é de 0,5. Se essa linha representar as condições operacionais verdadeiras, a classe positiva será duas vezes mais populosa (ou o custo de classificar erradamente um exemplo da classe positiva será duas vezes maior) que a classe negativa. Nessas condições, ambos os modelos, $B$ e $D$, são ótimos, *i.e.*, têm a mesma taxa de erro/custo de classificação global. Note que, no entanto, as taxas de erro separadas por classes são diferentes para cada um desses modelos.

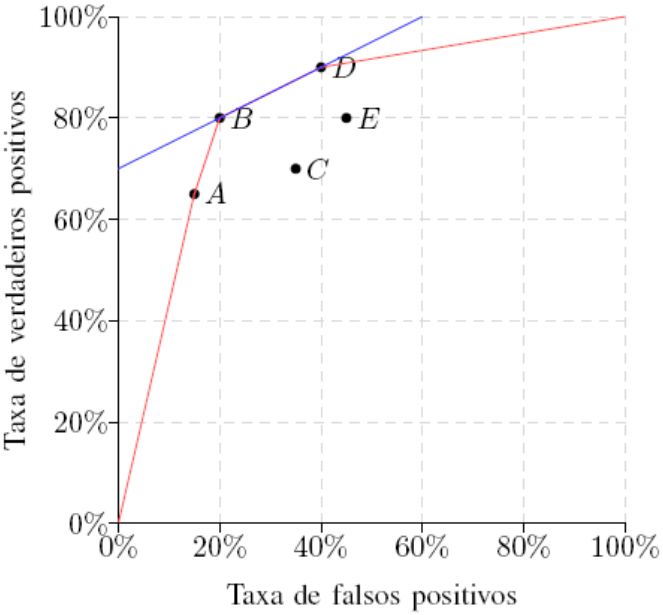
Uma outra vantagem de se utilizar a curva ROC está na avaliação da ordenação dos exemplos, ao invés da classificação. Nesse caso, como descrito na Seção III, o sistema de aprendizado não prediz a classe, mas um valor contínuo ou ordinal, e, para se criar o modelo de classificação, esse valor contínuo pode ser binarizado pela escolha de um limiar de classificação. Entretanto, quando se estabelece um limiar específico, assume-se um certo compromisso entre os acertos e os erros. Esse compromisso pode ser avaliado utilizando-se a análise de curvas ROC.

Ao invés de se escolher um limiar arbitrário e representar o desempenho do sistema para um dado domínio como um único ponto no espaço ROC, pode-se "simular" a escolha de vários limiares. Nesse caso, varia-se o limiar em todo o seu espectro, desde o valor mais restritivo até o valor mais liberal. Dessa maneira, a análise é feita independentemente da escolha do limiar. O desempenho do sistema é então representado por uma curva no espaço ROC – a curva ROC. A maneira mais eficiente de gerar essa curva é ordenar todos os casos de teste de acordo com o valor contínuo predito pelo modelo. A partir desse conjunto ordenado, para cada caso desse conjunto e seguindo-se essa ordem, dá-se um passo de tamanho $\frac{1}{POS}$ na direção do eixo $y$ se o exemplo for positivo ou um passo de tamanho $\frac{1}{NEG}$ caso o exemplo seja negativo.

Quanto mais distante a curva estiver da diagonal principal, melhor será o desempenho do sistema de aprendizado para aquele domínio. Ao se comparar duas (ou mais) curvas, caso não haja nenhuma intersecção, a curva que mais se aproxima do ponto $(0,100\%)$ é a de melhor desempenho. Caso haja intersecções, cada um dos sistemas tem uma faixa operacional na qual é melhor que o outro. Idealmente, a curva deve ser convexa e sempre crescente.




(a) Custos iguais para exemplos positivos e negativos.

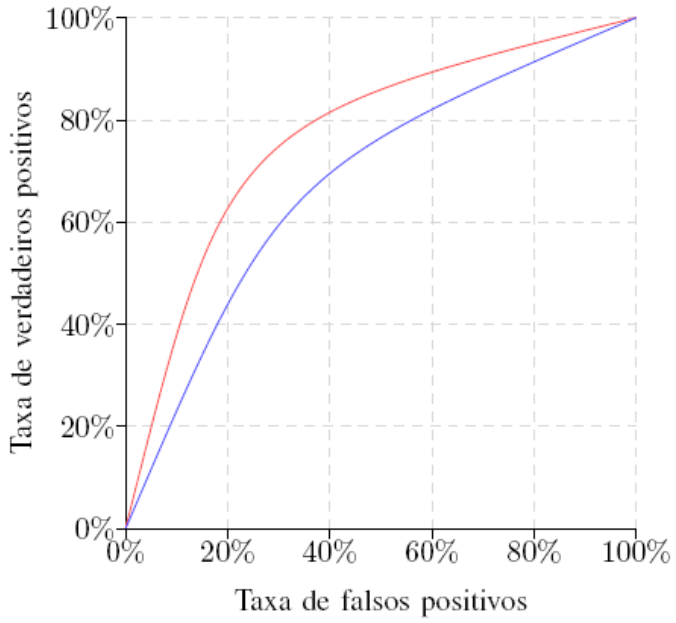



(b) A classe positiva é duas vezes mais custosa que a classe negativa.

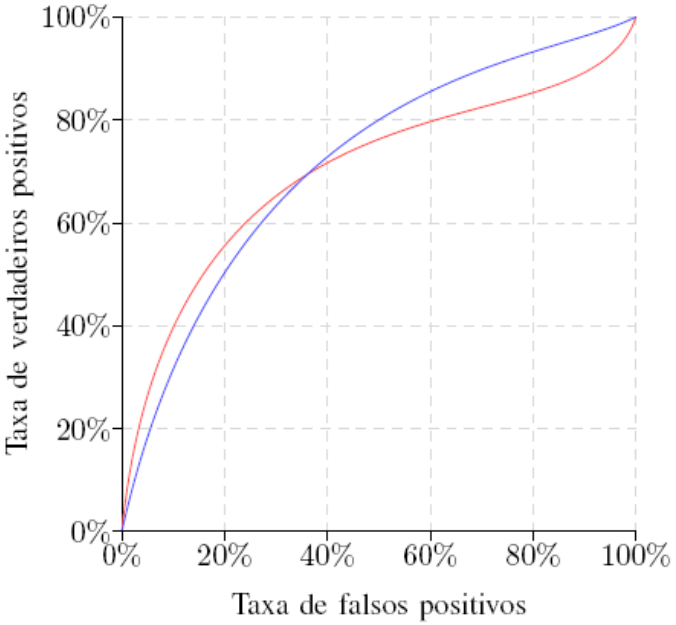
Fig. 3. Diferentes linhas de isodesempenho implicam diferentes modelos ótimos.

Na Figura 4, são mostrados alguns exemplos de curvas ROC. Note que na Figura 4(a) não há intersecções entre as curvas. Nesse caso, modelos de classificação derivados a partir da curva mais próxima do ponto $(0,100\%)$ serão sempre melhores do que os modelos derivados a partir da outra curva, independentemente das condições operacionais. Já no caso da Figura 4(b), em que há uma intersecção entre as curvas perto do ponto $(40\%,70\%)$, os modelos de classificação

derivados à esquerda desse ponto serão os melhores se forem derivados da curva que sobe mais próxima do eixo $y$, e da outra curva se forem derivados à direita.




(a) Curva sem intersecção.




(b) Curva com intersecção.

Fig. 4. Exemplos de curvas ROC.

Caso um único modelo seja realmente necessário, a sua derivação pode ser feita com base na análise da curva ROC, analisando-se os possíveis compromissos entre as classificações positivas e negativas. A idéia básica é a mesma das linhas de iso-desempenho. Dada uma condição operacional (e por conseqüência uma linha de iso-desempenho), é possível derivar o limiar apropriado para aquela condição operacional. Dessa maneira, é possível calibrar o modelo para as condições operacionais reais, ao invés daquelas utilizadas durante a indução. Esse fato é o segundo grande benefício da análise ROC.

Em suma, a primeira vantagem da análise ROC é que se pode fazer uma análise independentemente de certas condições, tais como o limiar de classificação, os custos relacionados às classificações errôneas e à distribuição *a priori* das classes. Essa análise é realizada visualizando-se o compromisso entre $tpr$ e $fpr$ em um gráfico bidimensional. A segunda vantagem é que a análise ROC pode ser utilizada para a calibração e ajuste de modelos de classificação, quando necessário.

Essas duas propriedades são às vezes mal interpretadas. É comum encontrar afirmativas de que a análise ROC é independente da distribuição de exemplos entre as classes. Na realidade, essa não é uma característica da análise ROC em si, mas uma conseqüência da aplicação das leis básicas da probabilidade citadas na Seção II. Ela somente é verdadeira se, como dito na Seção III, assumirmos a distribuição de $X$ como característica do domínio. A análise ROC é sim invariante à proporção de exemplos entre as classes desde que essa suposição seja verdadeira. Entretanto, essa propriedade não é válida em casos em que essa suposição não é válida, tais como em algumas aplicações em que há mudança de conceito (concept drift [17]). Muitas vezes, essa confusão é empregada como uma crítica à análise ROC [18]. No entanto, essa má interpretação está relacionada muito mais com a metodologia (decorrente de considerar verdadeira uma falsa suposição) do que com o método. Quando propriamente utilizada, a análise ROC é uma ferramenta poderosa para a avaliação de sistemas de aprendizado, principalmente em domínios para os quais não se podem definir as condições operacionais com precisão [19].

Uma outra conexão entre a curva ROC e a ordenação dos exemplos está relacionada com a área abaixo da curva ROC --- AUC (do inglês *Area Under Curve*). Uma vez que a área abaixo da curva ROC é uma fração da área de um quadrado de lado um, o seu valor está sempre entre 0 e 1. Em [20] é mostrado que essa área é numericamente equivalente à estatística de Wilconxon. Além disso, a AUC também é numericamente igual à probabilidade de, dados dois exemplos de classes distintas, o exemplo positivo seja ordenado primeiramente que um exemplo negativo [1]. A AUC também está correlacionada com o coeficiente Gini [21].

A AUC vem gradativamente ganhando espaço como medida de avaliação de modelos em aprendizado de máquina e mineração de dados. Apesar de, como discutido na Seção III, a avaliação de um modelo por uma única medida não ser a mais apropriada, a AUC tem menos deficiências do que a taxa de erro de classificação [22]. Entretanto, sempre que possível, é recomendável plotar e analisar a curva dos modelos.

Também é importante ressaltar que, assim como a taxa de erro de classificação, a AUC e a curva ROC são variáveis aleatórias e, portanto, devem ser estimadas e acompanhadas de alguma medida de variação. A AUC é normalmente estimada da mesma maneira que a taxa de erro (normalmente utilizando-se validação cruzada). Estimar a curva em si é um pouco mais complexo, uma vez que ela é bivariada, ou seja, envolve as duas variáveis, $tpr$ e $fpr$. Nesse caso, pode-se calcular tanto a variância com relação a $tpr$, $fpr$ ou ambas. Em [13], é apresentada uma descrição desses três métodos para se

estimar a curva média. Na Fig. 5, é mostrada uma curva ROC com estimativa de variância feita com o pacote ROCR (que pode ser obtido em http://bioinf.mpi-sb.mpg.de/projects/rocr/), desenvolvida para o software R (que pode ser obtido em http://www.r-project.org). O conjunto de linhas corresponde a diversas curvas calculadas com diferentes conjuntos de teste da validação cruzada, mostrando a variância da curva com respeito às duas variáveis.

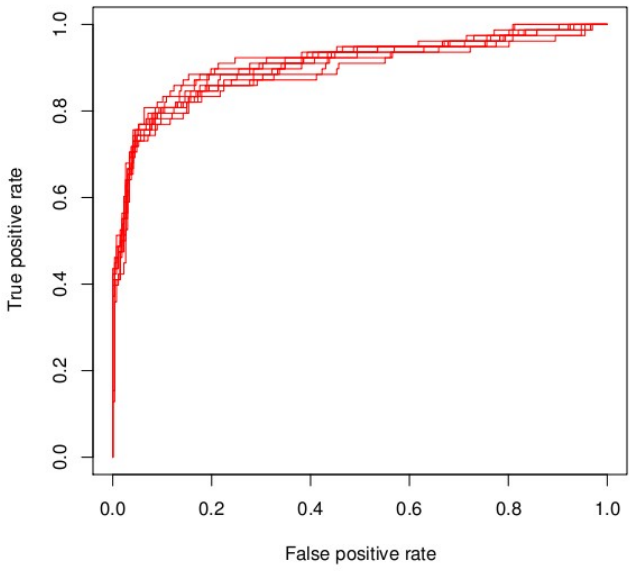



Fig. 5. A curva ROC com variância

Uma das principais desvantagens de se utilizar gráficos ROC é a sua limitação para apenas duas classes. Apesar de os princípios básicos serem os mesmos, o número de eixos cresce exponencialmente com o número de classes. No entanto, para reduzir a análise para duas classes, algumas aproximações são possíveis, como a um-contra-todos .

## V. Conclusões

Gráficos ROC constituem uma ferramenta muito útil para a visualização e avaliação de modelos de classificação. Eles também são utilizados para se avaliar como um sistema de aprendizado é capaz de ordenar os exemplos, permitindo uma análise independente do limiar de classificação. Caso seja necessário derivar um modelo de classificação, a análise ROC também permite que seja feita a calibração, por meio de linhas de iso-desempenho, para as condições operacionais mais apropriadas ao domínio da aplicação.

A análise ROC provê uma avaliação mais rica do que simplesmente avaliar o modelo de classificação a partir de uma única medida. Entretanto, para a sua correta utilização, é necessário que se conheçam suas características e limitações. Neste trabalho foi feita uma introdução à análise ROC dentro do contexto de aprendizado de máquina, bem como foram destacadas e exemplificadas algumas de suas limitações e erros de interpretação. Esperamos que os pontos aqui levantados sejam úteis para difundir ainda mais a utilização da análise ROC dentro da comunidade de AM e MD na América Latina.

## Referências


[1] J. P. Egan, *Signal detection theory and ROC analysis*. New York, USA: Academic Press, 1975.

[2] D. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. Los Altos, USA: Peninsula Publishing, 1989.

[3] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons Inc, 2002.

[4] A. C. Silva, M. Gattass, P. C. P. Carvalho, "Diagnosis of Solitary Lung Nodule Using Texture and Geometry in Computerized Tomography Images: Preliminary Results", IEEE Latin America Transactions, Vol. 2, No. 2, pp. 75-80, 2004.

[5] J. L. Gastwirth, "A general definition of the Lorenz curve", *Econometrica*, vol. 39, no. 6, pp. 1037–39, 1971.

[6] K. R. Mylne, "Decision-making from probability forecasts based on forecast value", *Meteorological Applications*, vol. 9, pp. 307–315, 2002.

[7] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[8] K. A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning", in *Proceedings of the 6th Int Workshop on Machine Learning (ICML'1989)*. Morgan Kaufmann, 1989, pp. 160–163.

[9] R. C. Prati and P. Flach, "ROCCER: an algorithm for rule learning based on ROC analysis", in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'2005)*, Edinburgh (UK), 2005, pp. 823–828.

[10] P. Flach and S. Wu, "Repairing concavities in ROC curves", in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'2005)*, Edinburgh (UK), 2005, pp. 702–707.

[11] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data", *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, 2004.

[12] C. Elkan, "The foundations of cost-sensitive learning", in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2001, pp. 973–978.

[13] J.L. Fleiss, *Statistical Methods for Rates and Proportions*, 2nd ed. New York (USA): John Wiley & Sons, 1981.

[14] T. Fawcett, "An introduction to ROC graphs", *Pattern Recognition Letters,* vol. 27, no. 8, pp 861-874, 2006.

[15] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms", in *Proceedings Fifteenth International Conference on Machine Learning (ICML'1998)*. WI: Morgan Kaufmann, San Francisco, CA, 1998, pp. 445–453.

[16] F. Provost and T. Fawcett, "Robust classifiers for imprecise environments", *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.

[17] J. Gama and G. Castillo, "Learning with local drift detection", in *Second International Conference on Advanced Data Mining and Applications (ADMA'2006)*,

ser. Lecture Notes in Computer Science, vol. 4093. Springer, 2006, pp. 42–55.

[18] G.I. Webb and K. M. Ting, "On the application of ROC analysis to predict classification performance under varying class distributions", *Machine Learning*, vol. 58, no.1, pp. 25–32, 2005.

[19] T. Fawcett and P. A. Flach, "A response to Webb and Ting's On the application of ROC analysis to predict classification performance under varying class distributions", *Machine Learning*, vol. 58, no. 1, pp. 33–38, 2005.

[20] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, vol. 143, pp. 29–36, 1982.

[21] D. J. Hand and R.Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems", *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[22] C. X. Ling, J. Huang, and H. Zhang, "AUC: a statistically consistent and more discriminating measure than accuracy", in *IJCAI*, G. Gottlob and T. Walsh, Morgan Kaufmann, 2003, pp. 519–526.

**Ronaldo Cristiano Prati** é doutor em Ciências da Computação e Matemática Computacional pela Universidade de São Paulo (2006). Atualmente, é pós-doutorando no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC/USP).

**Gustavo Enrique de Almeida Prado Alves Batista** possui doutorado em Ciências de Computação e Matemática Computacional pela Universidade de São Paulo (2003). Atualmente, é professor doutor do ICMC/USP.

**Maria Carolina Monard** é doutora em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (1980), livre-docente pela Universidade de São Paulo (1986) e pós-doutora pela University of Strathclyde (1987). Atualmente, é professora titular do ICMC/USP, coordenadora do Laboratório de Inteligência Computacional (LABIC) e líder do grupo de pesquisa em Inteligência Computacional.