COMPUTAÇÃO

Métricas de avaliação em machine learning

Acurácia, sensibilidade, precisão, especificidade e F-score

By **Diego Mariano**

19 de junho de 2021

Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e F-score

Diego Mariano

Revisão: Joicymara S. Xavier
BIOINFO – Revista Brasileira de Bioinformática. Edição #01. Julho, 2021.
DOI: 10.51780/978-6-599-275326-15

Ao construir um classificador usando *machine learning*, um desenvolvedor deve se perguntar o quão bom é seu modelo para predição. Assim, ao treinar um modelo de aprendizagem algumas métricas podem ser utilizadas para avaliação. A métrica utilizada para determinação do "melhor modelo" depende do problema analisado. Neste artigo, veremos as principais métricas para avaliação de modelos de classificação de dados, como acurácia, sensibilidade (*recall* ou revocação), especificidade, precisão e *F-score* (Tabela 1).

Método	Fórmula
Sensibilidade	VP / (VP+FN)
Especificidade	VN / (FP+VN)
Acurácia	(VP+VN) / N
Precisão	VP / (VP+FP)
F-score	2 x (PxS) / (P+S)

Tabela 1. Visão geral das métricas usadas para avaliar métodos de classificação. VP: verdadeiros positivos; FN: falsos negativos; FP: falsos positivos; VN: verdadeiros negativos; P: precisão; S: sensibilidade; N: total de elementos. Fonte: adaptado de Mariano (2019) [1].

Introdução

Um modelo de classificação de dados visa realizar uma previsão com base em ocorrências passadas. Para isso, o modelo utiliza um conjunto de dados com entradas (indivíduos) e atributos (propriedades). Além disso, é necessário conhecer o resultado esperado para esse conjunto de dados (rótulos). Todas essas informações serão usadas para treinar um modelo que será utilizado para predizer resultados esperados para novos dados que surgirem no futuro. Ao treinar esse modelo deve-se utilizar um conjunto de dados (não usados no treinamento) para testar o quanto o modelo acerta. Entretanto, não basta apenas contar a quantidade de acertos que seu modelo teve para dizer se ele é bom ou não. Dependendo do problema estudado, métricas diferentes devem ser utilizadas para essa avaliação. Entretanto, antes de apresentarmos essas métricas, precisamos entender alguns conceitos para classificação binárias: as classes que os dados preditos poderão receber.

Classes de dados preditos: VP, VN, FP e FN

Em um problema de classificação, há duas soluções possíveis: acerto ou erro. Entretanto, para um problema de classificação binária temos ainda duas outras classes possíveis, vamos chamá-las de classes positiva e negativa (elas podem receber quaisquer nomes). Por exemplo, digamos que desejamos construir um programa para predizer se irá chover. Os dias de chuva serão nossa classe positiva. Os dias sem chuva serão nossa classe negativa. Após construir nosso modelo, vamos usá-lo para predizer se amanhã poderemos ir à praia. Nosso modelo poderá dizer se irá chover ou não. No outro dia vamos a praia e observamos se choveu ou não, assim vemos se o programa acertou ou errou. Logo, há quatro resultados possíveis (Figura 1):

1. O programa disse que vai chover (positivo) e realmente choveu (predição verdadeira);
2. O programa disse que vai chover (positivo), mas não choveu (predição falsa);
3. O programa disse que não vai chover (negativo) e realmente não choveu (predição verdadeira);
4. O programa disse que não vai chover (negativo), mas choveu (predição falsa).

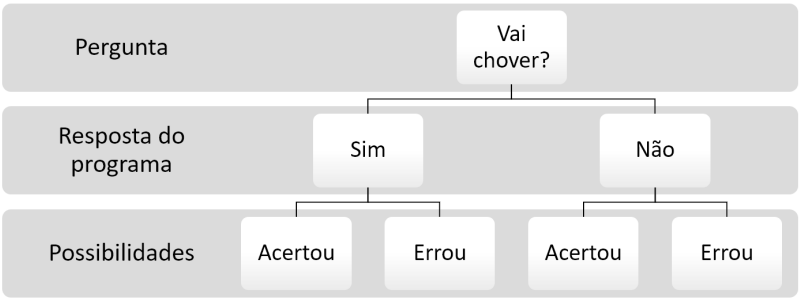

Figura 1. Resultados possíveis para um programa que realiza a previsão do tempo. Fonte: próprio autor.

De acordo com Ferrari & Silva (2017) [2], em problemas de classificação binária, predições podem ter quatro possíveis classes

- **Verdadeiro positivo (VP):** quando o método diz que a **classe é positiva** e, ao verificar a resposta, vê-se que **a classe era** realmente **positiva**;
- **Verdadeiro negativo (VN):** quando o método diz que a **classe é negativa** e, ao verificar a resposta, vê-se que **a classe era** realmente **negativa**;
- **Falso positivo (FP):** quando o método diz que a **classe é positiva**, mas ao verificar a resposta, vê-se que **a classe era negativa**;
- **Falso negativo (FN):** quando o método diz que a **classe é negativa**, mas ao verificar a resposta, vê-se que **a classe era positiva**;

Matriz de confusão

Uma maneira simples de se representar os resultados de um método de classificação de dados é através da chamada matriz de confusão (Tabela 2).

Matriz de confusão		**Classe predita**	
		Positiva	**Negativa**
Classe original	**Positiva**	VP	FN
	Negativa	FP	VN

Tabela 2. Matriz de confusão. Muitos autores costumam utilizar as siglas TP e TN (do inglês true positive e true negative) como sinônimos para VP e VN, respectivamente. Fonte: adaptado de Ferrari & Silva (2017) [2].

A matriz de confusão indica a quantidade de ocorrências que o programa teve para cada uma das quatro categorias.

Para ilustrar isso, digamos que nosso programa de predição de chuva foi usado durante 100 dias. Dos 100 dias, o programa disse que iria chover em 55 e que não iria chover nos outros 45 dias. Entretanto, após os 100 dias, percebemos que choveu em 50 e não choveu nos outros 50 dias. Vamos observar a matriz de confusão dos resultados do nosso programa (Tabela 3):

Matriz de confusão		**O que o programa disse:**	
		Vai chover	**Não vai chover**
O que aconteceu de verdade:	**Choveu**	40	10
	Não choveu	15	35

Tabela 3. Matriz de confusão que avalia o modelo de predição de chuva (n = 100). Fonte: próprio autor.

Com base nessa tabela, vemos que:

- **VP = 40:** o programa disse que em 40 dos 100 dias **iria chover** e realmente **choveu**.
- **FP = 15:** o programa disse que em 15 dos 100 dias **iria chover**, mas **não choveu**.
- **FN = 10:** o programa disse que em 10 dos 100 dias **não iria chover**, mas **choveu**.
- **VN = 35:** o programa disse que em 35 dos 100 dias **não iria chover** e **realmente não choveu.**

Veja que a soma dos valores dos quatro campos da tabela (VP = 40, FP = 15, FN = 10, VN = 35) deve ser igual ao total de dias (n = 100). Logo:

$$n = VP + VN + FP + FN \tag{1}$$

Para obtermos o total de predições realizadas em cada classe, podemos somar os valores presentes em cada coluna. Observe como obtemos o total de valores preditos como positivos:

$$pred_p = VP + FP \tag{2}$$

Para obter os valores preditos como negativos, usamos:

$$pred_n = VN + FN \tag{3}$$

Para obtermos o total de valores reais somamos os valores em cada linha. Assim, para obter os valores reais positivos, calculamos:

$$real_p = VP + FN \tag{4}$$

Para obter os valores reais negativos, calculamos:

$$real_n = VN + FP$$

(5)

Para calcular acertos do método, usamos:

$$acertos = VP + VN$$

(6)

Para calcular os erros, usamos:

$$erros = FP + FN$$

(7)

Agora veremos as métricas que podem ser utilizadas para avaliar a qualidade do classificador. São elas: acurácia, sensibilidade, especificidade, precisão e F-score [3].

Acurácia

A **acurácia** (*accuracy* ou ACC) é considerada uma das métricas mais simples e importantes. Ela avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas:

$$acurácia = \frac{Total\ de\ acertos}{Total\ de\ itens}$$

(8)

Utilizando como base a matriz de confusão, podemos obter a acurácia pela fórmula:

$$acurácia = \frac{VP + VN}{VP + FN + VN + FP}$$

(9)

Sensibilidade

Outra métrica que pode ser utilizada é a **sensibilidade** (também conhecida como *recall* ou revocação). Essa métrica avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. Ela pode ser obtida pela equação:

$$sensibilidade = \frac{VP}{VP + FN} \tag{10}$$

Especificidade

Por outro lado, a **especificidade** avalia a capacidade do método de detectar resultados negativos. Podemos calculá-la usando a equação:

$$especificidade = \frac{VN}{VN + FP} \tag{11}$$

Precisão

A **precisão** é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos:

$$precisão = \frac{VP}{VP + FP} \tag{12}$$

F-score

F–measure, F-*score* ou *score* F_1 é uma média harmônica calculada com base na precisão e na revocação. Ela pode ser obtida com base na equação:

$$f1 = 2 * \frac{precisão * sensibilidade}{precisão + sensibilidade}$$

(13)

Exemplo: previsão do tempo durante 100 dias

Para o exemplo apresentado anteriormente de um sistema de previsão do tempo, temos:

- n = 100
- VP = 40
- FP = 15
- FN = 10
- VN = 35

Vamos então calcular as métricas para nosso sistema de previsão do tempo:

$$acurácia = \frac{40 + 35}{100} => acurácia = 0{,}75$$

Vemos que nosso sistema possui uma acurácia de 0,75 (ou 75%). Vamos a seguir analisar a precisão, sensibilidade, especificidade e F-score (F_1):

$$precisão = \frac{40}{40 + 15} => precisão = \sim 0{,}73$$

$$sensibilidade = \frac{40}{40 + 10} => sensibilidade = 0{,}8$$

$$especificidade = \frac{35}{35 + 15} => especificidade = 0{,}7$$

$$F_1 = 2 * \frac{0{,}73 * 0{,}8}{(0{,}73 + 0{,}8)} => F_1 = 0{,}76$$

Podemos ver que nosso sistema de previsão de chuva possui como métrica mais alta a sensibilidade. Entretanto, todas as métricas avaliadas apresentam um resultado próximo, variando de 0,7 a 0,8.

Curva ROC

A curva ROC, do inglês *Receiver Operating Characteristic Curve*, ou na tradução "Curva Característica de Operação do Receptor" é um gráfico que permite avaliar um classificador binário. Essa visualização leva em consideração a taxa de verdadeiros positivos (TVP; ou sensibilidade) e a taxa

de falsos positivos (TFP; ou 1 – especificidade). Essas taxas também podem ser referidas pelas siglas TPR (*True Positive Rate*) e FPR (*False Positive Rate*), respectivamente. Esse gráfico permite comparar diferentes classificadores e definir qual o melhor com base em diferentes pontos de corte. Na prática, quanto mais próximo do topo do eixo Y melhor o classificador (Figura 2).

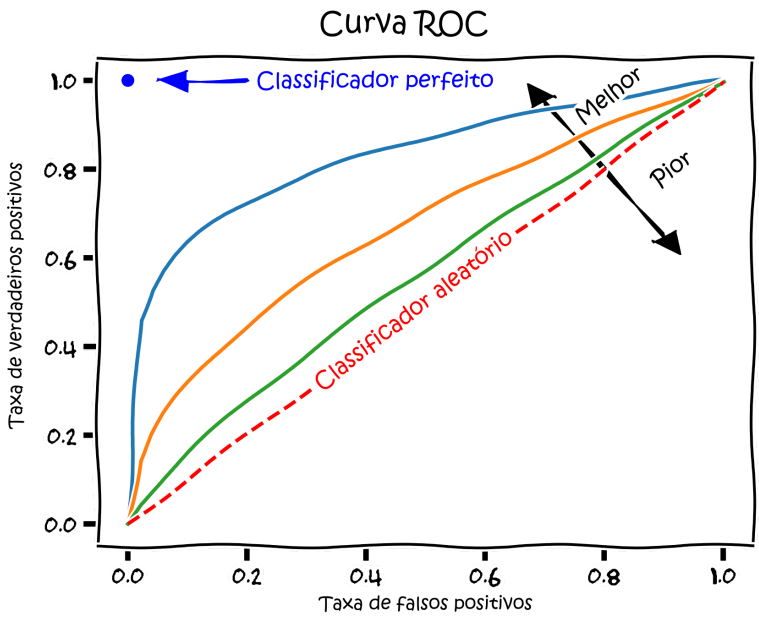

Figura 2. Ilustração de uma curva ROC. O eixo Y armazena a taxa de verdadeiros positivos (sensibilidade). O eixo X armazena a taxa de falsos positivos (1 – especificidade). O ponto azul representa um classificador perfeito, isto é, um classificador que atinge 100% de verdadeiros positivos e 0% de falsos positivos. A linha azul claro indica um resultado melhor do que os apresentados pelas linhas laranja e verde. A linha tracejada vermelha indica o limiar aleatório. Resultados abaixo da linha diagonal vermelha são considerados classificadores ruins. Fonte: adaptado e traduzido de MartinThoma (CC0 1.0 domínio público).

Uma curva ROC pode ser avaliada pela métrica AUC (*Area Under the Curve* ou "área sob a curva"). AUC calcula a área da forma bidimensional formada abaixo da curva. Essa métrica indica a probabilidade de duas previsões serem corretamente ranqueadas. A AUC será um valor entre 0 e 1. Quanto maior esse valor, melhor a capacidade do modelo em separar classes [4].

Quando usar cada uma das métricas

Ao usar *machine learning* para solução de problemas reais, obviamente desejamos construir classificadores perfeitos, que sempre acerta, mas no mundo real quase nunca isso é possível. Assim, ao construir um preditor, devemos visar o melhor resultado possível. Uma maneira simples de observar o quão bom é um modelo de classificação é usando a acurácia. A acurácia pode ser considerada uma métrica que nos dá uma visão geral do resultado, uma vez que ela mede o total de acertos considerando o total de

observações. Entretanto, outras métricas podem ser importantes dependendo de como o problema foi modelado.

O uso de cada métrica depende do objetivo do modelo que se deseja criar [3]. Por exemplo, suponha que desejamos criar um sistema que faça a detecção automática de spam. Nesse caso, um falso positivo pode ser considerado um problema mais crítico (uma mensagem importante ser considerada spam pode causar prejuízos ao usuário do sistema). Logo, a melhor métrica para comparação entre diferentes sistemas de detecção de spam seria a precisão.

Agora imagine um sistema que detecta falhas em um avião. Imagine que uma peça apresenta problemas, mas o sistema indica que não há nada errado. Isso poderia colocar vidas em perigo. Logo, para este exemplo um falso negativo seria um problema crítico. Portanto, um sistema construído para esse propósito deve levar em consideração uma taxa de falsos negativos próxima a zero. Uma métrica que poderia ser utilizada para comparar sistemas diferentes seria a sensibilidade. Valores altos de sensibilidade indicam altos valores de verdadeiros positivos mesmo quando se leva em conta o total de falsos negativos [3].

Referências

[1] MARIANO, D. C. B. Uso de assinaturas estruturais para proposta de mutações em enzimas β-glicosidase usadas na produção de biocombustíveis. 2019.

[2] FERRARI, D. G.; DE CASTRO SILVA, L. N. **Introdução a mineração de dados**. [s.l.] Saraiva Educação S.A., 2017.

[3] MARIANO, DIEGO; PAZ, F. J. . Data Mining. 1. ed. Porto Alegre: Sagah, 2020.

[4] Silva, Marcos. Matriz de Confusão e AUC ROC. Data Hackers – Medium. Disponível em: https://medium.com/data-hackers/matriz-de-confus%C3%A3o-e-auc-roc-f7e446dca107. 2019.