

Ensino Einstein



ALBERT EINSTEIN
INSTITUTO ISRAELITA DE
ENSINO E PESQUISA
CENTRO DE EDUCAÇÃO EM SAÚDE
ABRAM SZAJMAN

Métricas e Indicadores de Avaliação para Modelos de Classificação

Professor Rodrigo Signorini



Métricas e Indicadores de Avaliação para Modelos de Classificação

A **Matriz de Confusão** é muito utilizada para avaliações de modelos de classificação em *Machine Learning* e, sem dúvida, é o primeiro conjunto de métricas que temos que observar. É composta por quatro delas: **True Positive (TP)**, **False Negative (FN)**, **False Positive (FP)** e **True Negative (TN)**:

True Positive (TP) 150	False Negative (FN) 99
False Positive (FP) 60	True Negative (TN) 415

A **Matriz de Confusão** é uma tabela que mostra as frequências de classificação de cada registro para cada classe do modelo e é a partir dela que iremos obter demais indicadores.

Métricas e Indicadores de Avaliação para Modelos de Classificação

True Positive (TP): indica a quantidade de registros que foram classificados como positivos corretamente, ou seja, a resposta do classificador foi que o evento era positivo e o evento realmente era positivo.

True negative (TN): indica a quantidade de registros que foram classificados como negativos de maneira correta, ou seja, a resposta do classificador foi que o evento era negativo e o evento realmente era negativo.

False positive (FP): indica a quantidade de registros que foram classificados como positivos de maneira incorreta, ou seja, a resposta do classificador foi que o evento era positivo, mas o evento era negativo.

False negative (FN): indica a quantidade de registros que foram classificados como negativos de maneira incorreta, ou seja, a resposta do classificador foi que o evento era negativo, mas o evento era positivo.

Através dessas quatro métricas pode-se calcular indicadores como a **Accuracy, Precision, Recall, F1-Score, Sensibility, Specificity**, etc.

Accuracy

O **Accuracy** é a relação entre as previsões positivas e negativas realizadas corretamente (*True Positives* e *True Negatives*) e todas as previsões (*True* e *False Positives* e *True* e *False Negatives*).

True Positive (TP) 150	False Negative (FN) 99	Accuracy 0,7804
False Positive (FP) 60	True Negative (TN) 415	

$$(TP + TN) / (TP + FP + TN + FN)$$

O **Accuracy** nos informa o quanto o modelo acerta das previsões possíveis, isto é, o quão frequente o classificador está correto.

Métricas e Indicadores de Avaliação para Modelos de Classificação

*Considerações sobre uma alta pontuação do **Accuracy***

Esse indicador pode comprometer uma correta avaliação de um modelo. Para essa pontuação realmente significar algo, é necessário que a base de dados utilizada esteja razoavelmente balanceada.

IMPORTANTE:

Suponha que temos um problema de negócio em que estamos querendo prever se um paciente possui uma doença ou não e que nossa base de dados está composta por 95% de registros em que a doença a ser classificada não ocorre e que apenas 5% ocorre. Obviamente que o modelo tenderá a aprender quando a doença não ocorre e, sendo assim, irá apresentar um valor de **Accuracy** de ~95%!

Pelo resultado, o modelo está informando que pode prever pessoas doentes em ~95% das vezes, mas, no entanto, está fazendo o oposto. Está prevendo pessoas que não estão doentes em ~95%!

Seria esse um indicador correto para o nosso problema de negócio?

Precision

O **Precision** é a relação entre as previsões positivas realizadas corretamente (*True Positives*) e todas as previsões positivas (*True Positives* e *False Positives*).

True Positive (TP) 150	False Negative (FN) 99	Accuracy 0,7804	Precision 0,7143
False Positive (FP) 60	True Negative (TN) 415		

$$TP / (TP + FP)$$

O **Precision** é a capacidade do modelo de não classificar como positivo um evento negativo. A principal utilização desse indicador é para modelos onde seja necessário minimizar os **Falsos Positivos**. Quanto mais perto de 1, melhor.

Recall

O **Recall** é a relação entre as previsões positivas realizadas corretamente (*True Positives*) e todos os eventos **que realmente** são positivos (*True Positives e False Negatives*).

<i>True Positive (TP)</i> 150	<i>False Negative (FN)</i> 99	<i>Accuracy</i> 0,7804	<i>Precision</i> 0,7143
<i>False Positive (FP)</i> 60	<i>True Negative (TN)</i> 415	<i>Recall</i> 0,6024	

$$TP / (TP + FN)$$

O **Recall** é a capacidade do modelo de não classificar como negativo um evento positivo, assim como informar o quanto se está identificando corretamente os eventos positivos.

A principal utilização desse indicador é para modelos onde seja necessário minimizar os **Falsos Negativos**, especialmente em medicina onde 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. Quanto mais perto de 1, melhor.

F1 Score

O **F1 Score** mede o balanço entre o **Precision** e o **Recall** e é obtido através do cálculo da média harmônica entre eles.

<i>True Positive (TP)</i>	<i>False Negative (FN)</i>	<i>Accuracy</i>	<i>Precision</i>
150	99	0,7804	0,7143
<i>False Positive (FP)</i>	<i>True Negative (TN)</i>	<i>Recall</i>	<i>F1 Score</i>
60	415	0,6024	0,6536

$$(2 * Precision * Recall) / (Precision + Recall)$$

O **F1 Score** pode ser uma excelente alternativa em relação ao uso do **Accuracy** para muitos problemas de negócio onde **falsos positivos** e **falsos negativos** possam ser relevantes. Mas atente-se ao fato de que também não temos total capacidade para uma correta interpretabilidade, pois não temos visão suficiente sobre o que está sendo maximizado, se o **Precision** ou o **Recall**. Portanto, o uso combinado com outras métricas e indicadores de avaliação são cruciais.

Métricas e Indicadores de Avaliação para Modelos de Classificação

Area Under the ROC Curve (ROC-AUC)

Seu cálculo utiliza como base o **Sensitivity** (*True Positive Rate*) e o **Spesificity** (*True Negative Rate*).

True Positive (TP) 150	False Negative (FN) 99	Accuracy 0,7804	Precision 0,7143	Sensibility 0,6024
False Positive (FP) 60	True Negative (TN) 415	Recall 0,6024	F1 Score 0,6536	Specificity 0,8737

$$\text{Sensitivity} = (TP) / (TP + FN) = \text{Recall}$$

True Positive (TP) 150	False Negative (FN) 99	Accuracy 0,7804	Precision 0,7143	Sensibility 0,6024
False Positive (FP) 60	True Negative (TN) 415	Recall 0,6024	F1 Score 0,6536	Specificity 0,8737

$$\text{Spesificity} = (TN) / (TN + FP)$$

Area Under the ROC Curve (ROC-AUC)

Geralmente, o cálculo da **ROC-AUC** é obtido com o auxílio de funções pré-programadas em plataformas de programação, mas quando estamos tratando de problemas de classificação binária e assumindo que o ponto de corte (*threshold*) está definido como 0.5 (um outro parâmetro para o cálculo da mesma), podemos obter seu resultado utilizando a seguinte fórmula (como iremos comprovar durante o desenvolvimento em classe):

$$\text{ROC-AUC} = (\text{Sensitivity} + \text{Specificity}) / 2$$

O **ROC-AUC** é uma ferramenta útil e poderosa para a avaliação de modelos de classificação, particularmente em domínios nos quais existe uma grande desproporção entre as classes (desbalanceamento).

Para um dado indivíduo doente e outro saudável, a **ROC-AUC** é uma medida que permite aferir qual a probabilidade do indivíduo doente obter um resultado **True Positive** e do indivíduo saudável obter um resultado **True Negative**. Quanto mais perto de 1, melhor.



ALBERT EINSTEIN
INSTITUTO ISRAELITA DE
ENSINO E PESQUISA

CENTRO DE EDUCAÇÃO EM SAÚDE
ABRAM SZAJMAN