Métricas

Métricas são importantes pois ajudam a avaliar a performance dos algoritmos de ML.

Avalia a performance das predições comparando com o resultado esperado.

Matriz de Confusão

- Quando o problema de classifcação tem como saída 2 ou mais tipos de classes. Para facilitar, iremos apresentar uma matrix de confusão com apenas 2 dimensoes (ou seja, problemas com apenas dois tipos de classes).
- Nesse matriz, temos na nas rows a esquerda os valores reais e na colunas na parte superior, temos os valores preditos.


```
                                                                    (predição) -> (correto)
VP (VERDADEIRO POSITIVO) = uma predicao positiva que é verdade        ->  positivo  ->  positivo
FN (FALSO NEGATIVO) = uma predicao negativa que é verdade             ->  !negativo ->  positivo
FP (FALSO POSITIVO) = uma predicao positiva que é negativa            ->  !positivo ->  negativo
VN (VERDADEIRO NEGATIVO) = uma predicao que é negativa que é negativa ->   negativo ->  negativo
```

Acurácia

- Nos diz quantos de nossos exemplos foram de fato classificados corretamente, independente da classe.
- Acurácia = (VP+VN)/(VP+VN+FP+FV)
- Na fórmula podemos ver que é a razão entre o que o modelo acertou e todos os outros tipos de classes
- Uma das **maiores desvantagens** é que em alguns problemas a acurácia pode ser elevada mas, ainda assim, o modelo pode ter uma performance inadequada. Por exemplo, considere o modelo que classifica exames de câncer entre positivo ou negativo para a doença, e em nosso conjunto de dados temos 1000 exemplos, sendo 990 de pacientes sem câncer e 10 de pacientes com câncer [“dataset desbalanceado”]. Caso nosso modelo seja ingênuo e sempre classifique todos os exemplos com negativo (sem câncer)[“Chuta tudo como sem câncer”], ele ainda obteria uma acurácia de 99%. O que parece uma excelente métrica, mas na verdade não estamos avaliando nosso modelo de forma adequada. Para melhor avaliar modelos que lidam com conjuntos de dados desbalanceados como este, outras métricas que serão apresentadas em seguida devem ser utilizadas.
- Outro ponto de atenção em relação à acurácia é que ela atribui o mesmo peso para ambos os erros. Por exemplo, considerando o mesmo exemplo de exame de câncer, suponha que o modelo tenha acertado 950 exemplos do total de 1000 exemplos. Os 50 exemplos errados podem ter sidos da classe positiva (falsos negativos) ou da classe negativa (falsos positivos). Em ambos os casos a acurácia seria de 95%, porém como mencionado anteriormente, o erro por falso negativo é bem mais grave neste problema, e isto não é refletido nesta métrica.

Precisão

- Podemos entender a precisão como sendo a expressão matemática para a pergunta: dos **exemplos classificados como positivos, quantos realmente são positivos**?
- Esta métrica é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos (todos elementos tem o positivo)
- Precisão = VP / (VP+FP)

- Olhando esta fórmula, podemos ver que a precisão dá um **ênfase maior para os erros por falso positivo**.
- Voltando ao exemplo do modelo de câncer, se o valor para a precisão fosse de 90%, isto indicaria que a cada 100 pacientes classificados como positivo, é esperado que apenas 90 tenham de fato a doença.
- Precisão significa a porcentagem de seus resultados que são relevantes (porcentegem acertos de classes positivas)

Recall ou sensibilidade

- A revocação busca responder a seguinte pergunta: **de todos os exemplos que são positivos, quantos foram classificados corretamente como positivos**?
- Ao contrário da precisão, a revocação [3], ou recall em inglês e também conhecida como sensibilidade ou taxa de verdadeiro positivo (TPR), **dá maior ênfase para os erros por falso negativo**. Esta métrica é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e a quantidade de exemplos que são de fato positivos.
- Recall = VP/ (VP + FN)
- Considerando o exemplo do modelo de câncer, se o valor para a revocação fosse de 95%, isto indicaria que a cada 100 pacientes que são de fato positivos, é esperado que apenas 95 sejam corretamente identificados como doentes.

F1 score

- Leva em consideração tanto a precisão quanto a recall. Ela é definida pela média harmônica entre as duas
- F1 = 2*(precisao * recall)/(precisao + recall)
- Uma das características da média harmônica é que se a precisão ou a revocação for zero ou muito próximos disso, o F1-score também será baixo. Desta forma, para que o F1-score seja alto, tanto a precisão como a revocação também devem ser altas. Ou seja, um modelo que apresenta um **bom F1-score é um modelo capaz tanto de acertar suas predições (precisão alta) quanto de recuperar os exemplos da classe de interesse (revocação alta)**. Portanto, esta métrica tende

a ser um resumo melhor da qualidade do modelo. Uma desvantagem é que a F1 acaba sendo **menos interpretável que a acurácia**.

- No exemplo do modelo de câncer, se o valor para a precisão fosse de 90% e o da revocação fosse de 95%, o valor para a F1 seria de 92.43%.

AUV (Area under ROC curve)

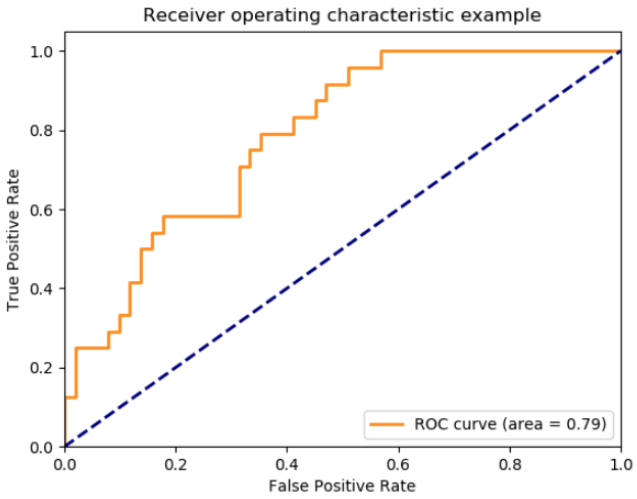
- Dado um modelo que atribui uma probabilidade para a classe positiva, é necessário definir um limiar de classificação. Acima deste limiar, um exemplo é classificado como positivo, caso contrário, é classificado como negativo. O limiar de classificação influencia o valor das métricas mencionadas anteriormente (acurácia, precisão, etc), e sua escolha deve levar em consideração o custo de cada erro.
- A curva ROC [3] (do inglês Receiver Operating Characteristic) pode ser utilizada para avaliar a performance de um classificador para diferentes limiares de classificação. Ela é construída medindo a Taxa de Falso Positivo (FPR — False Positive Rate) e a Taxa de Verdadeiro Positivo (TPR — True Positive Rate) para cada limiar de classificação possível, conforme as expressões abaixo.

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

- ROC é Receiver operating characteristic é um gŕafico para ilustrar a abilidade de um classificador binario em diagnosticar conforme seu limite de discrimicacao variavel.
- A curva ROC é então visualizada por meio de um gráfico, como mostra o exemplo abaixo na Figura 1. Ela mostra visualmente o compromisso entre falsos positivos e verdadeiro positivos na escolha do limiar. Quanto mais alto o limiar, maior é taxa de verdadeiro positivo (TP), porém a taxa de falso positivo (FP) também será maior. No caso extremo em que todos os exemplos são colocados na classe positiva, vemos que ambas as taxas chegam a 100%, enquanto no outro extremo, ambas ficam em 0%. Quanto mais próxima a curva estiver do canto superior esquerdo, melhor é a predição do modelo, dado que ele teria 100% de TPR e 0% de FPR. A

linha tracejada indica qual seria curva de uma classificador que prevê classes de forma aleatória, e serve como um baseline de comparação.

- A área sob a curva ROC (AUC — Area Under the Curve ou AUROC — Area Under the Receiver Operating Characteristic curve) pode ser utilizada como métrica de qualidade de um modelo, dado que quanto mais próxima a curva estiver do canto superior esquerdo, maior será a área sob a curva e melhor será o modelo. Uma vantagem desta métrica é que ela não é sensível ao desbalanço de classes, como ocorre com a acurácia. Por outro lado, a AUROC não é tão facilmente interpretável.



ROC é uma ótima maneira de visualizar o desempenho de um classificador binário e AUC é um único número para resumir o desempenho de um classificador avaliando a classificação em relação à separação das duas classes. Quanto mais alto, melhor.

Refinado by Lucas

Definir classes

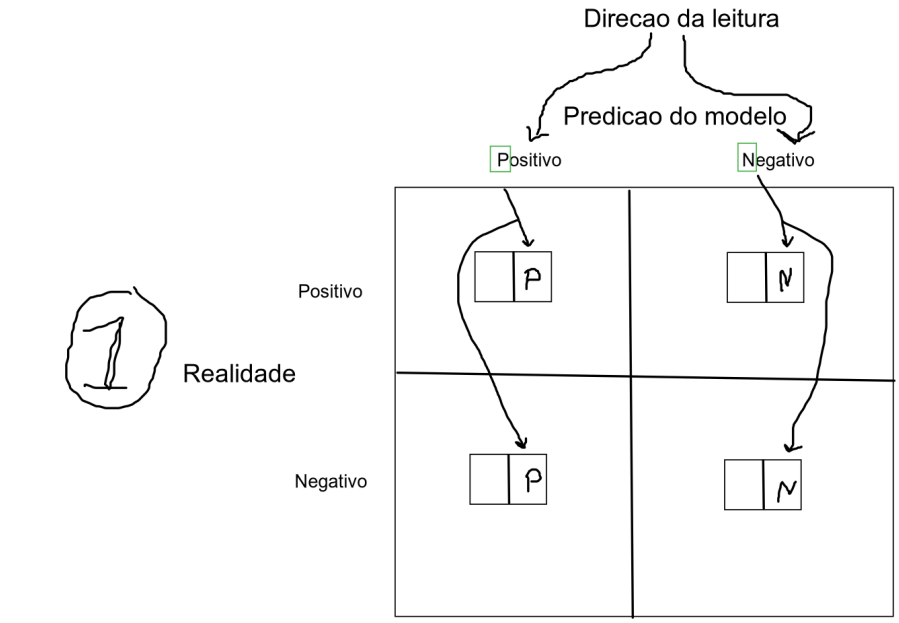
Primeiro passo é definir a classe positiva e a classe negativa do problema.
Exemplo:

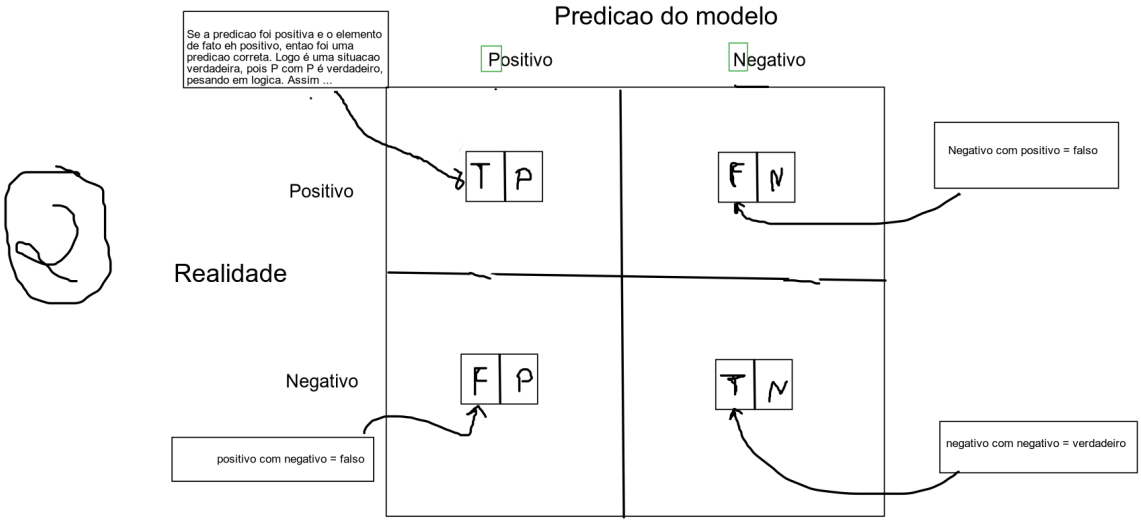
- Classe positiva = Email é span;

- Classe negativa = Email não é span;
  --
- Classe positiva = Estado não é anomalo;
- Classe negativa = Estado é anomalo;

Obs.: depois de definido esse mapeamento, nada impede de ser realizado uma inversão para serem realizados novos testes.

matriz de confusao





Metricas

Metricas ajudam a valiar o desempenho do modelo

Acuracia

Taxa de acertos de madeira geral, ou seja, acertos de classe positiva (TP) e acertos de classe negativa (TN).

formula-> (#TP+#TN/#total de dados)

#total dados = TP+TN+FP+FN

Exemplo:

Precisao

Dos valores preditos como positivo, quantos realmente sao positivos.

formula-> (#TP/#TP+#FP)

Alto valor de precisao significa fizer baixo valor de FP, i.e baixo erro do Tipo I

Exemplo:

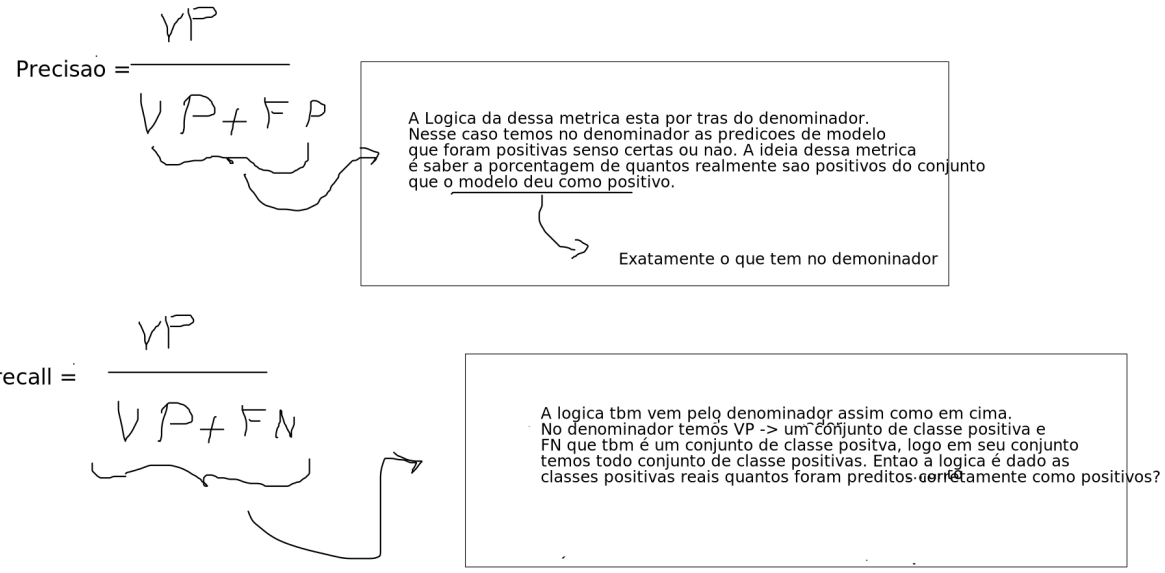
Recall ou Recuperar

De todos os valores positivos reais, quantos foram preditos corratamente positivos.

formula-> (#TP/#TP+#FN)

Alto valor de recuperacao significa fizer baixo valor de FN, i.e baixo erro do Tipo II

Exemplo:



F1 score

Um modelo que apresenta um bom F1-score é um modelo capaz tanto de acertar suas predições (precisão alta) quanto de recuperar os exemplos da classe de interesse (revocação/recall alta).

Alto valor no F1, podemos dizer que temos baixos valores para erros do Tipo I e II. Exemplo:

OBS.:
A precisão e a recall (e por extensão, a F1 score, que é uma função das duas) consideram uma classe, a classe positiva, como a classe em que estamos interessados. Eles usam apenas três dos valores na matriz de confusão: TP, FP e FN. O quarto valor - TN - não é usado nessas métricas. Você pode colocar qualquer valor na célula TN - 0, 100, infinito - e a precisão, a recuperação e a pontuação F1 não serão alteradas.

Portanto elas estão preocupadas em saber o quanto o modelo "entende" o que é a classe positiva e nem liga se o modelo entende o que é classe negativa.

Exemplo: supondo que a classe positiva seja cachorro e a classe negativa seja gato. Se o modelor der bons no Precisao, recall e F1 score, nao significa que ele sabe bem o que é um gato, pois eles nao usam o calculo o TF ou seja o acerto de ser gato.

Curva ROC (Receiver Operating Characteristic) e AUC (Area Under the ROC Curve)

A curva ROC mostra o quao bom o modelo pode distinguir entre a classe positiva e negativa. Logo para isso, a gente precisa saber a taxa que ele diz corretamente o que é uma classe positiva (VP) e a taxa que o que ele diz erradamente ó que a classe positiva (FP). Dado isso, podemos verificar se ele consegue distinguir bem o que é a classe positiva.

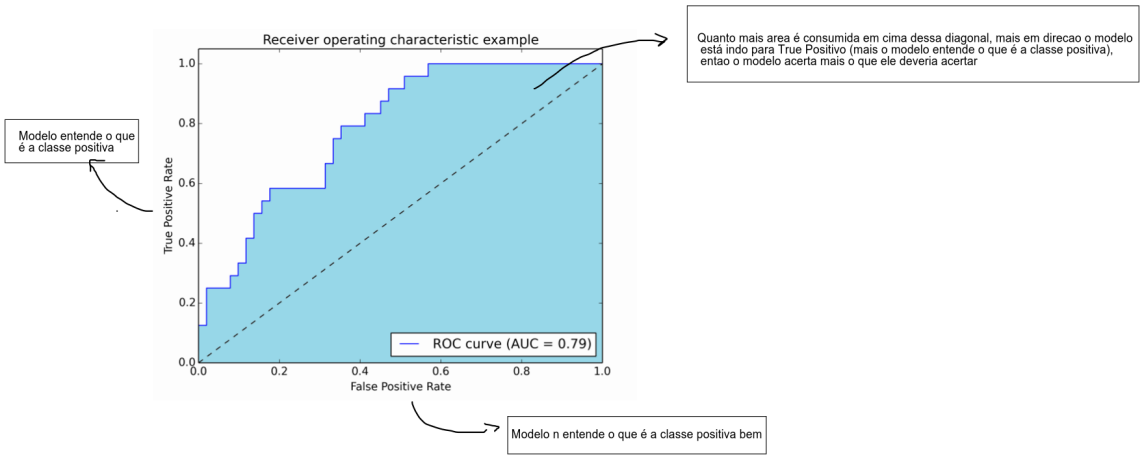
Logo os parametros da curva ROC são a taxa de VP e FP.

Assim a curva ROC geralmente é representada por um gráfico que tem um limite entre Taxa de verdadeiro positivo e Taxa de falso positivo.

Na tentativa de simplificar a análise da ROC, a AUC (“area under the ROC curve”) nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da ROC, calculando a “área sob a curva”.

O valor do AUC varia de 0,0 até 1,0 e o limiar entre a classe é 0,5. Ou seja, acima desse limite, o algoritmo classifica em uma classe e abaixo na outra classe. **Quanto maior a AUX, melhor.**

É um ótimo comparador de modelos.



Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1.

O interessante do AUC é que a métrica é invariante em escala, uma vez que trabalha com precisão das classificações ao invés de seus valores absolutos. Além disso, também mede a qualidade das previsões do modelo, independentemente do limiar de classificação.

Matthews Correlation Coefficient (MCC)

É especifica para classificação binária
Consegue me dá a melhor configuração para obter os melhors precision and recall;

Quando o classificador é perfeito (FP = FN = 0, ou seja, não existe erro de predição) o valor de MCC é 1.

Caso contrário, quando o classificador sempre erra (TP = TN = 0), é obtido um valor de -1. Nesta última situação representa uma correlação negativa perfeita (neste caso, você pode simplesmente reverter o resultado do classificador para obter o classificador ideal), ou seja, se o classificador desse que é cachorro, vc troca para gato e vice versa.

Com MCC 0 significando que o classificador não é melhor do que um lançamento aleatório de uma moeda justa.

MCC também é perfeitamente simétrico, então nenhuma classe é mais importante do que a outra.

Se você trocar a classe positiva e o negativa, ainda obterá o mesmo valor.

Uma correlação, por exemplo, de 0,17 significa que a classe prevista e a classe verdadeira estão fracamente correlacionadas.

Material de apoio:
https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a