Pontifícia Universidade Católica do Rio de Janeiro

Departamento de Engenharia Elétrica – Inteligência Computacional Aplicada

# Documento auxiliar - 1º Trabalho de Redes Neurais (Classificação)

## 1. Sobre o documento

Na área de ciência de dados, habitualmente nos deparamos com a necessidade de construir um modelo para auxiliar nas decisões de atividades. Esses modelos podem ter finalidades diferentes, como classificação de padrões ou detecção de fraudes. Contudo, qualquer modelo necessita de uma análise mais profunda sobre a capacidade de generalização, e podemos avaliá-lo através das **métricas de avaliação**.

Neste documento, iremos discutir um pouco sobre algumas métricas de avaliação essenciais para uma boa análise de resultados em problemas de classificação. Em cada métrica, será discutido como se calcula, como podemos extrair informações importantes e avaliar o modelo.

## 2. Métricas de avaliação

O projeto de Machine Learning envolve basicamente três grandes etapas: pré-processamento, desenvolvimento do modelo de Machine Learning e pós-processamento. Mas como podemos entender se alguma técnica de pré-processamento que utilizamos, ou o ajuste de hiperparâmetros do modelo, melhorou de fato o resultado do modelo? Neste caso, devemos analisar os resultados usando métodos de avaliação que auxiliam nesta tarefa.

Não existe uma métrica de avaliação mágica que consiga expressar todas as informações para a análise do resultado. Ainda assim, algumas métricas são bem úteis e, combinadas, podem auxiliar em uma avaliação mais completa do modelo. Além disso, nem todas as métricas são aplicáveis ou úteis a qualquer tipo de problema. Em um problema de regressão (no qual a saída do modelo é um valor numérico), não é possível realizar o cálculo de acurácia.

Nesta seção, iremos destacar algumas técnicas úteis para auxiliar na análise de resultados em problemas de **classificação**. O primeiro a ser abordado é sobre a Matriz de Confusão, visto que é essencial para a definição das outras métricas de avaliação citadas neste documento. Em seguida, explicaremos como se calcula a Acurácia, Precision, Recall e F1-Score, e que tipo de informação podemos extrair destes resultados.

## 2.1. Matriz de Confusão

A Matriz de Confusão é um método essencial para a análise de resultados em um problema de classificação, pois é a partir dele que outras métricas, como Precision e Recall, podem ser calculados. A Figura 1 mostra como é a matriz de confusão para duas classes, nominalmente chamado de Positivo (P) e Negativo (N).

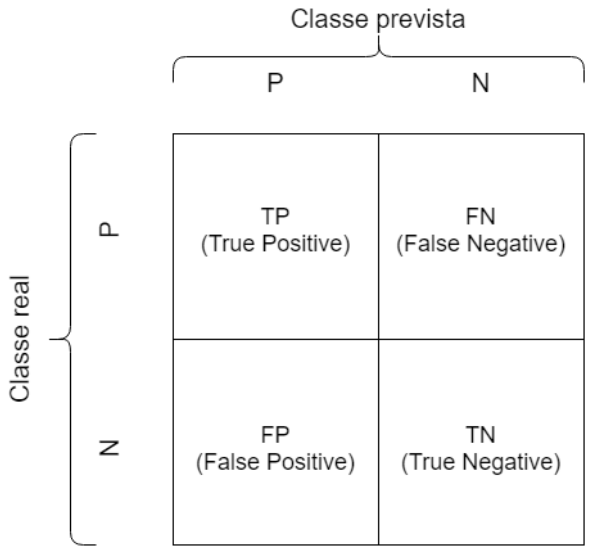


Figura 1 – Estrutura da matriz de confusão

Uma das grandes vantagens da matriz de confusão é a visualização do desempenho do modelo de uma forma mais detalhada. Além de mostrar as classificações corretas, também é possível observar como é ocorrem os erros de classificação.

Definimos como True Positive (TP) as classificações corretas da classe P, e False Positive (FP) as classificações que foram previstas como P, mas o rótulo verdadeiro é N. Analogamente, True Negative (TN) é definido como a classificação correta da classe N, enquanto False Negative (FN) é a classificação prevista como N, mas com classe real P.

As definições de classificação acima serão importantes para o cálculo das seguintes métricas: Acurácia, Precision, Recall e F1-Score. Vale ressaltar que, neste documento, todas as métricas foram definidas e calculadas em relação a classe Positiva (P). Caso exista um problema de classificação no qual seja interessante avaliar o Precision, Recall e F1-Score de mais de uma classe, podemos calcular estas métricas separadamente, transformando a classe de interesse como P.

Para exemplificar e motivar as métricas abaixo, usaremos como base este exemplo:

“Estamos interessados em descobrir se uma compra feita por uma pessoa no cartão de crédito é suspeita (i.e. pode ter sofrido golpe ou o cartão foi clonado). Para isso, foi criada uma base de dados contendo informações relevantes, como o valor da compra a ser analisada, o valor médio gasto no mês passado e o rótulo, indicando se aquela transação foi considerada suspeita ou não. Para este trabalho, treinamos um modelo de

Rede Neural MultiLayer Perceptron e observamos o resultado no conjunto de teste, como mostra a Figura 2.”

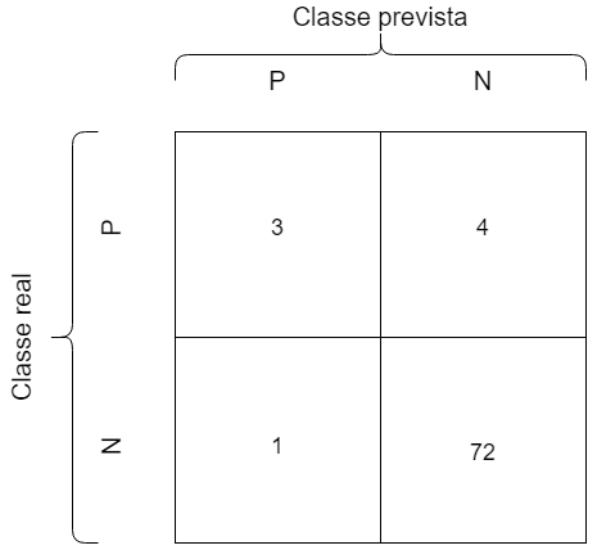


Figura 2 – Matriz de confusão do Exemplo

## 2.2. Acurácia

A acurácia é uma das métricas mais simples e, geralmente, é a primeiro resultado de interesse para avaliação do modelo. Esta métrica define quão bom o modelo é em classificar corretamente o problema. Sua fórmula geral é:

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$$

Onde P e N são todos os padrões referentes a classe P e N, respectivamente. É intuitivo pensar que a acurácia é um bom indicador para analisar um modelo. Esperamos sempre que o modelo consiga classificar corretamente o máximo de registros possíveis, e a acurácia revela exatamente isto: quantos padrões acertamos em relação a todos estes apresentados.

Embora seja muito utilizada, a acurácia também possui limitações, e por vezes pode confundir a análise do resultado. Em problemas envolvendo classes desbalanceadas, como o exemplo citado nesta seção, a métrica de acurácia pode ser alta (no caso, 92,5%), mas ainda sim observamos que o modelo não classifica corretamente as compras suspeitas. Precisamos, portanto, de mais informações para avaliar se o modelo treinado é adequado para este tipo de problema.

## 2.3. Precision

O Precision é uma métrica que indica a qualidade de previsão de classificação positiva de um problema, tendo a seguinte equação:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Em outras palavras, o Precision mostra, dentre todos os padrões que são rotulados como positivo, quais foram corretamente classificados. Deste modo, esta informação está relacionada com a taxa de falsos positivos do modelo. No nosso exemplo, a precision é de 75%.

Dependendo do problema, isso pode ser extremamente relevante para a análise. Um exemplo clássico é o de classificação de e-mails contendo spam. Neste problema, é desejável que o modelo desenvolvido consiga identificar os e-mails que são spam (TP). Contudo, classificar um e-mail genuíno como spam (i.e. Falso Positivo) pode ser prejudicial, e portanto uma precision alta é importante nesta tarefa.

## 2.4. Recall

Recall (ou sensibilidade) é uma métrica que indica a qualidade de classificação do modelo em relação aos padrões que possuem rótulos reais positivos. Em outras palavras, o Recall expressa a taxa de classificação correta do modelo para a classe atual P

Esta métrica pode ser calculada pela equação:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Assim como o Precision, a Sensibilidade é bastante útil para analisar quando estamos interessados em avaliar o desempenho do modelo quando o erro de um falso negativo pode ser muito prejudicial. Como exemplo, podemos citar casos em que queremos classificar quais pacientes estão doentes. Uma classificação de um paciente que é doente (i.e. tem rótulo Positivo) e é classificado como não doente (i.e. um Falso Negativo), este erro pode ser muito mais crítico do que classificar um paciente que não está doente (i.e. tem rótulo Negativo) como doente (i.e. classificar como um Falso Positivo). Portanto, avaliar este problema usando o Recall é essencial para selecionar um melhor modelo.

No nosso exemplo, o Recall é aproximadamente 42,9%. Este valor indica que, em casos que padrão pertencia a classe P, apenas em 42,9% dos casos o modelo foi capaz de classificar corretamente. Em termos práticos, onde essa classificação errônea pode ser custoso (como avaliar se um cliente pode ser fraudador), um valor baixo de Recall pode indicar que o modelo não obteve um bom desempenho, por mais que sua acurácia seja alta.

## 2.5. F1-Score

A métrica F1-Score utiliza informações de Precision e Recall para o cálculo da métrica, segundo a seguinte fórmula:

$$\text{F1} - \text{Score} = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Esta métrica informa a média harmônica entre Precision e Recall, onde o maior valor é 1 e o menor é 0 (quando pelo menos uma delas é igual a zero). Por conta disso, o F1-Score é uma boa opção quando precisamos também considerar estas duas outras métricas.

Além disso, o F1-Score é mais sensível em relação a dados desbalanceados. No Exemplo 1, temos:

$$\text{F1-Score} = 0{,}623$$

Embora a sua interpretação não seja tão simples quanto a acurácia, ela pode ser muito útil. Em casos onde é importante analisar os efeitos de falsos negativos e falsos positivos, ou quando a distribuição de padrões por classe é desbalanceada, o F1-Score é mais apropriado para ser utilizado.

No nosso exemplo, é possível ver que outras métricas (como o Recall, Precision e F1-Score) conseguem nos informar mais sobre o resultado obtido do que apenas a acurácia. Portanto, quando precisamos analisar o desempenho do modelo em um problema, devemos nos basear em métricas que sejam úteis e consigam expressar informações mais relevantes.