**TDE 02 – Comparação de Algoritmos de Detecção de Anomalias**

1. Problema

As anomalias podem indicar problemas de desempenho em sistemas computacionais. O algoritmo para detecção de anomalias pode ser construído a partir de dados usando uma técnica de classificação. Na atividade FORMATIVA 05 (semana 05), utilizamos um Classificador Bayesiano para identificar situações de carga anormal em servidores. Na atividade FORMATIVA 07 (Semana 07), utilizamos a distribuição normal multivariada para identificar situações de anormalidade em um sistema computacional usando a distribuição normal multivariada. O problema que vamos abordar nesse TDE é como comparar o desempenho dos dois algoritmos.

1. Métricas de avaliação de classificação

Na atividade “FORMATIVA 07” utilizamos a métrica F1 para ajustar o algoritmo de detecção de outliers baseado na distribuição normal multivariada. A métrica F1 é definida como a média harmônica entre as métricas precision e recall. As métricas accuracy, precision, recall e F1 são as mais utilizadas para comparar a eficiência de algoritmos de classificação.

Nessa atividade as métricas foram calculadas com relação aos valores verdadeiros, ou seja, as anomalias, da seguinte forma:

* Exatidão (accuracy): proporção dos verdadeiros entre todos examinados
* Precisão: proporção dos verdadeiros positivos entre todos os classificados positivos
* Recall: proporção dos verdadeiros positivos entre todos os realmente positivos
* F1: média harmônica entre precisão e recall

A métrica accuracy é sempre calculada em relação aos valores verdadeiros, mas as demais métricas podem ser calculadas com relação aos valores falsos, ou seja, com relação aos valores falsos:

* Precisão: proporção dos verdadeiros negativos entre todos os classificados negativos
* Recall: proporção dos verdadeiros negativos entre todos os realmente negativos
* F1: média harmônica entre precisão e recall

A partir das métricas para valores positivos e negativos podemos calcular as mesmas métricas como uma média ponderada dos valores encontrados para valores positivos e valores negativos, usando como valor a quantidade real de positivos e a quantidade real negativos. Vamos usar aqui a nomenclatura do sklearn (biblioteca usada na FORMATIVA 05):

* precision\_0, recall\_0 e F1\_0 as métricas calculadas em relação aos valores falsos
* support\_0 a quantidade real de negativos
* precision\_1, recall\_1 e F1\_1 as métricas calculadas em relação aos valores verdadeiros
* support\_0 a quantidade real de verdadeiros

Temos então as seguintes equações para precision, recall e F1:



Temos as seguintes equações para calcular as métricas ponderadas (mp):

* *support\_0 = fp + vn*
* *support\_1 = vp + fn*
* *prec\_mp = (prec\_0 \* support\_0 + prec\_1 \* support\_1) / (support\_0 + support\_1)*
* *rec\_mp = (rec\_0 \* support\_0 + rec\_1 \* support\_1) / (support\_0 + support\_1)*
* *F1\_mp = (F1\_0 \* support\_0 + F1\_1 \* support\_1) / (support\_0 + support\_1)*

DICA

Utilizar o código fornecido na FORMATIVA 07 para métricas ponderadas.

Código numpy para calcular as métricas ponderadas (mp):

# Accuracy

acc = (vp+vn)/(vp+vn+fp+fn)

# prec rec F1 para outliers

prec\_1 = vp/(vp+fp)

rec\_1 = vp/(vp+fn)

F1\_1 = (2\*prec\_1\*rec\_1) / (prec\_1+rec\_1)

# prec rec F1 para não outliers

prec\_0 = vn/(vn+fn)

rec\_0 = vn/(vn+fp)

F1\_0 = (2\*prec\_0\*rec\_0) / (prec\_0+rec\_0)

# prec rec F1 média ponderada

support\_0 = fp + vn

support\_1 = vp + fn

prec\_mp = (prec\_0 \* support\_0 + prec\_1 \* support\_1) / (support\_0 + support\_1)

rec\_mp = (rec\_0 \* support\_0 + rec\_1 \* support\_1) / (support\_0 + support\_1)

F1\_mp = (F1\_0 \* support\_0 + F1\_1 \* support\_1) / (support\_0 + support\_1)

1. Detecção de outliers com classificador Bayesiano

No arquivo classificadorNB\_Anomalias temos o código para implementação da classificação com algoritmo Bayesiano para os mesmos dados usados na FORMATIVA 07 (algoritmo com distribuição normal multivariada). O executar o notebook você deve obter os seguintes valores para as métricas ponderadas:

precision\_mp = 0.91

recall\_mp = 0.91

1. Comparação de algoritmos

Para comparar os dois algoritmos vamos observar as métricas accuracy e F1\_mp.

Para comparar será necessário calcular as métricas na Parte 3 de “DetecçãoOutliers.ipynb” da Formativa 07. A métrica accuracy já está calculada. Será necessário calcular a métrica F1\_mp.

5.1 Para realizar a comparação preencha a seguinte tabela:

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1\_mp |
| Classificação Bayesiana |  |  |
| Normal multivariada |  |  |

5.2 Faça aqui suas considerações sobre o desempenho dos algoritmos. Escreva um parágrafo sobre qual seria mais eficiente. Quando seria melhor utilizar um ou outro algoritmo. Não existe uma resposta correta. Procure entender as vantagens e limitações de cada algoritmo e escrever o que realmente achou. Um ponto a considerar é a distribuição das variáveis, já que o algoritmo que usa a Gaussiana multivariada funciona bem se as variáveis tiverem distribuição normal.

**Entrega**

Responder 5.1 e 5.2.

Salvar esse arquivo como pdf.

Enviar no AVA.