TDE 02 – Comparação de Algoritmos de Detecção de Anomalias

Equipe: Eduardo Eiji Goto, Gustavo Hammerschmidt, João Vitor Andrioli de Souza.

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 "Problemas em Equipe 05" da semana 05, utilizamos um Classificador Bayesiano para identificar situações de carga anormal em servidores. Na atividade "Problemas em Equipe 06" da 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.

2) Métricas de avaliação de classificação

Na atividade "Problemas em Equipe 06" 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 utilizados as métricas 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 em "Problemas em Equipe 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:

- $prec\_0 = \frac{vn}{vn+fn}$ $\quad rec\_0 = \frac{vn}{vn+fp}$ $\quad F_{1\_0} = \frac{2 \cdot pre\_0 \, rec\_0}{prec\_0 + rec\_0}$

- $prec\_1 = \frac{vn}{vp+fp}$ $\quad rec\_1 = \frac{vp}{vp+fn}$ $\quad F_{1\_1} = \frac{2 \cdot pre\_1 \, rec\_1}{prec\_1 + rec\_1}$

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)*

3) 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 no "Problemas em Equipe 06" da semana 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

F1_mp = 0.88

4) Comparação de algoritmos

Para comparar os dois algoritmos vamos observar as métricas accuracy e F1_mp.
Para comparar será necessário calcular as métricas em "Problemas em Equipe 06" da semana 07. A métrica accuracy já está calculada. Será necessário calcular a métrica F1_mp. Utilize o código disponível no notebook do TDE02 para calcular a métrica F1_mp para o algoritmo baseado na distribuição normal multivariada desenvolvido em "Problemas em Equipe 05" da semana 05.

Para realizar a comparação preencha a seguinte tabela:

	Accuracy	F1_mp
Classificação Bayesiana	0.9775	0.9784
Normal multivariada	0.892	0.891612

Faça aqui suas considerações sobre o desempenho dos algoritmos. Escreva um parágrafo sobe 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.

O trabalho pode ser feito em grupos de até 5 estudantes.

Considerações:

Segundo a biblioteca sklearn, a métrica F1_mp avalia a média ponderada pelo suporte e isso pode resultar numa métrica F1 que não está entre a métrica precision ou recall. Nos modelos testados, o modelo de classificação bayesiana (semana 07) aponta um melhor accuracy, logo

seria o algoritmo mais eficiente em responder o problema corretamente na maioria das vezes; o seu F1_mp ficou próximo de seu accuracy. Já o modelo da normal multivariada (semana 05) teve 0.892 em accuracy e seu F1_mp ficou próximo também. Porém, na ótica do accuracy, deve-se levar em conta que a métrica F1 constata o desempenho do modelo de uma forma mais harmônica. Neste cenário, contudo, tivemos que o modelo de classificação bayesiana obteve uma melhor métrica F1_mp do que o modelo normal variada. Portanto, para esse problema, o modelo de classificação bayesiana é melhor.