

MODELAGEM HÍBRIDA GENERATIVO-DISCRIMINATIVA PARA CLASSIFICAÇÃO BINÁRIA EM SAÚDE PÚBLICA

Ian Lucas Mendes Bezerra

Universidade Estadual de Campinas
R. Pedro Zaccaria, 1300, Limeira - SP, 13484-350
i291220@dac.unicamp.br

Cristiano Torezzan

Universidade Estadual de Campinas
R. Pedro Zaccaria, 1300, Limeira - SP, 13484-350
torezzan@unicamp.br

RESUMO

Este trabalho propõe e avalia um classificador binário híbrido para dados de saúde pública com atributos binários, como sinais e sintomas clínicos. Problemas de classificação com atributos exclusivamente binários apresentam desafios específicos, como limitação na expressividade individual das variáveis e maior sensibilidade a dependências entre atributos. A abordagem proposta combina o modelo generativo Bernoulli Naive Bayes, que assume uma estrutura causal simples entre diagnóstico e sintomas, com a regressão logística, modelo discriminativo amplamente utilizado para predição. A predição final é construída como uma combinação ponderada das saídas dos dois modelos, com pesos ajustados por meio da minimização da perda logística, resultando em um problema de otimização de segunda ordem. A motivação metodológica baseia-se na complementaridade entre a robustez estatística dos modelos generativos e o poder de discriminação dos modelos discriminativos. O modelo é aplicado a um conjunto de dados reais do DATASUS, com foco na classificação de Síndrome Respiratória Aguda Grave (SRAG). Os resultados são avaliados quanto à acurácia, calibragem probabilística e robustez em diferentes subconjuntos da população. A proposta destaca-se por incorporar uma leitura estrutural dos dados, aproximando modelagem estatística e aprendizado causal em contextos de saúde pública.

PALAVRAS CHAVE. Classificação binária, Modelos híbridos generativo-discriminativos, Saúde pública.

Tópicos (EST&AM – PO Analytics em Estatística e Aprendizado de Máquina; SA – PO na Área de Saúde; IA – PO e IA)

ABSTRACT

This work proposes and evaluates a hybrid binary classifier for public health data with binary attributes, such as clinical signs and symptoms. Classification problems involving exclusively binary features present specific challenges, such as limited individual expressiveness and greater sensitivity to attribute dependencies. The proposed approach combines the generative Bernoulli Naive Bayes model—which assumes a simple causal structure from diagnosis to symptoms—with logistic regression, a widely used discriminative model for prediction tasks. The final prediction is obtained through a weighted combination of the outputs of both models, with weights learned

by minimizing the logistic loss function, leading to a second-order optimization problem. The methodological motivation lies in leveraging the statistical robustness of generative models and the predictive power of discriminative models. The proposed model is applied to a real dataset from DATASUS, focusing on the classification of Severe Acute Respiratory Syndrome (SARS). Results are evaluated in terms of accuracy, probabilistic calibration, and robustness across different population subgroups. The proposal stands out by incorporating a structural interpretation of the data, bridging statistical modeling and causal learning in public health contexts.

KEYWORDS. Binary classification, Hybrid models, Public health.

Paper topics: EST & AM, SA, IA