

Classificação de Celulares por Faixa de Preço

Metodologia

Com o uso das bibliotecas *sci-kit learn* e *Tensorflow*, foi construído e avaliado um modelo para classificar o dataset do problema por faixas de preço. O trabalho foi desenvolvido usando na ferramenta Google Colaboratory.

Após a importação e leitura dos dados, a primeira etapa consiste na codificação *one-hot* das classes.

Em seguida, a biblioteca a amostragem é realizada com o auxílio da biblioteca *SMOTE*, realizando *over-sampling* na base de dados de modo a evitar problemas de desequilíbrio entre as classes. Com isto é gerada uma base de treinamento e uma de teste.

O modelo é construído de forma sequencial, com apenas uma camada escondida. A camada de saída utiliza a função de ativação *softmax*, o que justifica a codificação *one-hot* das classes.

Os hiperparâmetros avaliados para ajuste foram:

- número de neurônios na camada escondida (200, 100)
- função de ativação dos neurônios da camada escondida (*ReLU*, *tanh*)
- otimizador (*SGD*, *Adam*)
- tamanho de lote: número de registros apresentados a cada atualização dos pesos (256, 128, 64)

Para este ajuste, o modelo foi treinado e testado com cada combinação dos valores apresentados para estes hiperparâmetros. Os otimizadores foram empregados com suas configurações padrão.

Vale notar que a quantidade de épocas de treinamento não foi incluída nesta avaliação uma vez que a parada antecipada foi configurada de modo a selecionar a melhor quantidade dentre 50 e 500 épocas.

Para o teste, foram gerados gráficos de perda e acurácia ao longo do treinamento e validação, além da matriz de confusão e do relatório contendo múltiplas métricas como acurácia, *score F1*, *recall* e precisão. A métrica selecionada para a escolha do melhor modelo foi a acurácia. Após cada teste o modelo é gravado em disco caso obtenha a maior acurácia até então, de modo que a obtenção do modelo vencedor seja possível ao final.

Resultado

O melhor modelo avaliado foi o que continha 200 neurônios na camada escondida com a função de ativação *tanh*, otimizador *Adam* e tamanho de lote igual a 256. Este modelo alcançou uma acurácia de 94%.