Técnicas Clássicas de Reconhecimento de Padrões (2020/01)

Proposta de Artigo 02

Explorando Redes Profundas Utilizando o Método SHAP

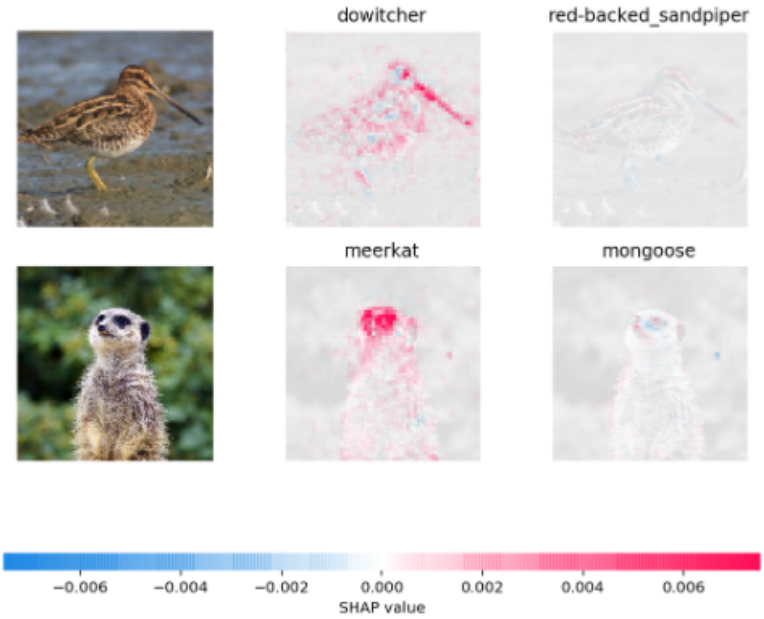
Aluno: Ramon Gomes Durães de Oliveira (2019720188)

Introdução

Nos últimos anos tem se tornado muito comum que modelos de aprendizado de máquina complexos sejam o estado-da-arte para diversas tarefas. Um exemplo são as redes convolucionais profundas, que são notórias pelo seu estilo "caixa-preta" de resolução de problemas. Estas técnicas de aprendizado "end-to-end" têm um forte ponto negativo: a falta de interpretabilidade. Uma vez que a predição da rede é obtida para uma nova observação, não se sabe o porquê da resposta obtida pelo modelo. Surge assim a necessidade de estudar técnicas para melhorar a explicabilidade dos resultados gerados por esses modelos e aumentar sua interpretabilidade.

Uma técnica que tem sido explorada nesse contexto é o "Shapley value", introduzido por Lloyd Shapley em 1951 [1, 2]. Resumidamente, dado um grupo de jogadores que se unem para desenvolver um trabalho e como consequência obtêm um ganho final geral, os Shapley values são uma forma de definir quantitativamente quão importante cada jogador foi para o resultado final. Num contexto de aprendizado de máquina, esses valores são utilizados para explicar o quão relevante cada feature do modelo foi para o resultado final. Esta técnica vem sendo explorada há alguns anos para classificadores em geral [4], e também em contextos mais específicos como árvores de decisão [3] e redes profundas [5, 6].

O método SHAP (SHapley Additive exPlanations), proposto em [4], se destaca em relação aos outros pois aumenta a explicabilidade dos modelos garantindo uma série de propriedades de forma matematicamente comprovada que outros métodos não garantem. Uma implementação em Python deste método está disponível no GitHub (https://github.com/slundberg/shap (https://github.com/slundberg/shap)). Na figura abaixo é mostrado um exemplo dos resultados obtidos no contexto de classificação animais. Utilizando um conjunto de dados e o modelo já treinado, os valores SHAP obtidos são altos (cor vermelha se positivos, azul se negativos) exatamente para as regiões da entrada que discriminam bem as classes.



Objetivos

O método SHAP será explorado neste trabalho para propor uma metodologia que combine os "SHAP values" e os graus de ativação de redes profundas para extrair mais informações a respeito destes modelos, extrapolando sua tarefa original de apenas classificar novas observações. Serão também avaliadas alterações no método SHAP que gerem resultados promissores, como a utilização de diferentes Kernels. A princípio, deseja-se explorar essas redes no contexto de imagens médicas tanto pela relevância de modelos interpretáveis nesta área quanto pela redução na complexidade computacional já que tipicamente estas imagens estão em tons de cinza.

Referências

[1] Shapley, Lloyd S. (August 21, 1951). "Notes on the n-Person Game -- II: The Value of an n-Person Game" (PDF). Santa Monica, Calif.: RAND Corporation.

[2] Roth, Alvin E., ed. (1988). The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511528446. ISBN 0-521-36177-X.

[3] LIME: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

[4] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.

[5] Layer-wise relevance propagation: Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.

[6] DeepLIFT: Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." arXiv preprint arXiv:1704.02685 (2017).