Adrian Marinovich
Springboard Data Science Career Track
Capstone Project #1 - Exploratory data analysis
October 12, 2018

Project: Detection of smiles in images of faces

Are there variables that are particularly significant in terms of explaining the answer to your project question?

As seen in the Jupyter notebook referenced below, the exploratory data analysis is tailored to review pixel and image characteristics that are important in the performance of an early stage machine learning approach using a random forest classifier. Please see the Jupyter notebook for the evaluation of pixel features importances.

Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

As seen in the contingency table in the Jupyter notebook, the random forest approach and analysis of feature importances by pixel, coupled with plotting the test faces on a contingency table, suggests some higher-level features that may have a role in random forest misclassification:

- False negatives may in part result from face rotation, scaling and centering/cropping differences.
- False positives may in part result from prominent nasolabial folds (as might be seen in smirking or grimacing), presence of facial hair, as well as face rotation, scaling and centering/cropping differences.

What are the most appropriate tests to use to analyse these relationships?

In the next machine learning steps involving support vector machines and neural networks, the important pixel regions may change, and with that the types of misclassification seen. This overall approach, however, will allow ongoing monitoring and adjustment of the models, as well as guide possible transformations and filtering of the image data to improve smile classification.

Submit a link to the document.

https://github.com/adriatic13/springboard/blob/master/dsct_capstone1/Adrian_Marinovich_Cap1_smiles_eda.ipynb