Exercise 4a

Deadline: 12.01.2021, 16:00.
Ask questions to #ask-your-tutor-(ulrich or yulia)

Regulations

Provide your comments for last week's homework in the files

- `tree-methods-commented.ipynb` and `tree-methods-commented.html` for feedback to your solution,
- `tree-methods-cross-commented.ipynb` and `tree-methods-cross-commented.html` for cross-feedback.

Hand-in your solutions to this week's tasks in notebooks `precision-recall.ipynb/precision-recall.html` and `red-cards.ipynb/red-cards.html`. Zip all files into a single archive with naming convention (sorted alphabetically by first names)

`firstname1-lastname1_firstname2-lastname2_ex04a.zip`

or (if you work in a team of three)

`firstname1-lastname1_firstname2-lastname2_firstname3-lastname3_ex04a.zip`

and upload this file to Moodle before the given deadline.

Recall that we will give zero points if your zip-file does not conform to this naming convention.

1 Comment on your and other's solution to Exercise 4

Similar to the exercises before, comment on your own and another group's solution. You will receive the latter via mail after the deadline of the previous sheet.

2 Red Cards Study

In this exercise, you will take a look at a recent experiment in crowdsourcing research. 29 teams of researchers were given the same dataset and the same question: *"Are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?"*. Interestingly, all 29 teams arrived at different conclusions (finding no bias or slight bias or severe bias in referee decisions), despite having identical data and instructions. Read `http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-1.18508` for a comment in the Nature journal by Raphael Silberzahn & Eric L. Uhlmann, the initiators of the experiment.

We ask you the same question: Given the dataset, can you confirm or refute the question? To do this, please download the dataset from `https://osf.io/gvm2z/` (`1. Crowdsourcing Dataset July 01, 2014 Incl.Ref Country.zip` contains the dataset, `README.txt` a detailed description of the data and `2._Crowdstorming_Pictures_Skin_Color_Ratings.zip` the images of the players[1]). Feel free to look at `Crowdsourcing Analytics - Final Manuscript.pdf` for a more detailed description of the experiment, its features, and the different methods participants applied to tackle the question.

[1] in case you want to improve skin color ratings on your own

2.1 Loading and Cleaning the Data (10 points)

The first step consists of loading the .csv file and preparing the data for the experiment. One participant of the official experiment provided a nice jupyter notebook demonstrating how the python library `pandas` can be utilized to achieve this: `http://nbviewer.ipython.org/github/mathewzilla/redcard/blob/master/Crowdstorming_visualisation.ipynb`. You should get inspiration from this example, but still choose your own data preparation steps. The following questions may guide you:

- What do the feature names (e.g. column `games`) stand for?
- Which irrelevant features might be dropped?
- What relevant features might be missing, but can be computed? E.g., you can obtain the age of a player (which might be relevant) from his birthday, or create entirely new features by non-linear combinations of existing ones.
- Are there missing data values (e.g. missing skin color ratings), and how should they be dealt with? (see `https://en.wikipedia.org/wiki/Missing_data`)
- How good are the skin color ratings? Do the raters agree?
- Should referees with very few appearances be excluded from the dataset?
- Should features be normalized and/or centralized?

Categorical features (e.g. `league`) should be transformed to a one-hot encoding (see `https://en.wikipedia.org/wiki/One-hot`). In case of `league`, you can also repreat the experiment independently for the different leagues to check if there are differences between countries. Provide a detailed description and justification of your data preparation.

2.2 Model Creation (8 points)

Given features X_i of player i, we want to predict $Y_i = N_{i,\text{red}}/N_i$, the fraction of games where the player will receive a red card. We will solve this problem using two model types: linear regression and regression forests.

Linear regression determines a weighted sum of the features $\hat{Y}_i = X_i\hat{\beta} + \hat{b}$, where optimal weights and intercept minimize the squared error:

$$\hat{\beta}, \hat{b} = \text{argmin}_{\beta,b} \sum_i (X_i\beta + b - Y_i^*)^2$$

A regression forests works similarly to a decision forest (reuse your code from exercise 4), but leaf responses and split criteria differ:

- The response of leaf b_l is the average response of the training instances assigned to this leaf:

$$\bar{Y}_l = \frac{1}{N_l} \sum_{i \in b_l} Y_i^*$$

- The optimal split into children b_λ and b_ρ minimizes the squared error:

$$\sum_{i \in b_\lambda} (Y_i^* - \bar{Y}_\lambda)^2 + \sum_{i \in b_\rho} (Y_i^* - \bar{Y}_\rho)^2$$

Moreover, the test `'not node_is_pure(node)'` makes no sense for regression trees and should be eliminated. The forest's reponse is the average response of its trees.

Implement both models and determine their squared test errors by means of cross-validation. Alternatively (or in addition – this will result in bonus points), you can also try to predict $Y_i = p(\text{red card} \mid X_i)$ via the posterior of a classification model.

2.3 Answering the Research Question (6 points)

Now perform a *permutation test* to answer the research question. To this end, create 19 new training sets where the skin color variable is randomly shuffled among the players. Each dataset uses a different permutation of skin colors, but keeps all other features and the response intact. This ensures that any possible association between skin colors and responses Y_i^* is destroyed, whereas the marginal skin color distribution gets preserved.

Determine the squared errors of the two model types on these new datasets by cross-validation as well. If all 19 datasets exhibit higher test errors than the original unscattered dataset, you can conclude that there is a skin color bias in red card decisions with a p-value of $p = 1/20 = 0.05$. If so, determine the direction of the bias by comparing the average of the Y_i^* for light and dark colored players.

2.4 How to lie with statistics (6 points)

Play with the data cleaning procedure with the following goal: Find two equally plausible cleaned datasets that give opposite answers to the research question, i.e. one uncovers a skin color bias, and the other does not.

If you succeed in finding such datasets, it demonstrates how easy it is in practice to tweak the data in the direction of the desired outcome, and how careful one needs to be conducting statistical research and interpreting published results.

2.5 Alternative hypotheses (6 points)

Keep in mind that a statistical analysis like this can only reveal *correlations* between features and response, but says nothing about the direction of causality (statistical analysis of causality is also possible, but requires more powerful methods and larger datasets). Provide two alternative plausible causal hypotheses, besides the obvious "referees discriminate against dark colored players", that might explain a possible correlation. Test your hypotheses with the data at hand.