# Advanced Machine Learning: Homework 2

*Note: this text is dummy translation of task formulated in Russian:*

*https://docs.google.com/document/d/1MapaSpMax0cEnnY2-QJ2XMgzVHpj_UrucJK-1r5FgA8/edit#heading=h.2yf7hlrm983*

The second assignment is a full-fledged data analysis project, from analyzing the problem statement to comparing the results of different models. The task is real and serious, although I chose an entertaining topic: we will build a probabilistic rating system for the sports “What? Where? When?" (ChGK).

Background: in the sporty “What? Where? When?" competing teams answer the same questions. After a minute of discussion, the teams write down and submit their answers on cards; the winner is the one who answered the most questions. The tournament usually consists of several dozen questions (usually 36 or 45, sometimes 60, more rarely). There are often synchronous tournaments, when teams on hundreds of playgrounds around the world answer the same questions. hundreds or even thousands of teams can play in one tournament. Accordingly, we need:

- build a rating list that is able to predict the results of future tournaments in a non-trivial way;
- at the same time, since ChGK is a hobby, and there are no contracts here, players are constantly moving from team to team, a strong player can sit down for one tournament to play for another team, etc .; therefore, the unit of the rating list should not be the team, but the individual player;
- And what greatly simplifies the task and translates it into the area of homework for the EM-algorithm is the nature of the data: from some point on, all question-by-question results of the teams began to be entered into the results base, i.e. the data will contain records like “which team answered which question correctly”.

I took only the first step for you: I downloaded all the necessary data through the API of the ChGK rating site so that the site would not fall under your numerous scrapers. :) The received data is in pickle format here:

https://www.dropbox.com/s/s4qj0fpsn378m2i/chgk.zip

1. Read and analyze the data, select tournaments that have data on team lineups and results by question (mask field in results.pkl). For unification, I suggest:

    - take tournaments with dateStart from 2019 to the training set;
    - in test - tournaments with dateStart from 2020.

2. Build a baseline model based on linear or logistic regression that will train the player ranking list. Notes and hints:

    - the results by question are actually the results of the toss of a coin, and their prediction is most likely related to the binary classification;
    - in different tournaments, questions of completely different levels of difficulty, so the model should take this into account; most likely, the model will have to explicitly train not only the strength of each player, but also the complexity of each question;
    - for the baseline model, you can forget about the teams and assume that the question-by-question results of a team simply refer to each of its players.

3. The quality of the rating system is assessed by the quality of predicting tournament results. But our models will hardly be able to predict the results themselves by question, because it is not known how difficult the questions will be in future tournaments; and these predictions are not needed by themselves. Therefore:

   - Suggest a way to predict the results of a new tournament with known squads but unknown questions, in the form of team rankings;
   - As a quality metric on the test set, let's consider the Spearman and Kendall rank correlations (they can be taken in the scipy package) between the real ranking in the tournament results and the predicted model, averaged over the test set of tournaments.

4. Now the main thing: ChGK is still a team game. Therefore:

   - suggest a way to take into account the fact that several players are answering the question at once; hidden variables are likely to be needed; feel free to make simplifying assumptions, but now the variables “player X answered question Y”, given the data, should become dependent for players of the same team;
   - develop an EM circuit for training this model, implement it in code;
   - train several iterations, make sure that the target metrics grow over time (most likely not much, but should grow), choose the best model using the target metrics.

5. And what about the questions? Build a "rating list" of tournaments by the difficulty of the questions Does it correspond to intuition (for example, at the World Championships in general there should be difficult questions, and at tournaments for schoolchildren - simple ones)? If it is interesting: build a top of complex and simple questions with links to specific entries in the ChGK questions database (this is a purely technical matter, there is no ML here).