Project 4: Analysis of marathon results [due May 14 by midnight]

Download the following csv file which contains race results of about 26,000 marathon runners: marathon_results.csv

Objectives

Analyze the marathon data. In particular:

1. Compute 1-dimensional kernel density estimate (KDE) of male and female runners using finish times, and use it together with the Bayes theorem to compute the probability that a runner with a given time was a female. Use this to make predictions if a runner was a male/female based on their finish times and check accuracy of these predictions.
2. Repeat part 1, but using 2-dimensional KDEs computed using finish times and ages of runners.
3. Compare accuracy of predictions obtained in parts 1 and 2 to the predictions made using k-NN with the same input data.
4. Use linear regression to predict finish times of runners based on their 5K times. Evaluate accuracy of these predictions. Then use other data beside the 5K time (the age of a runner, whether the runner was a male or a female) together with the 5K times to predict finish times using linear regression, and check if this meaningfully improves the predictions.
5. Add anything else that you find relevant and interesting.

Note: Tools for computing KDE are implemented by several Python libraries. You can use, for example, `scipy.stats.gaussian_kde` which is a part of the scipy library. On the other hand, you must not use ready-made Bayes classification tools implemented in `sklearn` and other machine learning packages. You can use `sklearn` to compute k-NN classification. For linear regression you can use the implementation provided by `sklearn`.