Titanic data classification via Naive Bayes classifier

Authors

Made by Kirill Prakhov and Danil Zakablukovskii.

Classification

Predictions were made using **Naive Bayes** classifier.
Achieved score: **0.77033**

For classification few attributes, which we considered most important, were used:

- `Pclass`, that represents the person ticket class. This is important because people with higher ticket class would have higher social status, that usually leads to higher change of surviving.
- `Sex`, that represents the person sex. Important because one can imagine that woman would have a higher change to survive.
- `Age`, represents the person age. Important since probably people older or too young could be not fast and/or coordinated enough to go to the boats.

After many attempts to create additional features via feature engineering or use other columns, we ended up using only three columns specified above, because usually we were getting results even worse than before.

Normalization

`Age` attribute has missing data for some persons, therefore we had to fill missing rows by using quite rough approximation — calculated mean age.

`Sex` attribute required us to convert string values to their number analogues, so `male` become `0`, `female` become `1`.

Training

Model was trained following the traditional Naive Bayes algorithm, where `Age` attribute was used as a continuous attribute, therefore instead of calculating probabilities for each possible value, we saved the `mean` and `std` values, which would be used later for classification via PDF.

Classification

Classification is pretty simple: for each class we calculate the probability of the attribute value having the value it has. In case of continuous attributes (as `Age` for example), we calculate the probability by using probability density function with the given `mean` and `std` (or variance) params determined on **Training** step.