Mathemaical Underpinnings of Machine Learning – project checkpoint

Jan Smoleń, Bartosz Siński

We have chosen **Project A – Feature selection.**

Progress

We have decided to implement methods using IT measures with following criteria:

- CIFE
- JMI
- Min- max Criterion
- Forward selection with Mutual Information

And following other feature selection methods:

- Boruta
- Variable importance based on mean decrease in impurity in the Random Forest.

For the evaluation we are using:

- **Artifical dataset where a set of significant features is known:** We generate features from the standard normal distribution and choose value k which will indicate the number of significant features. Then we set Y= 1 if $\sum_{j=1}^{k} X_j^2 > \chi_k^2(0.5)$ and Y = 0 otherwise, where $\chi_k^2(0.5)$ is the median squared distribution with k degrees of freedom.
- **Artifical example where MI method won't work:**
  Two approaches:
  - Random noise: we generate n features from standard normal distribution and Y from $Y \sim Bern(p)$ for some p
  - Random noise with two jointly-singificant features: we generate n-2 features from normal distribution and two i.i.d features $x_{n-1}, x_n \sim Bern(p)$. Then, we set Y to $I[x_{n-1} = x_n]$. With this approach, MI-based forward selection shouldn't be able to capture this dependence because of one-by-one selection of next added features.
- **And five real-world data set examples:**
  Variables will discretized if needed
  1. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
  2. https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

3. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
4. https://www.kaggle.com/datasets/mirichoi0218/insurance
5. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction