

QSAR biodegradation dataset classification using neural networks

Jonáš Kulhánek, Jan Uhlík

MFF, Charles University in Prague

December 2019

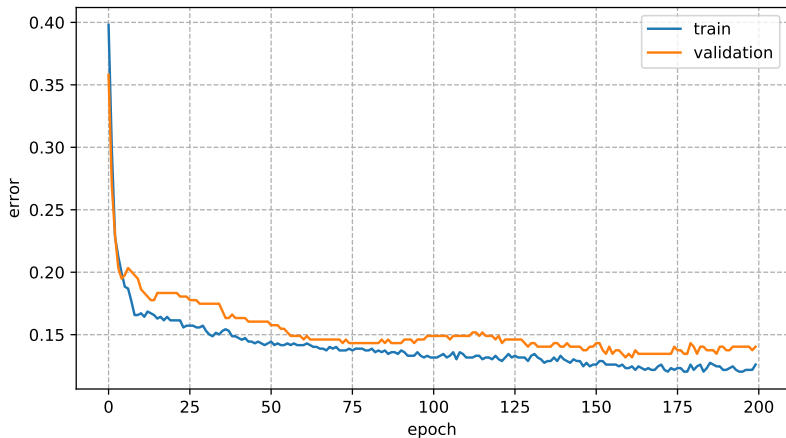
QSAR biodegradation dataset Kamel et al. (2013)

- ▶ Dataset from OpenML: qsar-biodeg
- ▶ Classification dataset where the task is to predict whether a protein is bio-degradable based on its chemical properties
- ▶ Published by Milano Chemometrics and QSAR Research Group
- ▶ 41 attributes in total, some real valued, other integer valued
- ▶ Number of proteins in the dataset is 1055
- ▶ Class 0 is bio-degradable

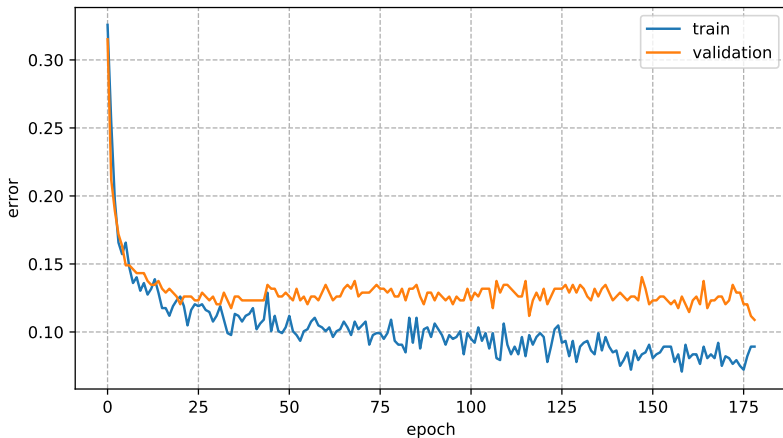
Method

- ▶ Multi Layered Perceptron (2 layers)
- ▶ 40 neurons in the hidden layer
- ▶ ReLU activation function
- ▶ trained as logistic regression with NLL loss and sigmoid activation function
- ▶ optimized using Adam TODO: add citation
- ▶ used dropout (0.5) and weight decay (0.01) for regularization
TODO: add citation
- ▶ trained for 200 epochs

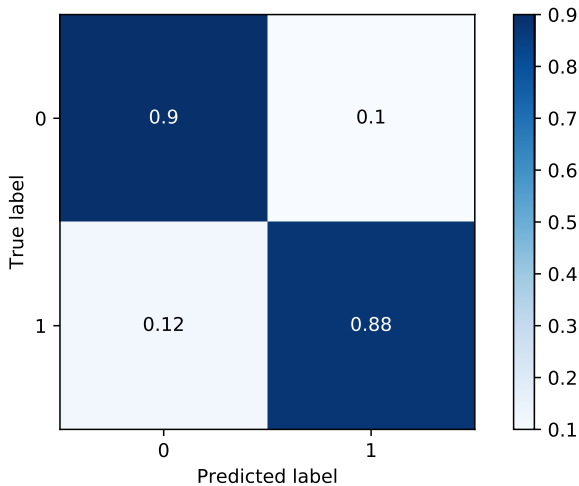
Single layer network training



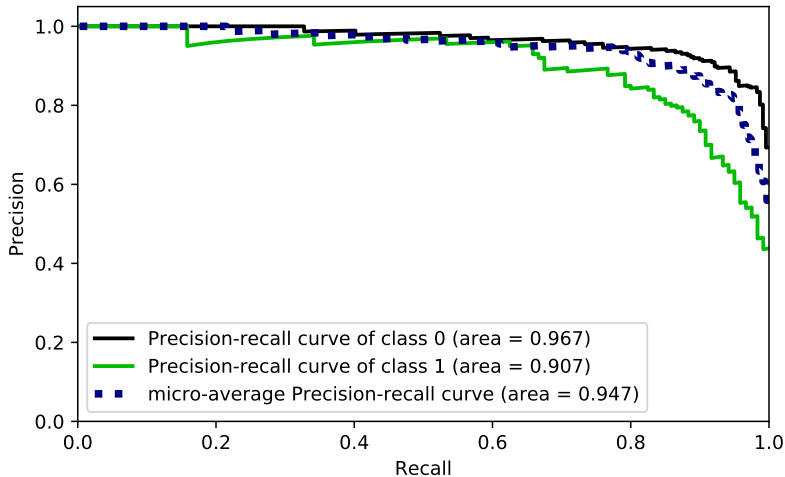
Two layers network training



Confusion matrix



Precision recall



Our results and comparisons

Method	Predictive accuracy
<code>svm.classes.SVC</code>	89.19%
Ours - 2 layers	89.12%
<code>xgboost</code>	87.68%
<code>weka.RandomForest</code>	86.67%
Ours - 1 layer	85.96%
<code>weka.VotedPerceptron</code>	79.43%
<code>weka.NaiveBayes</code>	75.64%
<code>neural_network.MLPClassifier</code>	66.26%

Mansouri Kamel, Tine Ringsted, Davide Ballabio, Roberto Todeschini, and Viviana Consonni. Quantitative structure-activity relationship models for ready biodegradability of chemicals. *Journal of chemical information and modeling*, 53: 867–878, 03 2013. doi: 10.1021/ci4000213.