QSAR biodegradation dataset classification using neural networks

Jonáš Kulhánek, Jan Uhlík

MFF, Charles University in Prague

December 2019

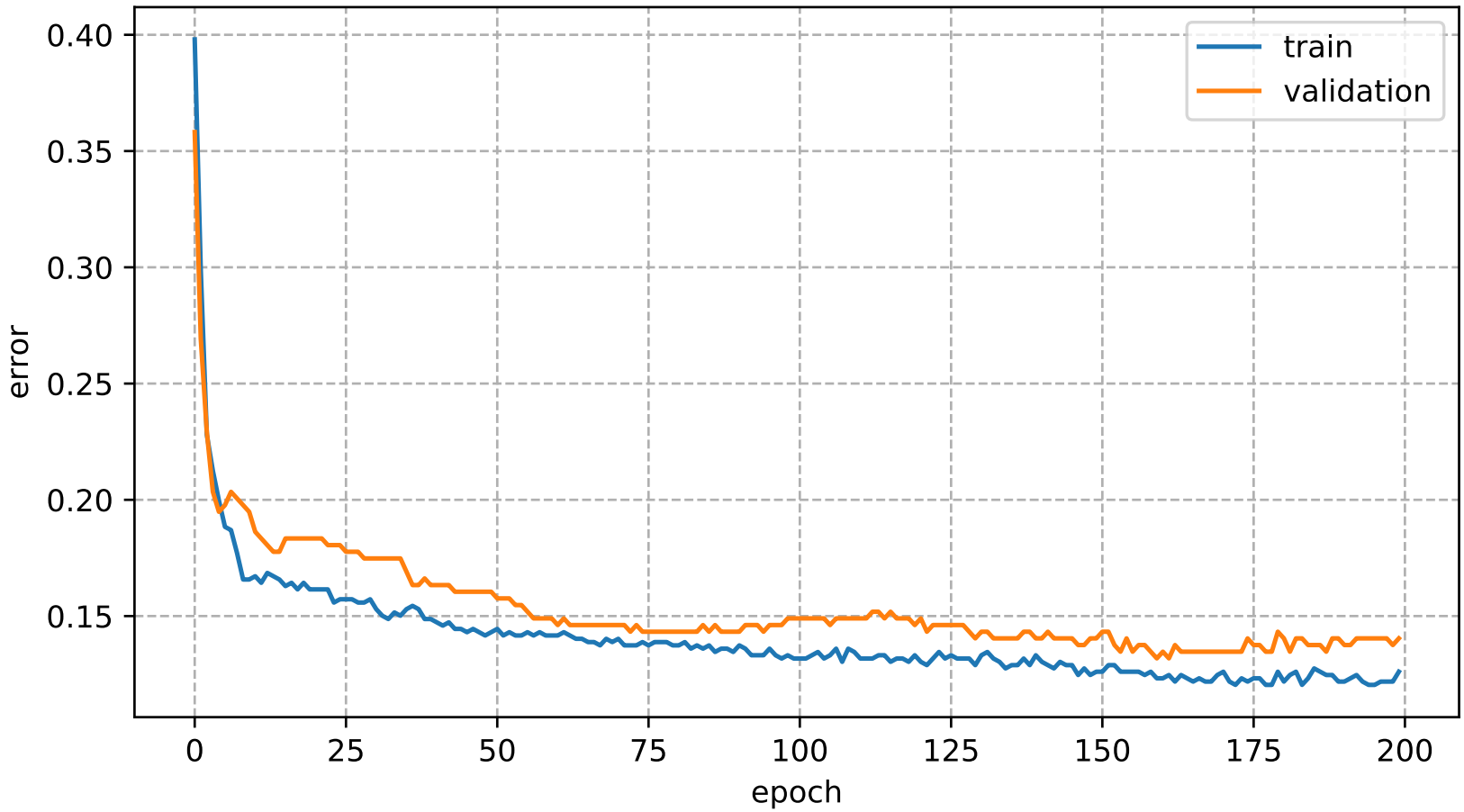
QSAR biodegradation datasetKamel et al. (2013)

- 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

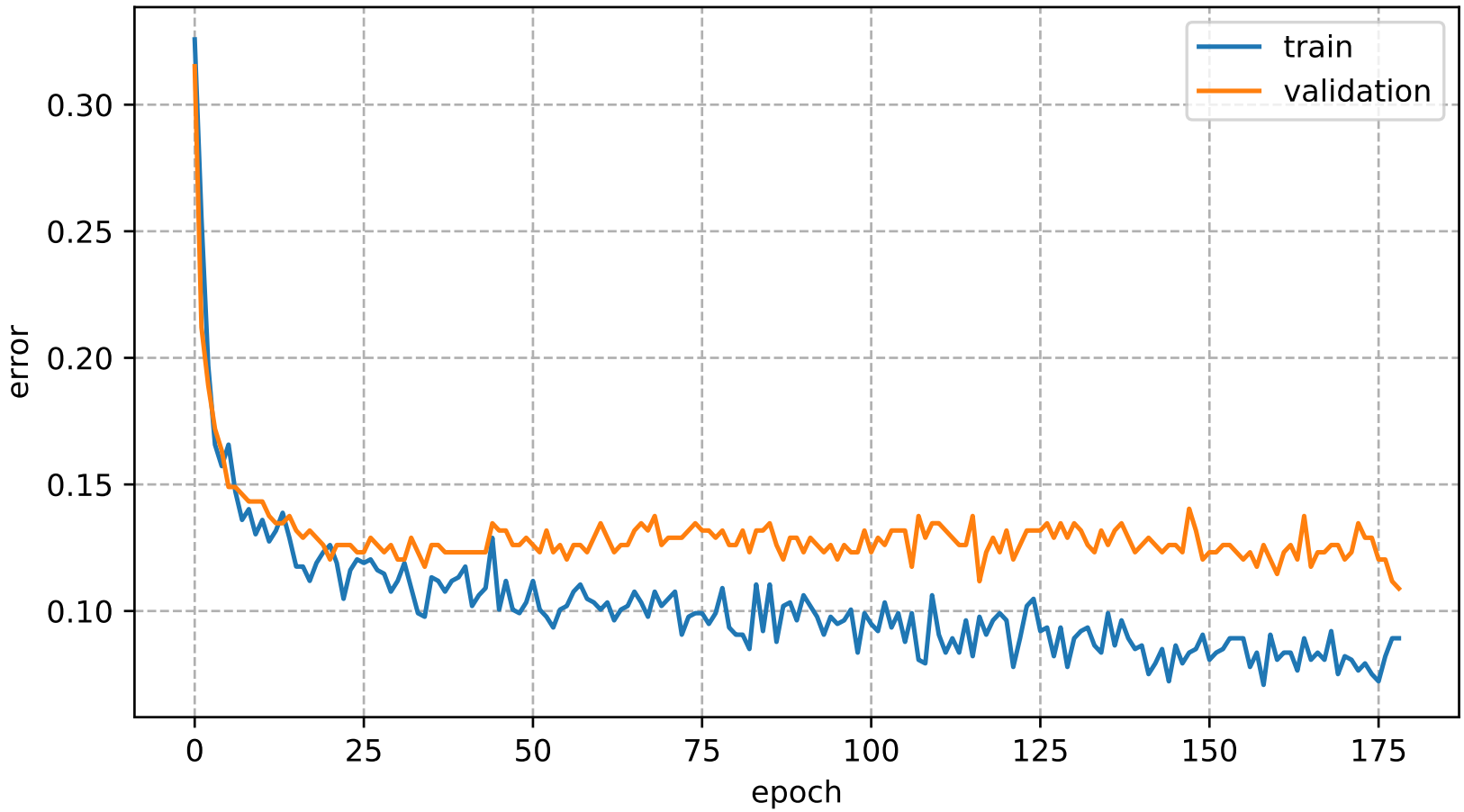
Method

- 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 citatoin
- trained for 200 epochs

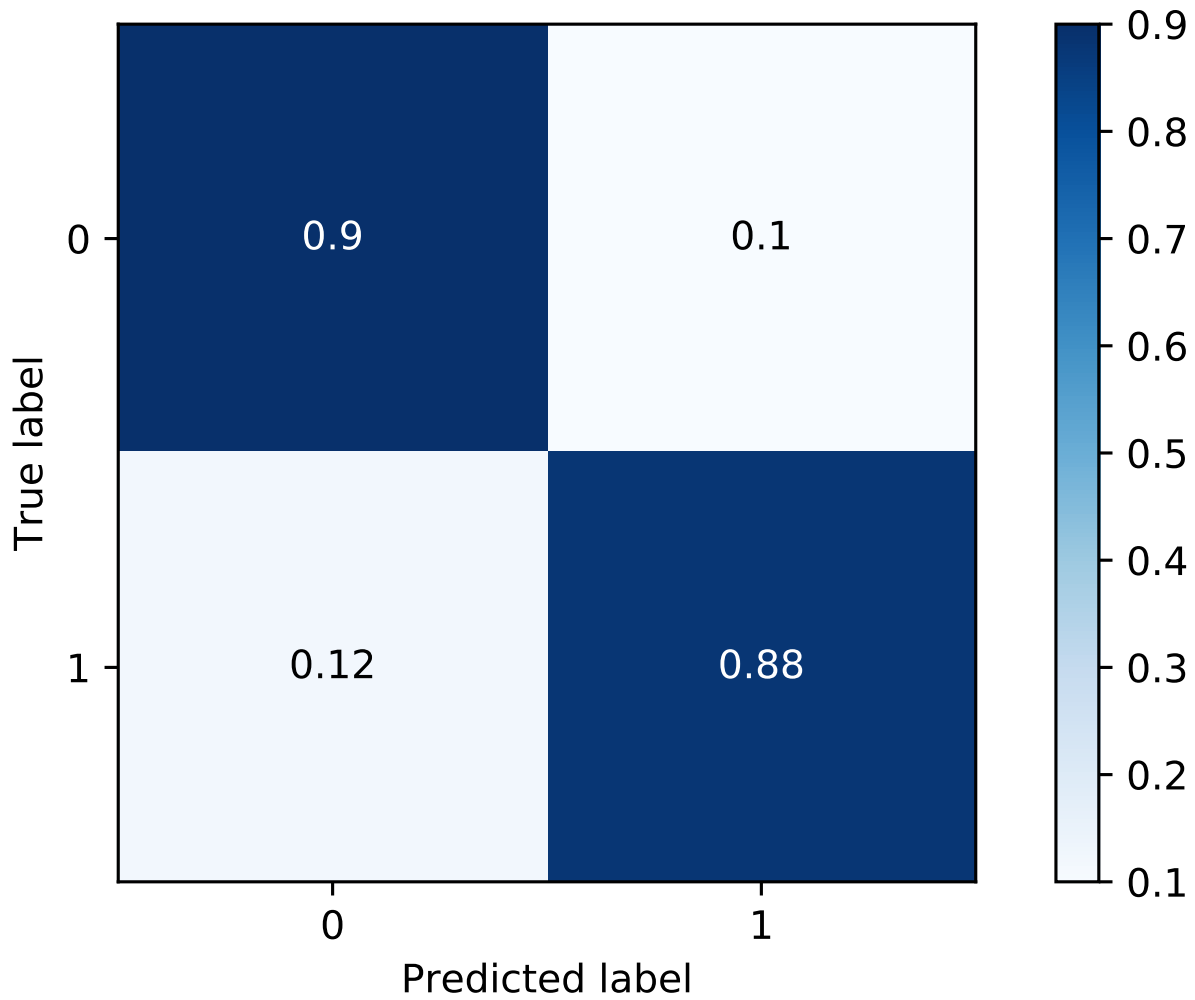
Single layer network training



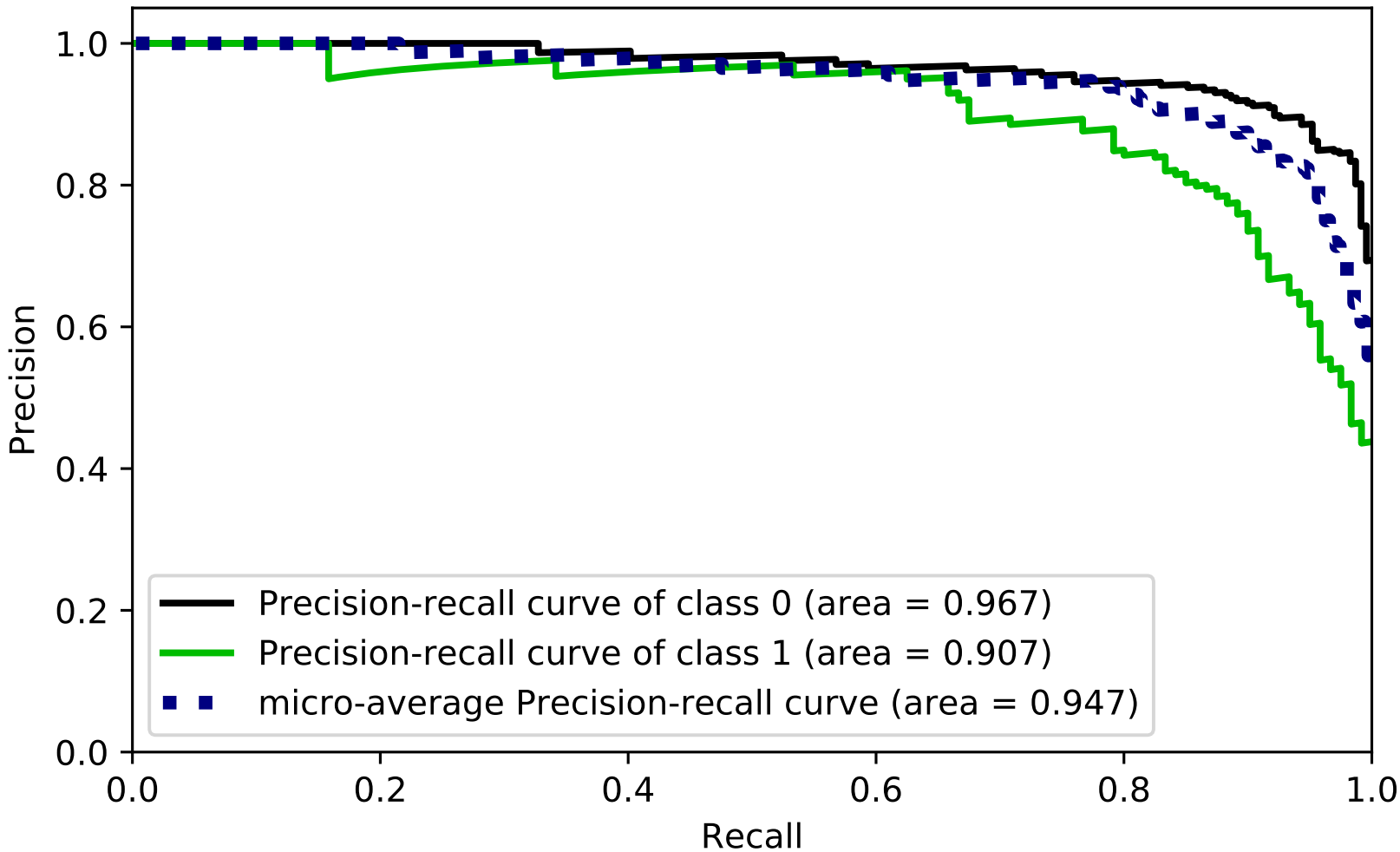
Two layers network training



Confusion matrix



Precision recall



Our results and comparisons

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