Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 6

07.06.2018
Deadline: 14.06.2018 - 16:00

Task 1: Metaparameter Optimization *(10 Points)*

In this task, you are supposed to apply some of the concepts from evaluation and validation to the parameter optimization of classifiers. The following subtasks have to be performed on three datasets: Car, Diabetes, and CoverType (`https://archive.ics.uci.edu/ml/datasets/Covertype`).

(a) Use the confidence parameter for pruning of `J48` classifier, which is `-C`, in WEKA and perform a 10 fold cross-validation for different values of that parameter. Plot the accuracy for values in $\{0.05 \cdot i \mid i \in \{0, 1, 2, ..., 10\}\}$.

(b) Implement a `CVparameterSelection` method, that takes as an input a dataset, one of `J48` options, and the possible range of that parameter (e.g., `start:steps:end` for real-valued options or `{a,b,...}` for nominals). Then runs a k-fold cross-validation on the data and trains a `J48` classifier for each possible value in the range. It returns the best value of the parameter as the output.

(c) Using your code from part (b), implement a classifier, called `OptimalJ48`, that performs a parameter optimization on the $-C$ parameter of the `J48` and classifies the instances using the resulting `J48` tree. Dataset path and parameter range is given as input parameters.

(d) Evaluate your `OptimalJ48` classifier on the proposed datasets. Use accuracy as the performance measure.

(e) Compare the parameter that has been selected by your `OptimalJ48` with the parameters in (a). Discuss the results.

Task 2: Mean Squared Error *(5 Points)*

Assume that the true relationship between the instances $x \in X$ and the target value $y \in Y$ is given by a function $f : X \to Y$, so that $y = f(x)$ for all instances. Both, X and Y are finite. A user samples a training set T of a fixed size n and uses a regression algorithm R to generate the regression function $g_T : X \to Y$ that is supposed to resemble the true relationship. He then applies the generated function to a fixed test instance x_0: $y_0 := g_T(x_0)$. The *expected mean squared error* of this prediction is given by

$$\text{MSE}(x_0) := \mathbb{E}_T[f(x_0) - g_T(x_0)]^2,$$

where E_T denotes the expectation over all possible training sets. Show that

$$\text{MSE}(x_0) = \text{Var}_T(y_0) + \text{Bias}^2(y_0)$$

where $\text{Var}_T(y_0) := \mathbb{E}_T[g_T(x_0) - \mathbb{E}_T g_T(x_0)]^2$ and $\text{Bias}^2(y_0) := [(\mathbb{E}_T g_T(x_0)) - f(x_0)]^2$.

Task 3: Multivariate Linear Regression *(5 Points)*

Given is the data in Table 1. It represents the annual expenses of various livestock markets. How do these depend on the number of animals sold? Create a multivariate linear regression model (with constant term) as described in the lecture for this task. You should use only basic linear algebra to build the model (matrix multiplication, transposition, solving linear equations etc., but *not* a complete linear least square fitting procedure). Give results and procedure for each step you performed (by hand or implemented).

Cattle	Calves	Pigs	Lambs	Expenses
(thousands)	(thousands)	(thousands)	(thousands)	(1000*dollars)
3.437	5.791	3.268	10.649	27.698
12.801	4.558	5.751	14.375	57.634
6.136	6.223	15.175	2.811	47.172
11.685	3.212	0.639	0.964	49.295
5.733	3.220	0.534	2.052	24.115
3.021	4.348	0.839	2.356	33.612
1.689	0.634	0.318	2.209	9.512
2.339	1.895	0.610	0.605	14.755
1.025	0.834	0.734	2.825	10.570
2.936	1.419	0.331	0.231	15.394
5.049	4.195	1.589	1.957	27.843
1.693	3.602	0.837	1.582	17.717
1.187	2.679	0.459	18.837	20.253
9.730	3.951	3.780	0.524	37.465
14.325	4.300	10.781	36.863	101.334
7.737	9.043	1.394	1.524	47.427
7.538	4.538	2.565	5.109	35.944
10.211	4.994	3.081	3.681	45.945
8.697	3.005	1.378	3.338	46.890

Table 1: Expenses of livestock markets.

(a .csv version of this table is available in /datasets)