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

Machine Learning - Sheet 5

25.05.2017
Deadline: 01.06.2016 - 23:55

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 publicly available datasets. Use the Car dataset from the previous tasks and two additional datasets (e.g., `http://archive.ics.uci.edu/ml/` or `http://www.cs.waikato.ac.nz/ml/weka/datasets.html`).

(a) Make yourself familar with the `J48` classifier of WEKA and find out how to set confidence parameter for pruning, which is `-C`.

(b) Perform a 10 fold cross-validation for different values of the pruning confidence and plot the accuracy for each value. You have to use at least the values $\{0.05 \cdot i \mid i \in [0; 10]\}$.

(c) Make yourself familar with `CVParameterSelection` of WEKA (`https://weka.wikispaces.com/Optimizing+parameters#CVParameterSelection`) and re-implement the `CVParameterSelection`. Notice that you must not copy any code!

(d) 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.

(e) Randomize the datasets and split it into training set, validation set, and test set.

(f) Evaluate your `OptimalJ48` classifier on the public datasets. Use at least the accuracy as performance measure.

(g) Compare the parameter that has been selected by your `OptimalJ48` with the parameters in (b).

(h) Discuss the results.

Task 2: Naive Bayes for Text Categorization *(10 Points)*

The goal of this exercise is to apply the Naive Bayes Classifier algorithm to text categorization.

(1) Recall the definition of conditional independence: Given some finite set of elementary events Ω with probability mass function $\mathbb{P}[\{\bullet\}]$, and three events $A, B, C \subseteq \Omega$, we call A and B *conditionally independent given* C, if it holds:

$$\mathbb{P}[A \cap B \,|C] = \mathbb{P}[A \,|C]\,\mathbb{P}[B \,|C]\,.$$

Furthermore, we call A and B just *independent*, if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$.

Let Ω be the set consisting of eight binary strings:

$$\Omega := \{000, 001, 010, 011, 100, 101, 110, 111\}$$

and let $\mathbb{P}[\{\omega\}] = \frac{1}{8}$ for all $\omega \in \Omega$. Find events $A, B, C \subseteq \Omega$ such that A and B are independent given C, but not independent.

(2) What is the Naive Bayes assumption of conditional independence? How does it help us to design practical classification algorithms?

(3) Demonstrate on a small artificial example why it is a good idea to add a weighted prior when estimating the conditional probabilities of words given the document class.

(4) Implement the Naive Bayes Classifier (`Exercise_05_02.java`). Try it out on the 20Newsgroups dataset *(training and test dataset are provided, loading and rudimentary preprocessing is already implemented)*. Report your results.

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 (thousands)	Calves (thousands)	Pigs (thousands)	Lambs (thousands)	Expenses (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)