Assignments 7, 8 and 9
Machine Learning Project

Machine Learning, Summer term 2015
Prof. Ulrike von Luxburg, Morteza Alamgir, Tobias Lang

May 28, 2015

Total number of points: 60 (25+25+10)

Overview

Assignments 7, 8 and 9 are the final assignments for the lecture part of Prof. von Luxburg. These assignments are different from the previous assignments: you will not need to present any results in class. Your task is to **investigate a data-set by yourself**, write two reports about your studies and **grade reports** of fellow students:

1. **Assignment 7** (25 points, due June 17) (Sec. 2): Perform an unsupervised data analysis study where you study the general characteristics of a data-set; write a report about your study.

2. **Assignment 8** (25 points, due June 24) (Sec. 3): Perform a predictive modeling study on the same data-set; write a report about your study.

3. **Assignment 9** (10 points, due July 1) (Sec. 5): Grade reports of fellow students. *This is mandatory for each student to pass the whole course.*

For Assignments 7 and 8, you can work in a **team with up to 3 people** and hand in jointly made reports. Assignment 9 has to be done by each student individually.

Participation in the **exercise classes** on June 8/10, June 15/17 and June 22/24 is optional. You can discuss your current work with your fellow students and the teaching assistants there.

Note that each of these assignments is work for about one week, but we shifted the due dates due to the mini-exams. In particular, also the maths students need to work on assignments 7 - 9, even though they don't need to attend the classes after June 15.

1 Data-sets

We use data-sets from **Kaggle** (`www.kaggle.com`). "Kaggle is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models." (Wikipedia)
Please choose *one* of the following data-sets available at Kaggle for your studies in Assignment 7 and 8 (the same data-set for both assignments). To access these data-sets, one student per group needs to register at Kaggle. Some of these challenges are not active and can be accessed in the section for completed challenges.

- Otto Group Product Classification Challenge: multiclass classification with vectorial features, convenient for matlab users
- Bike Sharing Demand: regression problem with seasonal data, convenient for matlab users
- Restaurant Revenue Prediction: regression problem, convenient for matlab users

---

Web page: `http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/teaching/2015-ss-vorlesung-ml/`
Login: "machine", Password: "learning"

- Africa Soil Property Prediction Challenge: regression problem on high dimensional data, convenient for matlab users
- Search Results Relevance: needs pre-processing the textual features, better to use python or R

Take your time for your decision: the data-sets involve different sub-problems (natural language processing, time series etc.) and the respective prediction problems vary in difficulty. Which data-set is a good choice depends also on which software and toolbox you want to use (e.g., Matlab is less suitable for natural language processing). You can find plenty of information on the data-sets together with helpful hints, scripts and tutorials on the **Kaggle forum and scripts pages**. Use these information sources!

2 Assignment 7: Unsupervised Data Analysis (25 points, due to June 17, 20:00h.)

In the first study, you have to analyze the basic characteristics of your chosen data-set. The goal is to sharpen your general understanding of the data. This is an **unsupervised** task: do not take into account too much the predictive modeling task of Assignment 8.

2.1 Tasks and Report

You have to write a report of up to **2,000 words** about your study. Take the general remarks in Sec. 4 on how to write the report into account. Your report will be used to grade your work. It should describe your work along the following tasks:

Short Introduction (2 points) Give a *short* overview over the data-set. Explain the features and their types.

Data Analysis (20 points) Describe your findings of the data analysis part along the following points:

- Calculate **basic statistics** (means, medians, variances, empirical probabilities) (8 points).
  - Calculate statistics for appropriate subgroups.
  - Identify class inbalances and non-normalized features.
- Identify potentially **interesting features** and subgroups of features (2 points).
  (Are there features which are correlated with each other? Can you identify groups of uncorrelated features, for example by clustering features with respect to the correlation coefficient as similarity function? Which features seem to have a lot of explanatory or predictive potential for other features?
- Run at least one **clustering** method on the data (3 points).
  If necessary, preprocess the data. You may focus on specific feature subgroups for clustering (this means that you only use some of the features to cluster your data). Interpret the found clusters and identify meaningful clusters.
- Run at least one **dimensionality-reduction** method (3 points).
  Can you simplify the problem with dimensionality reduction methods? What might be a good choice for the dimension parameter?
- Are there any other aspects of the data that you find interesting?

In general, note that you can analyze your data-set in many different ways. There is not a single correct expected analysis. You are also not expected to perform all possible analyses one can conceive. Your report shall show that you know how to deal with quantitative data-sets, that you manage basic data analysis methods and that you know how to interpret and convey your findings.

You can use whichever **toolboxes, software libraries and programming and scripting languages** you like. Popular choices include: Matlab, Weka, Orange, sklearn, Octave, Python (numpy, pandas), R, libsvm ...

Conclusions (3 points) Provide concluding remarks about your findings. Describe what your findings suggest for the predictive modeling task in Assignment 8.

References When you use special algorithms, techniques, or software libraries, provide references to the papers, journals and websites.

2.2 What to hand in

- Report: You need to hand in your report as a PDF-file (not larger than 5 MB) on an online plattform, the URL will be provided soon on the course webpage. Your report has to be anonymous (due to peer grading).
- Please send your complete code for data analysis or screenshots (when using a toolbox) as a ZIP file to your teaching assistant by email. This file needs to make clear that you have done the data analysis yourself.

3 Assignment 8: Predictive Modeling (25 points, due to June 24, 20:00h)

Each Kaggle data-set comes with a well-defined predictive modeling task. In Assignment 8, you are asked to perform a study on the predictive modeling task of your chosen data-set.

3.1 Tasks and Report

You have to write a report of up to **1,000 words** about your study. Take the general remarks in Sec. 4 on how to write the report into account. Your report will be used to grade your work. It should describe your work along the following tasks:

Short Introduction (3 points) Describe the prediction task. Describe precisely the evaluation metric used in the Kaggle competition. Use this evaluation metric in your study (this might be an unusual metric which you might need to implement yourself).

Predictive Modeling (20 points)

- Describe data preparation steps including normalization. (1 point)
- Describe clearly which data you use for training, validation and testing and how you optimize parameters. (2 points)
- Investigate at least three different **prediction methods**. (12 points (4+4+4))
  For instance, for classification you might use SVMs, logistic regression, kNN, decision trees etc. Among them must be at least one method which was presented in class. Investigate different parameterizations (kernel type, distance measure etc.).
- Illustrate your results with plots.
- **Discuss** your findings: (5 points)
  - Which model would you recommend to use?
  - Are the overall results satisfactory? Could your predictive model be used in practice?
  - Use a lot of skepticism and point to limitations, potential problems and shortcomings.
  - Can you interpret your models? What do they tell you about the data?
  - Can you hypothesize any causal effects? Or do important explanatory variables seem to be missing?
  - If possible, compare your results to the results of other participants in the Kaggle competition.

Conclusions (2 points) Provide concluding remarks about your findings. What types of methods achieved the best results? Where would you continue future work? How could your methods be improved?

3.2 Remarks on Predictive Modeling in Practice

- Make sure to distinguish between **training, validation, and test data**. In particular, do not optimize your models on the test data. Use cross-validation.
- **Test data:** Each Kaggle data-set contains official test data. The correct test labels are not disclosed. You can evaluate your models on this data by making submissions to Kaggle. In your report, you have to provide results on the Kaggle test data. However, you can usually make only a limited number of submissions to Kaggle per day. Thus in case you require more test runs, it might make sense to split the training data into your personal training and test data and use this personal test partition for further evaluation of your models. This allows you to perform several train/test splits and report results over multiple runs.
- Use the same **evaluation metric** as used in the Kaggle competition. This might involve implementing it yourself.
- Make sure to **preprocess and normalize** your data before learning.
- Pay attention to **feature selection**. The following procedure is a common error: You take the whole data set and select features that are nicely correlated with the labels. Then you separate your data in train and test, train on train and test on test. This will look great, but is completely wrong (why?).
- Include **variances** in your results whenever possible:
  - When using your own train/test partitions, run your methods multiple times on different partitions.
  - When your algorithm depends on initializations with different random seeds (e.g., kNN), run it multiple times with different random seeds.
- Again, you can use whichever **toolboxes, software libraries and programming and scripting languages** you like.

3.3 What to hand in

You need to hand in:

1. Your full report as a PDF file, not larger than 10 MB, to the same online platform as last week. The report should now contain **both** the unsupervised part (which you already wrote last week) and the supervised part. Your report has to be anonymous (due to peer grading).
2. Your complete code (for predictive modeling, running experiments, preparing figures) or screenshots (when using a toolbox) as a ZIP file by email to your teaching assistant. This file needs to make clear that you have done the predictive modeling yourself.

4 Report Requirements

- The report should be written in English (German only in exceptional cases). If possible, use Latex to write your report.
- The goal is NOT to present the most competitive results (if you can do so, this is great, but we don't expect this). Instead, someone who reads your reports should be able to
  - understand what you did,
  - judge whether your approach makes sense and your results seem correct, and
  - **evaluate your findings**.

- Most importantly, the reader needs to be able to **reproduce** the results from reading only your report (not from looking at your code!). Therefore, you need to provide accurate descriptions of what you did, in particular report all necessary details on the chosen methods, parameters, train and test data etc.
- **Plots:**
  - To report results, provide *meaningful* plots and discuss their contents in the text. Think well which are plots that are helpful to convey your story. The point of the report is not to provide a collection of 50 uncommented plots. The purpose of the plots is to illustrate what you write in the text.
  - Avoid tables, they are hard to digest.
  - All plots need to have captions, axis descriptions and full legends.
  - Each plot needs to be discussed within the text. **Always discuss whether your findings are expected or unexpected, surprising or obvious, good or disappointing, etc.** If you find something unexpected, you should try to explain the reason in the text (but sometimes this is impossible, and then you might write that you find it unexpected and why, but that you don't have a good explanation for the behavior).
  - Potentially interesting plots:
    * Empirical probability distributions of the values of features (bar plots for discrete features, bar plots using bins for continuous features)
    * Two-dimensional PCA plots of the data
    * How does the accuracy of a method vary with a parameter? (e.g., σ on x-axis, accuracy on y-axis)
    * Bar plots to compare the accuracy of different methods

5 Assignment 9: Peer Grading (pass (=10 points) or fail, due to July 1, 20:00h)

In Assignment 9, you are asked to evaluate the studies of other students.
To pass the whole course, you need to participate in Assignment 9 with reasonable performance. In other words, every student is required to participate in peer grading and deliver satisfactory results (i.e., not giving random scores, provide reasons for the own judgement, etc). The gradings will be controlled by the teaching assistants.
Peer grading is going to be performed on our online platform. Every student gets an assignment of a small number of reports to grade. In the studies for Assignments 7 and 8, one can achieve a total maximum of 25 points each. How the total points are divided over subtasks is outlined above. In grading, you are asked to give points according to the following three criteria:

1. Work: Have the authors performed the required subtasks? Has the work been carried out correctly?
2. Discussion: Do the authors describe well the implications of their findings? Whether they were expected/unexpected, good/disappointing? Are the plots well-chosen?
3. Writing: Can the text be understood without problems? Can the plots be fully understood? Do the plots contain axis labels and captions?

We will provide more details about the peer grading on the online platform once it is running.