Exercises in Machine Learning
Introduction and Visualizing

Zdeněk Žabokrtský, Ondřej Bojar
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Tue Mar 7, 2017
Outline

- Motivation: Why machine learning (ML)?
- Quick overview of typical ML tasks
- Why ML with Python?
- Random Tips for Python.
- Data and visualization HWK.
- Your data suggestions.
Motivation: Why ML?

- Lazy programmers.
- Too complex problem.
  - Too difficult to describe, easier to show examples.
  - Conflicting idea or unclear what is a “correct” result.
- Easier to customize or incorporate feedback.
  - Simply add more examples.
  - But there is a risk of reaching a plateau.

- Win and earn:
  - http://www.kaggle.com/
Typical ML Tasks

- Classification.
  - Should I go to cinema or cycling?
  - Is this beer or wine?
  - Which song is this?
  - Who is it behind the door?
    Is he going to give me money or ask for some?

- Regression.
  - How much should this whisky cost?
  - How many flowers should I buy for this lady?
  - When am I going to die?

- Clustering.
  - Divide dirty laundry from dirty dishes.
  - Divide students into groups: geeks, nerds, lamers...
    ...and recommend each group what they should drink.
Why ML in Python?

- Python is said to be:
  - Clean, elegant, easy to learn (i.a. because interactive)...
  - Good for teaching.

- Great Toolkits:
  - http://scikit-learn.org/ (advanced ML)
  - http://matplotlib.org/ (plotting graphs)
  - http://nltk.org/ (natural language processing)
  - http://tensorflow.org/ (deep learning; there are many more)
Random Python Tips

- Show which methods are supported by an object:
  ```python
x=[1,2,3]
dir(x)
```

- Jump into python console from a running program:
  ```python
  import pdb
  pdb.set_trace()
  ```
Dataset (HWK #03)

- Prepare (in pairs) a dataset:
  - Line-oriented CSV.
  - The last item is the expected answer.
  - Divided into train+test.

- We want to have the following types covered:
  - few observations, many features, 2 classes are enough
  - very few real-valued features (2D only), 5 to 10 classes
  - many observations, many features, many classes (>10)
  - many observations, many features, just a few classes
  - many real-valued features, some categorial features, 2 classes are enough

- Possible domains: medicine; demography, sociology; graphics (analysis or generation); speech processing; text processing

Due: 2 weeks from now, i.e. March 21.
Data Visualization

- Data visualization is extremely useful.
- Always plot the data when working with a new dataset.
- This is inherent part of your HWK #03.
  - Make sure to include an interesting curve fit to some aspect of your data.
- Suggested gradual steps on the following slides. (Python source on the seminar web page.)
Scatter Plot of Random Values
Scatter Plot of Activity – Heart Rate
Box Plot of Activity – Heart Rate
Histogram of Random Values
Histogram and Gaussian Function (not Fit!)
Histogram of Heart Rate when Lying

Heart rates hist. for activity 1
Hists of HR when Lying, Walking, Running
Least Squares for Linear Fit
Cubic Fit of a Histogram