Exercises in Machine Learning
Datasets and Visualization

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

Tue Mar 15, 2016
Outline

▷ Motivation: Why machine learning (ML)?
▷ Quick overview of typical ML tasks
▷ Why ML with Python?
▷ Random Tips for Python.
▷ Data HWK.
▷ Visualization HWK.
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:
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?

- Ondřej and Zdeněk want to learn 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)
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 #Data)

- Prepare (in small groups) 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 29.
Data Visualization

- Data visualization is extremely useful.
- Always plot the data when working with a new dataset.

Our dataset for this class:
- Subject 1 of the PAMAP2 dataset:
  - physical activity monitoring
  - URL: Google for PAMAP2

Read the description of the dataset on the web.
To avoid downloading the 600 MB, get the data from:
- http://ufallab.ms.mff.cuni.cz/~bojar/NPFL104/subject101.dat.gz (52 MB)
Homework #03

Make a python script and a wrapping Makefile so that:

- 'make' will plot **preferably one graph with three histograms**
  - Histograms of the heart rates of Subject 1 when Lying, Walking, and Running.
  - Each histogram with the least-square fit of the gaussian.
- Commit your solution to your SVN repo:
  - `svn co https://svn.ms.mff.cuni.cz/svn/undergrads/students/<your-login>/2015-npfl104/hw03/`
- Notify Ondřej by e-mail.

Due: 3 weeks from now, i.e. Apr 5.

Suggested gradual steps on the following slides.
Scatter Plot of Random Values
Scatter Plot of Activity – Heart Rate
Box Plot of Activity – Heart Rate
Gaussian Function
Histogram of Random Values
Histogram and Gaussian Function (not Fit!)
Histogram of Heart Rate when Lying

heartrates hist. for activity 1
Hists of HR when Lying, Walking, Running
Least Squares for Linear Fit

Data points are plotted on a graph with the x-axis ranging from 60 to 130 and the y-axis ranging from 50 to 250. The data is represented by red dots, and two lines are shown: one blue and one red. The blue line is labeled 'initial' and the red line is labeled 'optimized.'
Cubic Fit of a Histogram