Data Mining Assignment 1

Erik Zeitler

Uppsala Database Laboratory

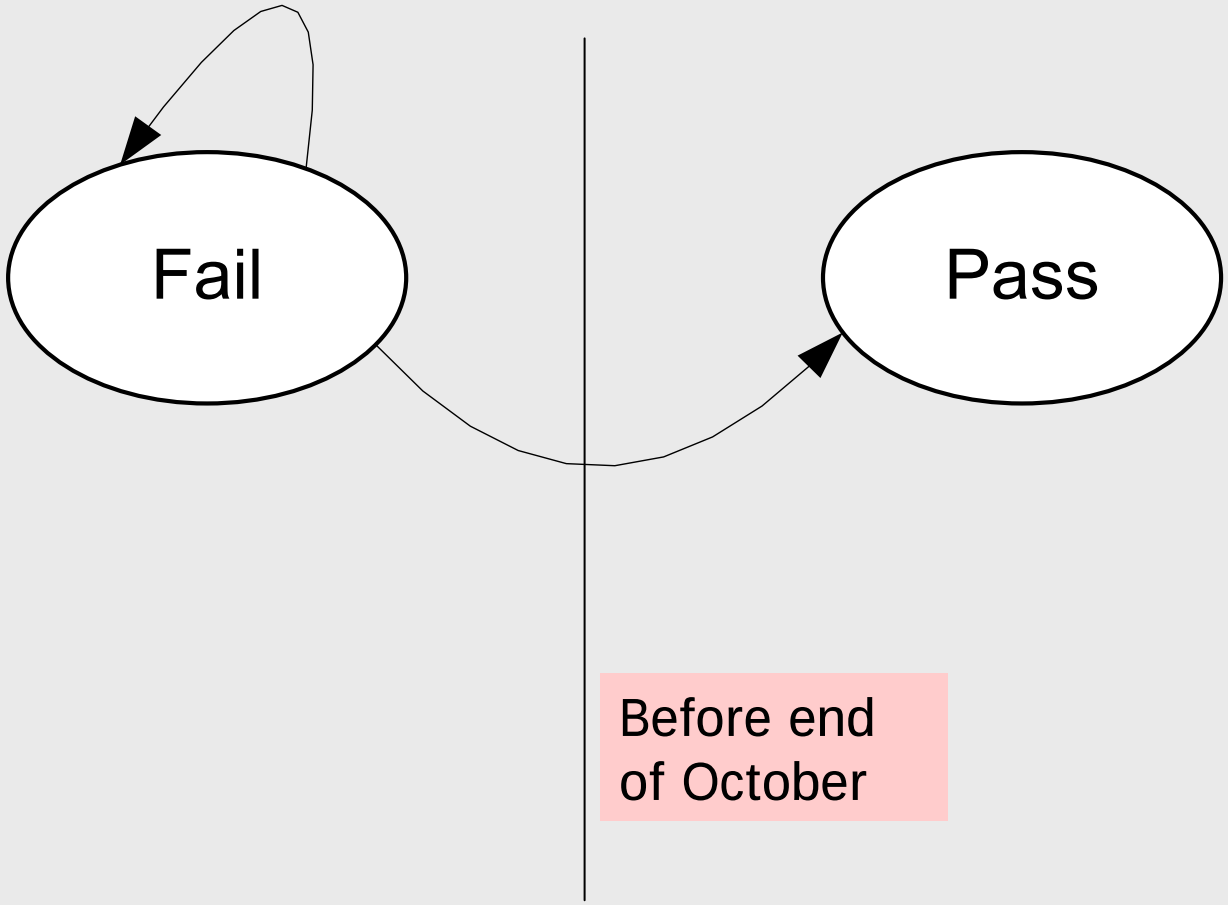
UPPSALA UNIVERSITET

Oral exam

- Two parts:
  1. Validation
     - Your solution is validated using a script
       - If your solution does not work
         → the examination ends immediately (“fail” grade is given)
         → you may re-do the examination later
  2. Discussion
     - Prepare answers to the written questions
     - The instructor will ask additional questions
       - about your solution
       - about the method
- All group members must be able to answer
  - Group members can get different grades on the same assignment

UPPSALA UNIVERSITET

Grades



UPPSALA UNIVERSITET

What you need to do

20, 22 Sep

- Sign up for labs and examination

28 Sep

  - Groups of 2 – 4 students
  - Forms are on the board outside 1346
- Implement a solution
  - Deadline: Submit by e-mail no later than 24h before your examination
    - 1: thanh.truong@it.uu.se. Subject: **`DM1-A1 Solution`**
    - 2, 3: andrej.andrejev@it.uu.se
- Prepare answers to the questions
- Prepare for the discussion
  - Understand the theory

UPPSALA UNIVERSITET

Assignment 1 in a nutshell

- You will get
  - 163 data points with *known* class belonging
  - 30 data points with *unknown* class belonging
  - A kNN implementation
- Improve classifier performance
  - Find the best *k*
  - Normalization (max-min, Gaussian)
  - Metrics (Minkowski *r* = ?)
  - Vote weighting
  - *Optional:* Attribute weighting
- Using your parameter settings, classify the 30 data points

UPPSALA UNIVERSITET

Things to consider: Know your data!

1. Normalize $\subset$ Pre-process
   - What is the range of each attribute?
   - Is one attribute more important than another?
     - If so, what should we do?
     - If not, should we do anything else?
   - You can assume: no missing points, no noise.
2. Select training + testing data
   - Is the data sorted?
     - Does it matter? If so, is this good or bad?
   - Are there any alternatives to leave-one-out cross validation?
3. Choose *k*
   - How do you know if the value of *k* is good?

UPPSALA UNIVERSITET

Know your data!

4. How many points of each class are there?
   - Should this observation affect the choice of *k*?
5. Choose distance measure
   - What distance measure is suitable? Why?
     - Euclid, Minkowski, and maxnorm are available in AmosII.
     - You can implement other distance measures, similarity measures, *etc*...
6. Classify unknown data
   - Should the unknown data be normalized? How?
   - Which data set should be used to classify the unknown data?