Coding data-mining assignement

A not so brief guide

Devert Alexandre
School of Software Engineering of USTC

Submitting your code

Name of the archive

If you are asked to submit your code as an archive named *dm-assignment-1-xxx*

$\Rightarrow$ *HOHAI-MyPrOJEct-LOL* is not correct

Name of the archive

If you are asked to submit your code as an archive named *dm-assignment-1-xxx*

$\Rightarrow$ *dm_assignment-1-xxx* is not correct

Name of the archive

If you are asked to submit your code as an archive named *dm-assignment-1-xxx*

$\Rightarrow$ *dm-assignment_1-xxx* is not correct

Name of the archive

If you are asked to submit your code as an archive named *dm-assignment-1-xxx*

$\Rightarrow$ *dm-assignment-1-xxx* is correct

Writing an email

Typical mail from a student submitting code

Title : *myself*
THIS IS MY ASSIGNEMENT LOL KTHXBAI

Ok, so 30 students will call themselves *myself* ? Also, email is not a phone message to your best friend.
(Note : It really happened. Several times.)

Writing an email

Title : *data-mining assignement 3 / SA111564987*
Hi,
I tested my work on the examples you gave me. I noted a problem with XXX. I tried YYY and ZZZ but somehow, I did not find out how to fix it. To change XXX in my code do BLABLABLA.
Cheers

A useful, clear title. A message with actual useful informations. Something that looks professional.

Making an archive of your work

Common issues when student archive their work

- 5 Ko of useful content, but the archive weights 20 Mo
- A lot of totally useless files
- Assumes I use the same tools as them
- Opening the archive is likely delete things by accident

Making an archive of your work

Some tips on how to archive your work

- Keep only the source files
- Remove the binary files (*.class*, *.o*, ...)
- Remove the backup files (*MyCrappyCode.java*$\sim$)
- Remove the Eclipse, Visual Studio, ... whatever project files your favorite IDE created
- Put your code in a directory with a sensible name, like *dm-assignment-1-xxx*

I don't force you to use my coding tools, don't force me to use your own tools. And emails that takes 20 mn to download are not cool.

Following requirements

Following requirements

What you are supposed to do at 20 years old or more

- Read carefully the assignment instructions
- Take your time for reading
- Do the questions in order
- Ask me if you are not sure about one point

Following requirements

What a lot of students usually do

1. Read half of each question only
2. Do half of each questions only
3. Do the questions in random order
4. Then call me because something does not work (how surprising !)

Following requirements

Caring for details makes customers happy and is a key to quality work

Caring for details is a training. Training starts *NOW*

Common coding errors

Computing the minimum of a list

Most of you do did this

```
double min(double* u, int size) {
  double min = 999999999.0;
  for(int i = 0; i < size; ++i)
    if (u[i] < min)
      min = u[i];
  return min;
}
```

Hard-coded limits . . . What if the data decide to not respect your limit ?

Computing the minimum of a list

A correct approach, assuming we have at least 1 element

```
double min(double* u, int size) {
  double min = u[0];
  for(int i = 1; i < size; ++i)
    if (u[i] < min)
      min = u[i];
  return min;
}
```

No hard-coded limits !

Computing the minimum of a list

A better version, using proper types

```
double min(const double* u, size_t size) {
  double min = u[0];
  for(size_t i = 1; i < size; ++i)
    if (u[i] < min)
      min = u[i];
  return min;
}
```

Some compilers can take advantage of constness. It can also help the compiler to find some errors.

Computing the minimum of a list

A shorter version, using the standard library

```
#include <algorithm>

double min(const double* u, size_t size) {
  double min = u[0];
  for(size_t i = 1; i < size; ++i)
    min = std::min(min, u[i]);
  return min;
}
```

Some compilers can take advantage of constness. It can also help the compiler to find some errors.

Computing the minimum of a list

An even shorter version, using the standard library

```
#include <algorithm>

double min(const double* u, size_t size) {
  return *std::min_element(u, u + size);
}
```

Reading about the language you use, reading the standard library documentation is *NOT* a waste of time.

Using the standard library

Nobody will teach you every single detail of a programming environment.

Practice, curiosity, self–motivation, or stay a low-grade coder

Using the standard library

If you are coding in Java

- *java.util.Arrays* to fill, sort, print, reverse, copy, (and more) arrays of any type
- *System.arrayCopy* for fast copy between arrays
- *Math* have min and max functions
- *java.util.Collections* have also some nice things to help
- *Double* to parse strings as double, maximum possible value, and more

Using the standard library

If you are coding in C++

- *algorithm* for fill, sort, partial sort, reverse, copy, (and more) any kind of collection of any type
- *math* for sin, cos, fabs, floor, ceil, pow, (and more) for float and double
- *limits* for highest possible value of a given type, and more

Computing the median of a list

Most of you did this

```
#include <algorithm>

double median(double* u, int size) {
  std::sort(u, u + size);
  return u[size / 2];
}
```

This is wrong. This is *not* the definition of median !

Computing the median of a list

A corrected version

```
#include <algorithm>

double median(double* u, int size) {
  int m = size / 2;
  std::sort(u, u + size);

  if (size % 2 == 0)
    return 0.5 * (u[m - 1] + u[m]);
  else
    return u[m];
}
```

Note that we use *size_t* now ...

Computing the median of a list

A faster version

```
#include <algorithm>

double median(double* u, size_t size) {
  size_t m = size / 2;
  std::partial_sort(u, u + m + 1, u + size);

  if (size % 2 == 0)
    return 0.5 * (u[m - 1] + u[m]);
  else
    return u[m];
}
```

Partial sort will sort just the number of elements we need to have sorted.

Comparing floating point numbers

Most of you did this

```
bool equals(double a, double b) {
  return a == b;
}
```

Floating point computations are not exact. Exact comparisons makes little sense.

Comparing floating point numbers

A better version (but can be much better)

```
#include <math>

bool equals(double a, double b, double tol=1e-8) {
  return std::fabs(a - b) < tol;
}
```

At least, here, we can adjust the level of tolerance.

Reading data files

To read a matrix stored in a text file, typical student code

- Create an array with 1000 elements
- Read the file line by line
- For each line, read 5 numbers

So, if you have more than 1000 lines ? Or more than 5 values per line ? Program crashes, of course !

Reading data files

A better way to read a matrix stored in a text file

- Create an empty list A
- For each line
  - Read the numbers in a list B
  - Append B to A
  - Empty B

After 4 years of university, can not imagine this by yourself ?! Or just too lazy ? Both way, no excuses for not doing it.

Data–mining algorithms

Coding data–mining algorithms requires precision

Being sloppy won't make you finish sooner.

Coding discipline

Writing portable C/C++

Do not use the *system* function.

You have no idea if the command you call with *system* exist on others computers. And you do not need this for assignment work.

Writing portable C/C++

Something I see often

```
int
main(int argc, char* argv[]) {
  // Do things

  // End of the program
  system("pause");
  return 0;
}
```

pause does not exist on Linux. Besides, it's terrible coding practice. Don't be lazy, open a console and run your program from there.

Writing portable C/C++

Do not use hard-coded path for input/output files

C:/xiao-huang-gua/dm/test.txt might exist on your computer. But it is just *YOUR* computer, not *OTHER* computers. If I ask for command line tools which take paths as command line parameters, there is a reason for this !

Writing portable C/C++

If you use command-line parameters, you will not have this problem

```
int
main(int argc, char* argv[]) {
  // Read command line
  if (argc != 3) {
    cerr << "Wrong number of parameters." << endl;
    cerr << usage_str << endl;
    return EXIT_FAILURE;
  }

  input_file_path  = argv[1];
  output_file_path = argv[2];
  cluster_count    = str_to_int(argv[3]);

  // Do stuffs
}
```

Writing portable C/C++

The standard, portable *main* function is not

```
void
main() {
  /* My wonderful program */
}
```

Writing portable C/C++

The standard, portable *main* function is not

```
void
main(int argc, char* argv[]) {
  /* My amazing program */
}
```

Writing portable C/C++

The standard, portable *main* function is not

```
int
main(int argc, char* argv[]) {
  /* My incredible program */
  return 0;
}
```

Writing portable C/C++

This is the standard, portable *main* function

```c
#include <stdlib.h>

int
main(int argc, char* argv[]) {
  if (problems)
    return EXIT_FAILURE;

  /* Job done, no problems */
  return EXIT_SUCCESS;
}
```

Writing portable C/C++

If you code in C++, use the C++ headers.

```
#include <stdlib.h>
#include <stdio.h>
#include <string.h>
#include <time.h>
```

In C++ you do this

```
#include <cstdlib>
#include <cstdio>
#include <cstring>
#include <ctime>
```

And then you are sure it will compile and link properly everywhere

Writing portable C/C++

No, not everybody will compile your code with Visual Studio.

```
#pragma once

#include <stdafx.h>
```

Tying your code to the software of one particular company is not a very smart move in general

Writing portable C/C++

This is not supported by the C++ standard, as it introduce nasty ambiguities

```
std::vector<std::vector<double>> matrix;
a = b >> 3;
```

Do this, and your code will compile everywhere

```
std::vector< std::vector<double> > matrix;
a = b >> 3;
```

Avoiding to write inefficient code

Passing parameters by *value*

```
double
sum(std::vector<double> x) {
  double ret = 0.0;
  for(int i = 0; i < x.size(); ++i)
    ret += x[i];
  return ret;
}
```

Copy the whole data at every function call ! 10000 values ? Copy 10000 values !

Avoiding to write inefficient code

Passing parameters by *reference*

```
double
sum(std::vector<double>& x) {
  double ret = 0.0;
  for(int i = 0; i < x.size(); ++i)
    ret += x[i];
  return ret;
}
```

Only a *reference* to the data is passed, no copies involved.

Avoiding to write inefficient code

One keyword is here to help you $\Rightarrow$ *const*

```
double
sum(const std::vector<double>& x) {
  double ret = 0.0;
  for(int i = 0; i < x.size(); ++i)
    ret += x[i];
  return ret;
}
```

const for data which are not changed

- compiler can exploit this to increase efficiency
- compiler will tell you if you try to change what should not change

Avoiding to write inefficient code

Use the proper data types

```
double
sum(const std::vector<double>& x) {
  double ret = 0.0;
  for(size_t i = 0; i < x.size(); ++i)
    ret += x[i];
  return ret;
}
```

It will remove some common compiler warnings

Avoiding to write inefficient code

Iterators are your friends

```
double
sum(const std::vector<double>& x) {
  double ret = 0.0;
  std::vector<double>::const_iterator it = x.begin()
  for(; it != x.end(); ++it)
    ret += (*it);
  return ret;
}
```

This code save one addition per iteration, compared to above version

Avoiding to write fragile code

Your code should not crash for datasets you did not used for test

- Loading of datasets $\Rightarrow$ use dynamic data-structure
- Avoid hard-coded limits
- Avoid parameters which depends on a particular dataset
- Fix compiler warnings
- Free the memory you allocate

Avoiding to do unreadable code

- C, C++, Java have *separate compilation*. Put in different files things which need to be reused.
- C, C++, Java have *functions*. Don't copy-paste the same code multiple time Use a function for that repeated code.
- C, C++ and Java are *object-oriented*. Avoid having 10 functions that are all related, and all have the same 8 parameters. Those parameters and functions can be part of an object.

Care

Who is going to a good work mate ?

- The one who don't pay attention to anything ?
- The one who try to make the life easier for others ?

Which work mate you will be ?

Do not do to others what you do not want done to yourself – Kong Fuzi

Coding for an assignement

An assignement is an opportunity to improve your skills

Give me the best of you, I will give you the best of me