CS1632: TESTING THEORY AND TERMINOLOGY

Wonsun Ahn

Expected vs. Observed Behavior

- *Expected behavior*: What “should” happen
- *Observed behavior*: What “does” happen

- *Defect*: when expected != observed behavior
- *Testing*: checking expected == observed behavior

- Expected behavior is also known as *requirement*

Example

- Suppose we are testing a function `sqrt`:
  ```
  // returns the square root of num
  float sqrt(int num) { … }
  ```

- When `float ret = sqrt(9);`,
  Expected behavior: `ret` == `3`

- When `float ret = sqrt(-9);`,
  Mathematically, square root of `-9` can't be a real number, but requirements should still specify some behavior

THE IMPOSSIBILITY OF EXHAUSTIVE TESTING

- Let's say we want to ensure that `sqrt` is defect-free for all arguments (both positive and negative)
- Assume arg is a Java `int` (signed 32-bit integer)
- How many values do we have to test?

4,294,967,296

What if there are two arguments?

- Suppose we are testing a function `add`:
```
// return the sum of x and y
int add(int x, int y) { … }
```

- How many tests do we have to perform? (Hint: all combinations of `x` and `y`)

4,294,967,296 ^ 2

What if the argument is an array?

- Suppose we are testing a function `add`:
```
// return sum of elements in A
int add(int[] A) { … }
```

- How many tests do we have to perform?
(Note: array `A` can be arbitrarily long)

4,294,967,296 ^ Infinity

Would testing all the combinations of arguments guarantee that there are no problems?

LOL NOPE

- Issues causing defects even after exhaustive testing
  - Compiler issues
  - Systems-level issues (e.g. OS/device-dependent defect)
  - Parallel programming issues (e.g. data races)

- The same input must be tested multiple times
  - On different compilers, OSes, devices, …
  - (Potentially) many times on same compiler / OS / device

Compiler Issues

- The compiled binary, not your source code, runs on the computer
- What if compiler has a bug? (Rare)
- What if compiler *exposes* a bug in your program? (More frequent)

```
int add_up_to (int count) {
  int sum, i;    /* Is sum == 0? Not necessarily! */
  for(i = 0; i <= count; i++) sum = sum + i;
  return sum;
}
```

  - Behavior is undefined according to C language specifications
  - Compiler can generate code with arbitrary behavior

☞ Need to exhaustively verify with all compilers and compiler options!

Parallel programming issues

```
class Main implements Runnable {
    public static int count = 0;
    public void run() {
        for(int i=0; i < 1000000; i++) { count++; }
        System.out.println("count = " + count);
    }
    public static void main(String[] args) {
        Main m = new Main();
        Thread t1 = new Thread(m);
        Thread t2 = new Thread(m);
        t1.start();
        t2.start();
    }
}
```

```
$ javac Main.java
$ java Main
count = 1868180
count = 1868180
$ java Main
count = 1033139
count = 1033139
```

Why?

Parallel programming issues

- Why does this happen?
  - Threads `t1` and `t2` execute concurrently
  - Two threads try to increment `count` at the same time
  - Often, they step on each other's toes (a *data race*)
- If there is a data race, result is undefined
  - Java language specifications say so!
  - Every time you run it, you may get a different result
  - Result depends on relative speed of threads `t1` and `t2`

☞ Running 1000+ times may not cover all behavior

For the purposes of this Chapter…

- Let’s ignore these issues for now
  - Combinatorial testing issues
  - Compiler issues
  - Systems-level issues
  - Parallel programming issues

- Exhaustive input value testing is hard enough
  - a.k.a. “test explosion problem”
  - This is what we will focus on in this chapter

- We will address the other issues later 😉

Equivalence Classes

Achieving Test Coverage Efficiently

Defining Test Coverage

- Goal of testing: achieve good *test coverage*
  - Test coverage: measure of how rigorously code has been tested

- Ideally, *test_coverage = defects_found / total_defects*
  - But is there a way to measure *total_defects*?
    (Hint: if we knew, we wouldn't need to do any testing!)
  - Impossible to measure true test coverage

- Then how do you know you've achieved good coverage?
  - Use a proxy coverage metric that estimates true coverage
  - *statement_coverage = statements_tested / total_statements*
  - Rationale: if a high percentage of statements are tested
    ☞ likely that a high percentage of defects have been found

Improving Test Coverage

- QA engineers have a limited testing time budget
  - Since true test coverage is impossible to measure, must choose tests maximizing a proxy coverage metric
  - Most commonly, maximizing statement coverage

- Which tests are likely to maximize statement coverage?
  - Tests that exercise all required program behaviors
  - If tests exercise only one specific program feature or program behavior → likely to have low statement coverage
  - This is the idea behind *equivalence class partitioning*

Equivalence Class Partitioning

- Partition the input values into "equivalence classes"
  - Equivalence class = group of values with similar behavior
- E.g. equivalence classes for our `sqrt` method: {*nonnegative_numbers*, *negative_numbers*}
- Behavior for each equivalence class:
  - *nonnegative_numbers*: returns the square root of number
  - *negative_numbers*: returns NaN (not a number)

Equivalence Classes should be *Strictly* Partitioned

- *Strictly*: each value belongs to one and only one class

- If an input value belongs to multiple classes
  - Means requirements specify two different behaviors for the input
  - Either requirements are inconsistent, or you misunderstood them

- If an input value belongs to no class
  - Means requirements do not specify a behavior for the input
  - Either requirements are incomplete, or you misunderstood them

Values can be Strings

- For a spell checker, input values are strings
- Equivalence classes:
  {*strings_in_dictionary*, *strings_not_in_dictionary*}
- Behavior for each equivalence class:
  - *strings_in_dictionary*: do nothing
  - *strings_not_in_dictionary*: red underline string

Values can be Any Object

- Input values can be tuna cans
- Equivalence classes:
  {*not_expired*, *expired_and_not_smelly*, *expired_and_smelly*}
- Behavior for each equivalence class:
  - *not_expired*: eat
  - *expired_and_not_smelly*: use it in your rat trap
  - *expired_and_smelly*: discard

Test Each Equivalence Class

- Pick at least one value from each equivalence class
- Ensures you cover all behavior expected of program
- Gets you good coverage without exhaustive testing!

- How to pick the value? Well, that is part of the art.
  - However, there are some good guidelines!

Interior and boundary values

- Empirical truth:
  - Defects are more prevalent at boundaries of equivalence classes than in the middle.
- Why?
  - Due to the prevalence of off-by-one errors

Off-by-one Error

- Suppose expected behavior is:
    - Method shall take the age of a person as argument
    - Method shall determine whether person can be US president
    - Rule: Person must be 35 years or older to be US president

- Suppose code implementation is:

```
boolean canBePresident(int age) {
   return age > 35;
}
```

- Is observed behavior the same as expected behavior?

Equivalence class partitioning

CANNOT_BE_PRESIDENT =
[…19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34]

CAN_BE_PRESIDENT =
[35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50…]

Always Test Boundary Values

CANNOT_BE_PRESIDENT =
[…19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34]

CAN_BE_PRESIDENT =
[35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50…]

- Always test boundary values (shown in red).
- In fact, there is a bug at 35: `age > 35`

Also Test a few Interior Values

CANNOT_BE_PRESIDENT =
[…19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34]

CAN_BE_PRESIDENT =
[35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50…]

- Testing interior values (in green) is also important.
- Who knows? There may be a non-off-by-one error.

Are we done?

CANNOT_BE_PRESIDENT =
[…19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34]

CAN_BE_PRESIDENT =
[35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50…]

- Input values so far: {26, 30, 34, 35, 39, 42}

“Hidden” (IMPLICIT) boundary values

- Some boundaries are implicit – they come from the language, hardware, domain, etc.:
  - Language boundaries: MAXINT, MININT
  - Hardware boundaries: memory space, hard drive space
  - Domain boundaries:
    weight can’t be negative, score can’t exceed 100, etc.

- Why do we check implicit boundaries?

Crossing implicit boundaries should not change behavior

- Why do we check **explicit boundaries**?
  - To verify behavior **changes** when boundary is crossed.

- Why do we check **implicit boundaries**?
  - To verify behavior **does not change** on the boundary.
  - Suppose requirement does not specify change in behavior even when system runs out of memory space. → Good idea to check behavior remains the same!

Add implicit boundary values

CANNOT_BE_PRESIDENT =
[MININT,...-2,-1,0,1,...,25,26,27,28,29,30,31,32,33,34]

CAN_BE_PRESIDENT =
[35,36,37,38,39,40,41,42,43,44,45,46,47,...,MAXINT]

- MININT, MAXINT: language boundaries
- -1, 0: domain boundaries (age can't be negative)
- Inputs: {MININT, -1, 0, 26, 30, 34, 35, 39, 42, MAXINT}

Finding the Off-by-one Error

- Now let's feed these inputs to our code:

```
boolean canBePresident(int age) {
    return age > 35;
}
```

  Inputs: {MININT, -1, 0, 26, 30, 34, 35, 39, 42, MAXINT}
- Remember, requirement was:
  - Person must be 35 years or older to be US president
- An off-by-one-error is found with input 35:
  - Expected behavior: Can be president
  - Observed behavior: Cannot be president

Base, edge, and corner cases

- **Base case**: An expected use case
  - Interior value of equivalence class for normal operation
- **Edge case**: An unexpected use case
  - Boundary value of equivalence class for normal operation
- **Corner case** (or **pathological case**):
  - Value far outside of normal operating parameters
  - OR multiple edge cases happening simultaneously

Base, edge, and corner cases: Example

- Suppose a cat scale has these operating envelopes:
  - Weight between 0 – 100 lbs
  - Temperature between 0 – 120 F
- Base cases: (10 lbs, 60 F), (20 lbs, 70 F), …
- Edge cases: (**100 lbs**, 70 F), (10 lbs, **0 F**), …
- Corner cases: (**300 lbs**, 70 F), (**100 lbs**, **120 F**), …
- Why test corner cases?
  - Even if scale isn't expected to operate correctly for 300 lbs, user still cares what happens (i.e. does it break the scale?)

Categories of Testing: Black / White / Gray Dynamic / Static

Black-, white, and gray-box testing

- **Black-box testing**:
  - Testing with no knowledge of interior structure or source code
  - Tests are performed from the user's perspective
  - Can be performed by lay people who don't know how to program

- **White-box testing**:
  - Testing with explicit knowledge of the interior structure and codebase
  - Tests are performed from the developer's perspective
  - Test inputs are crafted to exercise specific lines of code

- **Gray-box testing**:
  - Testing with some knowledge of the interior structure and codebase
  - Knowledge comes from partial code inspection or a design document
  - Performed from the user's perspective, but informed by knowledge

Black-box testing examples

- Tests are performed using only UI

- Examples:
  - Testing a website using a web browser
  - Testing a game by actually playing it
  - Testing a script against an API endpoint
  - Any type of beta test
  - Penetration testing on a website

White-box testing examples

- Tests are performed by both…
  - Using UI to exercise specific program paths
  - Explicitly calling methods from a testing script
- Examples
  - Choosing inputs to exercise specific parts of an algorithm
  - Choosing inputs causing exceptions and checking handling
  - Testing that a method call returns the correct result
  - Testing that instantiating a class creates a valid object
- Unlike black-box, can measure **statement coverage**

Static vs dynamic testing

- We talked a great deal about choosing good inputs
  - But is this all there is to testing?
- Dynamic testing = code is executed
  - Relies on good inputs for good coverage
- Static testing = code is not executed
  - There are no inputs since code is not executed
  - Relies on analyzing the code to find defects

Dynamic testing

- Code is executed under various test scenarios
  - Varying input values, compilers, OSes, etc.
  - Observed results are compared with expected results
  - Hard to achieve 100% test coverage
- Examples:
  - Manual testing
  - Unit testing
  - System testing
  - Performance testing

Static testing

- Code is analyzed by a person or testing tool
  - While checking whether correctness rules are followed
  - 100% test coverage achieved for all code analyzed
  - Even when check passes, defects can still occur at runtime
- Examples:
  - Code reviews by a person
  - Code analysis using a tool
    - Compilers
    - Linters
    - Bug pattern finders
    - Code coverage analysis
    - Model checkers

Now Please Read Textbook Chapters 2-4