

Notes on Machine Learning Projects

Charles Elkan
elkan@cs.ucsd.edu

November 8, 2007

These notes are based on issues that I noticed in the reports submitted in October 2007 by CSE 250B students for the first project. I have tried to group together related points, but the issues are definitely not mentioned in any order of importance.

On any report, include the date and some sort of contact information for the authors, in addition to the names of the authors. This extra information lets you claim priority for your work, and lets readers reach you.

Make anything you write easy to read for a busy reader. Don't make it forbiddingly long, don't use heavyweight paper, don't use an unusually large or small font; do make margins be at least one inch, and do have page numbers. Have a table of contents if and only if the documents is too long to flip through quickly. To make quick scanning easier, section headings should usually be informative short phrases (with or without verbs), not single words or names such as "Results" and "Perceptron."

Always summarize your findings in an abstract and/or in the introduction. Don't write sentences that are almost content-free such as "Experiments and results are discussed." Instead, describe the experiments and state the results briefly. Remember that many readers will read only the abstract and/or the introduction; maximize the useful information that these readers get. Don't bother suggesting an objective for future work (e.g. "preprocessing to reduce noise") unless you also have a specific suggestion for how to achieve the objective.

Write in a plain, taut, precise style that is neither colloquial nor stilted. For example, change "To test an example, you would only need to take the inner product ... " to "To test an example, compute the inner product ... " Organize your report logically, not chronologically. Write in the present tense as much as

possible. Don't use past or future tenses unless what you are describing is time-specific in a central way.

Writing manuals often say to avoid the passive voice. This is usually good advice, but the more important goal is to avoid sentences that do not have a meaningful subject. Consider "Perceptron classification has been done of the MNIST data." Make it clear *who* did the classification.

If you derive a mathematical result that requires several steps of argument, then express the result as a lemma or theorem, with a clear statement and a separate proof. This makes the result easier to check, easier to think about intuitively separate from its proof, and easier to reuse.

Make sure that the logic of your work and of your arguments is clear and correct. The following are examples of mistakes in logic.

- A conceptual claim: "*k*-nn performs better than 45-way perceptron because *k*-nn is directly multiclass."
- Real-world thinking: "92% accuracy seems useful for most applications."
- Algorithm analysis: "*k*-nn is faster for smaller *k*."
- In experiment design, randomizing the order of training or test data is pointless with a nearest-neighbor classifier.
- In algorithm design, breaking ties randomly for the majority of 45 binary choices is not sensible.

Implementation efficiency is important in order to make your results firmly established, even if no one else will ever reuse your code. For example, one team reported 10-way classification results based on 60,000 training and 10,000 test examples for a perceptron method but based on only 3000 and 500 examples for their nearest-neighbor method. This discrepancy makes it impossible to conclude anything useful about the relative accuracy of the methods.

Use general knowledge and back-of-the-envelope calculations to evaluate the efficiency of your code. Consider the statement "62 million distance calculations took close to two hours on a desktop computer." Approximately, $62 \cdot 10^6 \cdot 784 = 50 \cdot 10^9$ floating point operations, in 7000 seconds. A modern 3 GHz processor can do at least one gigaflop per second, so a speedup of 100 should be achievable.

Get details correct in your writing. For example, "total error percentage is 0.146" is 100 times smaller than "error rate is 0.146." State findings quantitatively, e.g. don't just write "Nearest-neighbor takes much longer." Do not be imprecise

when being precise takes no more space, e.g. do not write “There are about 6000 training examples.” Note that many popular datasets are available in multiple versions on the web. The exact number of examples, of features, etc., is a useful indicator of exactly which version is used in a given paper.

Use words in precisely correct ways. All the following are incorrect: “both algorithms yield incredibly low error rates” (do you not believe your own results?) “increase the number of the datasets” (should be cardinality, not number) “error rates depend on the demographics of the training and test sets” (human populations have demographics, not datasets), “using an arbitrary definition of distance” (arbitrary means not well-motivated; “an arbitrary” should be just “any” here). Use standard spelling, including “Euclidean” and “Laplace” not “Euclidian” and “LaPlace.” Remember the difference between “principle” and “principal.”

Avoid cryptic phrases like “may lead to noise interference.” This example has the “nouns in a row” problem: does it mean “interference caused by noise” or “noise caused by interference” or what? Also, it uses words that are fundamentally vague: noise and interference can each have many different technical meanings.

Follow the norms for grammar and punctuation. Avoid sentence fragments; conversely, avoid run-on sentences. The subject and the main verb of a sentence should not be separated by a comma; do not write “All 30 runs, converge at six epochs.” Also follow the norms for mathematical notation. The standard symbols to use for multiplication are \cdot and \times depending on context. The symbols $*$ and \bullet are almost never correct.

Think about how to present experimental results in ways that are easy to understand. For example, state standard deviations, not variances. Be consistent in how you present results: for example, don’t switch back and forth between accuracy and error rate. Discuss (and, ideally, explain) all unanticipated results, for example much lower accuracy on some digits than on others.

Tables and figures show mappings, i.e. functions from x values to y values. Do not show mappings that are not semantically meaningful. Example 1: Do not show a mapping from j to the accuracy achieved on trial number j , where the only difference between trials is randomization. In this case, the ordering of the trials has no significance. Example 2: If you have a mapping from digit pairs $\langle x_1, x_2 \rangle$ to accuracies, do not impose a linear order on the set of digit pairs.

The information shown in a table or figure should all be meaningful. Do not show too few or too many digits of precision. Eliminate “chartjunk” such as colored backgrounds, boxes around charts, labels on every individual point, etc.