Crafting Papers on Machine Learning

Pat Langley LANGLEY@RTNA.DAIMLERCHRYSLER.COM
Adaptive Systems Group, DaimlerChrysler Research and Technology Center, 1510 Page Mill Road, Palo Alto, California 94304 USA

Abstract

This essay gives advice to authors of papers on machine learning, although much of it carries over to other computational disciplines. The issues covered include the material that should appear in a well-balanced paper, factors that arise in different approaches to evaluation, and ways to improve a submission's ability to communicate ideas to its readers.

1. Introduction

Although machine learning has become a scientific discipline, the effective communication of its ideas remains an art. Nevertheless, there exist rules of thumb even for practicing art, and in this essay we present some heuristics that we maintain can help machine learning authors improve their papers. Much of this advice applies equally well to other branches of artificial intelligence and even to scientific fields in general, but we will cast it in terms specific to our discipline.

Each section addresses a different facet of publications on machine learning. We first address the content appropriate for papers, considering briefly the topics that should appear in any scholarly work. After this, we discuss issues of evaluation at greater length, as they have come to play a central role in papers on machine learning. In closing, we give advice about matters of communication, ranging from high-level organization to the individual words used in the text. We hope that, taken together, these suggestions help authors to convey their ideas effectively to their audience.

2. Content of the Paper

A well-crafted paper on machine learning should cover a number of topics that communicate essential items to the reader. Different manuscripts may well organize this information in quite different ways, but the ideal paper should:

- *State the goals of the research* and the criteria by which readers should evaluate the approach. Categorize the paper in terms of some familiar class; e.g., a formal analysis, a description of some new learning algorithm, an application of established methods, or a computational model of human learning.
- *Specify the performance and learning tasks* that are the focus of the research, clearly distinguishing between the two aspects. If there is no performance system, propose some other means of evaluating the learning behavior.
- *Describe the representation and organization* of the system's knowledge, along with the representation of training data. Include examples of each in the paper, unless the approach is a standard one and thus familiar to most readers.
- *Explain both the performance and learning components* of the system in enough detail that readers can reimplement them (again, unless they are familiar to most readers). Ideally, use some metaphor (like search through a hypothesis space) to describe the learning algorithm.
- *Evaluate the approach to learning*, avoiding unsubstantiated or rhetorical claims. If stating that one approach is better than others, include evidence or at least careful arguments to support these claims. For example, present experimental or theoretical evidence of performance improvement, show successful accounts of psychological phenomena, or give evidence of new functionality.
- *Relate the approach to other methods*, discussing similarities, differences, and advances over previous research. Do more than simply list references to relevant work. Place the method in historical context and clearly specify intellectual debts, including work on the same task done within other paradigms.
- *State the limitations of the approach* and suggest directions for future research. Go beyond a list of problems to propose tentative solutions.

Of course, covering each of these will not ensure a high-quality paper, but omitting even one of them will weaken the manuscript and should be addressed before it is ready for publication.

3. Evaluation in Machine Learning

Evaluation has a central role to play in any publication on machine learning, but it is important to remember that many types of evaluation are possible. At the highest level, this can take any form that attempts to support the basic claims made by the author, but different sorts of claims can lead to distinct forms of research. Here we consider briefly the evidence appropriate to different types of evaluation.

3.1 Experimental Approaches to Evaluation

Certainly the most common approach to evaluation in machine learning relies on experimental studies. Many of the same issues arise here as in the natural sciences, including the need to identify clearly one's dependent measures and independent variables, the importance of careful experimental design, and the need to average across random variables outside one's control. These have become almost obvious features of a careful experimental investigation in our field.

Thus, a paper should state precisely the dependent variables in each study. Typically, these will be some measures of performance (i.e., behavior when learning is disabled), but other metrics, including characteristics of the learned knowledge, are also legitimate. However, as Provost, Fawcett, and Kohavi (1998) have argued, it is important that these variables make direct contact with the goals of the research. Using a measure like classification accuracy, despite its popularity, can be misguided for domains with skewed error costs or class distributions. In such cases, it may be better to invoke ROC curves, which report separately each type of error at different cost tradeoffs. Figure 1 shows such a curve for the task of rooftop detection in aerial images, taken from Maloof et al. (1998).

In a similar vein, an experimental report should state clearly the independent variables controlled in each study. Typical independent factors in research on supervised learning include the induction method – often some new algorithm being compared against more established ones – and the domain on which induction occurs – often data sets taken from the UCI repository. Most such studies aim to establish the new method as superior to existing techniques, which means they treat the domain as a random variable over which to average results, rather than interesting in its own right.

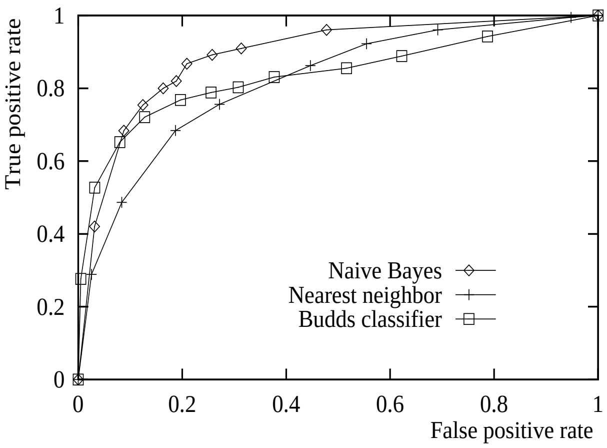

Figure 1. ROC curves for building detection in nadir images when trained and tested on different locations.

By themselves, such 'bake offs' tell one very little about the reasons why one method outperforms another, and thus do not provide the insight about causes that we expect in science. Insight is best obtained by running additional experiments on synthetic domains designed to test explicit hypotheses, typically motivated by the intuitions behind the original extension. The importance of using synthetic data sets is not because they provide new tasks, but because they let one vary systematically domain characteristics of interest, such as the number of relevant and irrelevant attributes, the amount of noise, and the complexity of the target concept. Thus, they let the researcher test hypotheses about each method's ability to scale under conditions of increasing difficulty.

Of course, insights about the sources of an algorithm's power are as important as insights about the effects of domain characteristics. Thus, a well-rounded experimental paper will also include lesion studies, which remove algorithm components to determine their contribution, and studies that examine sensitivity to specific parameter settings. Experiments that systematically vary external resources, such as the number of training cases available for learning, can also contribute important insights into an algorithm's behavior. Typical empirical papers report results on training sets of fixed size, which tells one nothing about how the methods would fare given more or less data, rather than collecting learning curves like those in Figure 1, taken from Langley and Sage (1999).

In recent years, the machine learning community has become increasingly concerned with statistical tests to establish that differences between observed experimen-

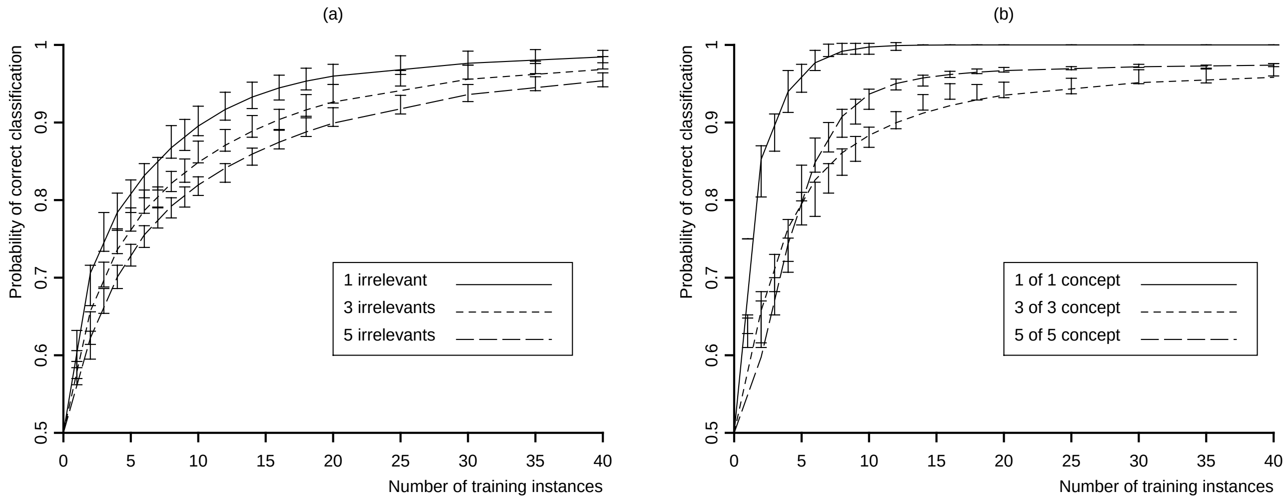

Figure 2. Theoretical and experimental learning curves, with 95% confidence intervals, for naive Bayes when (a) the domain involves a '2 of 2' target concept and varying numbers of irrelevant attributes, and (b) for a domain with one irrelevant attribute and a conjunctive concept with varying numbers of relevant features.

tal conditions are not accidental. Clearly, one should be careful not to draw unwarranted conclusions from experimental results. But it is even more important that these differences reveal some insight into the nature of the learning process, and we encourage authors to move beyond simple 'bake offs' to studies that give readers a deeper understanding of the reasons behind behavioral differences.

3.2 Alternative Forms of Evaluation

Although experimentation is the most popular style of evaluation in machine learning, it clearly is not the only valid approach. Perhaps the closest alternative involves the use of learning algorithms as models of human behavior (Langley, 1986). In this context, evaluation also involves running an algorithm, preferably many times and on different tasks, to determine its average behavior under various conditions. However, the goal is not for learning to improve performance as much as possible, but rather to improve it the same amount, under comparable conditions, as does human learning. Yet apart from this difference, the same issues arise as in experimental studies. Thus, the ideal evaluation of such a computational model will identify which components are most responsible for the ability to match human behavior and will examine the influence of domain characteristics on learning.

A third approach to evaluation revolves around the formal analysis of learning algorithms or tasks. Here the aim is to derive statements that, under precise conditions, relate aspects of learning behavior to characteristics of the learning task. For most such analyses, careful reading can determine whether the derivation or proof is correct, and thus whether the evidence supports the claim. However, there exist many true statements about learning that hold little intrinsic interest, making relevance to experimental findings an important factor. Also, some average-case analyses introduce approximations that require direct comparisons between predicted and observed behavior, as Figure 1 illustrates for an analysis by Langley and Sage (1999).

Certain claims are best backed by experimental evidence, comparison to human behavior, or formal analysis, but others require quite different types of support. For example, Dietterich (1990) has proposed criteria for *exploratory research* on machine learning. He maintains that papers on such work should identify, and state precisely, a new learning problem, show the inability of existing methods to solve this problem, propose novel approaches that show potential for solving it, discuss the important issues that arise in tackling this problem, and suggest an agenda for future research on the topic. Exploratory research, by its very nature, is not ready for careful experimental studies or final formal analyses, but it has an essential role to play in the field. Without such contributions, researchers would continue to spend their energies on minor variations of established tasks.

Another, related, approach to evaluation concerns the demonstration of new functionality. In this setting, the researcher claims that some new approach has capabilities not available to existing systems, which he then demonstrates by illustrating its ability to handle a number of challenging tasks. Such claims often occur in the context of systems that involve interaction of mechanisms not typically used together. Nordhausen and Langley (1993) present one such evaluation, in which they demonstrate that an integrated system for computational scientific discovery can handle tasks not accessible to any of its component algorithms.

Superficially, applications of machine learning appear to fall at the spectrum's other end, focusing on how one can use established methods to solve challenging problems that arise in the real world. However, as Provost and Kohavi (1998) note, a more common outcome is the identification of difficulties in applying these techniques, leading us to question assumptions made by basic researchers. For instance, problem reformulation, representation engineering, data manipulation, introduction of background knowledge, and dealing with error costs often play an important role in machine learning applications. The ideal applied paper examines their importance to the problem at hand, characterizes the key issues in more general terms, and challenges the research community to address those issues. The result is more akin to an exploratory research paper than one might expect.

Naturally, most publications in machine learning will focus on only one or two of these approaches to evaluation, but it seems equally clear that each such paradigm has an essential role to play in the field. The success of any given paper should be judged, not on which type of evaluation it embraces, but on the extent to which its evaluation provides evidence that supports its central claims.

4. Issues of Communication

The purpose of a scientific paper is to communicate ideas to the reader. To this end, you should craft your text to convey the key ideas to your audience clearly, so they can comprehend them with minimal effort.

4.1 Title and Abstract

Readers will remember your paper by its title. Thus, you should use a title that is informative but not overly long. If possible, describe different aspects of the research like the approach, domain, or factor of interest, as in "Genetic Induction of Planning Heuristics" or "Ensemble Methods in Noisy Domains". If you want to say more, add a brief subtitle, but be succinct.

The ideal abstract will be brief, limited to one paragraph and no more than six or seven sentences, to let readers scan it quickly for an overview of the paper's content. Do not repeat text from the abstract in your introduction; they should serve different purposes, with the former summarizing the text and the latter introducing the reader to the research.

4.2 Partitioning the Text

The organization of a paper into sections and paragraphs helps readers place a structure on the material and understand its contributions. Thus, you should put some effort into designing a good organization.

For instance, your paper will benefit from informative section and subsection headings, rather than generic ones like 'Representation' or 'Evaluation', as they let the reader browse through the paper and still have some idea what it is about. For the same reason, never use pronouns (such as 'it') in headings, and do not treat a section heading as if it were part of the text.

Include introductory sentences at the beginnings of sections and subsections to help readers make the transition. Make your sections roughly the same length, except possibly for the introduction and conclusion. Be consistent about whether you include an introductory paragraph before the first subsection. Also, never include only one subsection in a section, since subsections are designed to divide a section into components. For the same reasons, avoid subsections that contain only one paragraph; if you have only one paragraph's worth of material, embed it in another subsection.

Within each section or subsection, you should further partition the paper into paragraphs, each of which should discuss a distinct idea and flow naturally from its predecessor. The ideal paragraph will run no more than six sentences and no fewer than three sentences. Neither should the sentences themselves say too much or too little; rather, they should convey ideas in bites the reader can digest.

On occasion, you may want to use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. For the sake of readability, footnotes should take the form of complete sentences.

4.3 Continuity and Flow

In a well-written paper, each part of the text seems to flow naturally from the material that precedes it. We have already mentioned the need for transition sentences at the outset of sections, but you can take other steps to improve the continuity of your paper.

One important factor is that the text should treat conceptually distinct topics separately and in their proper

[1] However, keep your footnotes to a reasonable length, say one to three sentences.

order. For example, you should not talk about the performance element or learning mechanism while discussing representational issues. In general, ensure that earlier text lays the groundwork for what comes later.

Itemization can highlight important points or steps, but make sure the list improves clarity rather than reduces it. Be careful not to overuse itemizations; often a paragraph with the same material will communicate as well. You should also itemize at the right level, giving neither too much nor too little detail; ideal items are shorter than paragraphs but more than a few words. Close off each list with a concluding sentence.

Similarly, parenthetical expressions are useful for making side comments, but be wary of overusing them, as anything longer than a few words will upset the sentence's flow. In such cases, place the information in a footnote instead.

Readers usually understand active sentences more easily than passive ones, so use active constructions whenever possible. This is easier to achieve by writing in the first person or by using the system name for a sentence's subject. For instance, "ID5 constructs decision trees incrementally ..." is better than "Decision trees are constructed incrementally ...".

At the sentence level, you should avoid long chains of adjectives, such as "incremental instance-based learning algorithms". Instead, break such chains into more manageable chunks, as in "incremental algorithms for instance-based learning". Also, avoid using contractions in the text, since such expressions sound overly chatty in a technical paper.

4.4 Figures and Tables

You may want to include figures in the paper to help readers visualize your approach and your results. Make sure you label all distinct components of each such figure. For example, Figure 1 assigns a letter to each graph, gives labels for each axis, and includes a legend that briefly describes each curve.

Below each figure, include a caption with enough detail to give readers an idea of the contents without reading the text. For instance, "Improvement in classification accuracy for three induction methods on the congressional voting domain" is better than "Learning curves on the voting domain" or "Behavior of three induction methods". However, do *not* include a title above the figure, as the caption already serves this function.

You may also want to include tables which summarize textual material that can be typeset, in contrast to figures, which contain graphical material that must be drawn. As in figures, label all distinct components clearly. For example, if the form is tabular, then specify the contents of each row and column in the topmost row. Above each table, include a title that briefly explains the content to readers.

Ensure that you refer to each table (and figure) in the text and that you discuss them, at least briefly. Their purpose is to augment the text, not to replace it. In such discussions, do not refer to their location, as in "the table below", since their exact position may change during typesetting.

4.5 Describing Your System

Many papers in machine learning center around a new system or algorithm, and clear descriptions of this method are crucial to a paper's success. Thus, you should put careful thought into communicating the essential features of its operation.

Naming your system will give your text more variety, and it will let other authors refer to something concrete in their review of your work. However, do not overuse this label; instead, alternate between using the system name and an equivalent term, like "the system". If the system name appears more than three times in one paragraph, you should remove some occurrences. Also, never end one sentence and then start the following one with the system's name.

In addition, you should avoid language-specific terms and formalisms when describing your system, as many readers will not know your implementation language. Reformat representations that involve list structures to make them more readable. And when referring to subroutines or system parameters, use mnemonic names rather than internal system names, so that your description does not read like a core dump.

Also remember that, although detailed program traces can be helpful, they are not a replacement for a careful system description. If you do include one, make sure you paraphrase it in English and include running commentary. Consider placing the trace in an appendix where it will not hurt the flow of the paper.

4.6 Terminology and Notation

Successful communication begins at the level of individual words, so the choices you make here also influence the readability of your paper. Remember that precision is not enough; your audience must be able to recall what they have read.

One important step is avoid abbreviations, especially if you invoke them only a few times. Even if you use an abbreviation repeatedly, you should not use more than a few distinct abbreviations in a given paper. If your goal is to save keystrokes, you can use other means, such as defining a macro or making a global substitution. And you should never include an abbreviation in a title or heading, however often it appears elsewhere.

For similar reasons, you should avoid needless jargon. Whenever possible, use terminology shared by other researchers in the field rather than inventing your own. If you must coin new terms, explain their relation to existing concepts. Even borrowing legitimate but unfamiliar terms from another field can make a paper very difficult for readers who are unacquainted with that area. More generally, avoid terms that lend themselves to confusion, especially when other words would serve equally well. Think carefully about the specialized terms that you employ.

Finally, you should omit unnecessary formalism that does not occur in proofs or otherwise aid communication. If you do decide to use formal notation, make sure to clarify its meaning in the text. Even readers with mathematical sophistication will appreciate the effort. Your goal is to convey ideas and evidence to the audience, not to overwhelm them with your arcane language and formal prowess.

5. Concluding Remarks

As in other sciences, research in machine learning is a complex endeavor that includes identifying new problems, developing new frameworks and methods, evaluating those approaches, and conveying the results to other scientists. Writing and publishing papers is only one stage in this process, but one that must put all the previous steps into an integrated, comprehensible package. Nor does it constitute the final step, since good papers attract the attention of readers and foster additional research.

In closing, remember that successful communication is central to the scientific process, and that few readers – including reviewers and editors – are willing to wade through difficult text. Spending an extra hour or two making your paper clear and easy to read can save many more hours across the entire research community, as well as increase the paper's chances of publication and influencing your colleagues.

Acknowledgements

Thanks to Claude-Nicolas Fiechter, Mehmet Göker, Stephanie Sage, and Cynthia Thompson for comments that improved the essay. Portions of this document are borrowed from earlier editorials in *Machine Learning* (Langley, 1990) and *IEEE Expert* (Langley, 1996).

References

Dietterich, T. G., (1990). Exploratory research in machine learning. *Machine Learning*, *5*, 5–10.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338–345). San Francisco: Morgan Kaufmann.

Maloof, M. A., Langley, P., Binford, T. O., & Nevatia, R. (1998). Generalizing over aspect and location for rooftop detection. *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision* (pp. 194–199). Princeton, NJ: IEEE Press.

Nordhausen, B., & Langley, P. (1993). An integrated framework for empirical discovery. Machine Learning, 12, 17-47.

Langley, P. (1986). Human and machine learning. *Machine Learning*, *1*, 243–248.

Langley, P. (1990). Advice to authors of machine learning papers. *Machine Learning*, *5*, 233–237.

Langley, P. (October, 1996). Relevance and insight in experimental studies. *IEEE Expert*, 11–12.

Langley, P., & Sage, S. (1999). Tractable average-case analysis of naive Bayesian classifiers. *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 220–228). San Francisco: Morgan Kaufmann.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). San Francisco: Morgan Kaufmann.

Provost, F., & Kohavi, R. (1998). On applied research in machine learning. *Machine Learning*, *30*, 127–132.