A Book Proposal
The Practical Guide to Machine Learning, Data Mining, and Computer Vision

What the book is good for

There have been many improvements in our ability to make computers understand large and complicated data sets over the past fifteen years. Most, if not all, of these strides have been made in academia. There is still a large gap between the potential for the use of machine learning in industry and its current level of play. The goal of this book is to close that gap by educating computer hackers about modern machine learning techniques.

(The terms 'machine learning' and 'data mining' usually refer to the same thing. Academics tend the use the former term, while people in industry use the latter. I prefer the term 'machine learning,' since its open ended nature encourages more creativity. ML is used as an abbreviation for 'machine learning.')

The machine learning techniques described in the book will allow hackers to infer predictions about the world based on a training data set. For example, the tools described might be used to label an incoming email as spam or ham.

The book will also offer a surface-level treatment of computer vision. The machine learning and computer vision camps do not use the same books or go to the same conferences, but I think that computer vision will become increasingly useful to hackers as pictures and videos become more prolific in day-to-day computing. Fortunately, many of the ML concepts and techniques apply to computer vision; I believe that I can touch on computer vision without getting side-tracked or diluting the usefulness of the book.

The book is targeted to those interested in leveraging machine learning in the real world. I have focused on two subsets of this audience while planning the contents and layout – hackers and technology consultants. The computer hacker represents the skilled and inquisitive reader, while the consultant speaks for the practical and results driven reader. Both of these perspectives are necessary; ignoring intellectual curiosity prohibits having fun in going over the necessary details, while ignoring the results blurs the book's ultimate purpose.

The stated goal of the book is to close the gap between potential and actual usage of ML. If describing machine learning techniques is the primary strategy, then the secondary strategy is to convince people that ML has a large amount of potential to create value. I want to ~~convince~~ demonstrate to readers and their friends that machine learning will be critical to the next generation of innovation. This demonstration will be done through the third part of the book which will contain case studies of ML in action. The studies will be down to earth, fun, and interesting. For example, one study will focus on an algorithm to buy and sell concert and sporting event tickets on EBay so as to make a profit, playing EBay like a stock market. Aside from demonstrating the utility of ML, the studies will help educate the reader and help us market the book.

The market

Machine learning is becoming more popular. People in vertical markets of all kinds are starting to understand ML's power to solve their estimation and prediction problems. For example, ML is being used to detect credit card fraud, breast cancer, and to calculate insurance rates. ML is also beginning to appear in broadly consumer markets. For example, Riya is a startup company whose product supports photo searching based on facial recognition; they are currently running invite-only alpha testing.

The usual story that machine learning books open with is about the amassing of data. Companies and consumers are getting cheaper access to storage, which has encouraged the digital equivalent of pack rat behavior. ML advances have started to offer leverage on the problem of extracting intelligence from these data stores. Rutherford Roger is quoted as saying "We are drowning in information and starving for knowledge" in the beginning of Hastie, et al (a popular statistical learning book).

Even though the popularity of ML has been increasing, the democratization process is just starting. For example, my father buys and sells used heavy truck equipment; he and his partners have in-house telephone operators who call dealers to ask if they want to do any deals. They currently go down the list alphabetically, even though it seems likely that there are predictors in the data waiting to indicate which phone numbers are the hottest prospects. There are countless similar situations in the world today; the adoption curve is on the upswing.

Other Books

I searched Amazon for all of the machine learning and data mining books that I could find. The resulting list contains 54 books.

Many of these books are narrowly focused on one machine learning technique, such as neural networks or support vector machines.

Some significant fraction of the books out there do not contain any pseudocode; these are usually the ones that are more academically or mathematically focused. For example, the main textbook for my MIT Machine Learning course (6.867), *The Elements of Statistical Learning* by Hastie, et al., contains little to none pseudocode.

One of the top two books that covers ML, by Russell and Norvig, is the canonical book for college AI classes. If Cormen, Leiserson, et al. is the standard for algorithm courses, this book is the equivalent in AI. Although a few sections focus on ML, the book itself covers a very large variety of AI-related topics.

The other of the top two books, by Witten and Frank, is more practically focused. The authors of the book are staff at the University of Waikato in New Zealand, and work on Weka, a popular open source ML software package. The sales and popularity of the book seem to be driven by Weka. The book has two parts – one describes machine learning algorithms, and the other describes Weka and how to use it. The first section of the book seems disorganized and confusing from my perspective as a ML enthusiast. For example, decision trees are the topic of discussion in sections 3.2 ("Decision trees"), 4.3 ("Divide-and-conquer: Constructing decision trees"), 6.1 ("Decision trees"), *and* 7.4 ("Improving decision trees"). The book should have been organized by technique instead of by step. I can follow the book as someone who knows the ML landscape, but I am worried about the non-sequential reader who is not familiar with the enterprise.

The Outline

The book will have an introduction and three parts. The first part will deliver the big picture; it will be a short overview of the ML world, our goals, and a two page description of each ML technique and concept. The second part will grow the ML toolbox one tool at a time – introducing it, describing it, discussing its pros/cons, and providing simple examples. The third part will leverage the reader's newly developed ML toolbox against interesting and cool problems that everybody can relate to. (As opposed to, for example, the use of ML to identify DNA sequences of protein binding sites.) We will now dive deeper into each of these parts.

Introduction
The introduction will bring the reader into the frame of mind of a ML person. This frame of mind calls for an understanding of what our goals are, what assumptions are made throughout the book, and what the basic trade-offs are. The introduction will also give a roadmap for the book, as simple as it will be.

Finally, logistics about the book's website will be provided. I plan on providing keywords in the book that can be used to find material on the web site. For example, "neuronvideo" might link to a video of neuron activity in the brain of a mouse.

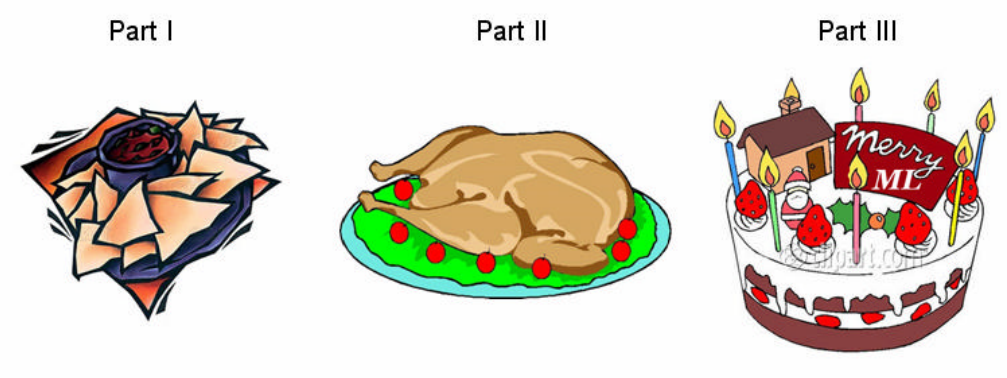

Figure 1. An overview of the three parts of the book.

Part I – Overview of our enterprise
Part one of the book is the appetizer; it will give the reader a taste for our enterprise. This will be done through two page summaries of the classes of tools, the tools themselves, and the trade-offs that we'll be dealing with.

The first chapter will address the classes of tools. It will contain two pages of discussion on each of the following topics, describing what it is, when it should and should not be used, and what the tools are that implement it.

- Supervised learning
- Clustering
- Filtering
- Dimensionality reduction / feature selection
- Filling in missing values
- Working with large data sets and programs
- Model and parameter selection
- Multi-modal estimation
- Feature extraction in images
- Object recognition in images
- Object tracking in video

For example, the goal of supervised learning is to predict an output given some inputs and a training set of input-output pairs. There are two kinds of supervised learning – classification and regression, (…). Tools that perform supervised learning are decision trees, neural networks, support vector machines, etc.

The second chapter will address individual tools. One or two page descriptions will be given for each tool or technique, describing the idea behind it, what it's useful for, and what gotchas to look out for. Each of the following tools will be discussed.

- Nearest neighbor

- Linear Discriminate Analysis
- Decision trees
- Bayes nets
- Naïve bayes
- Linear separators / perceptrons
- Support vector machines
- Regressions
- Neural networks
- K-Means clustering
- (a few other clustering techniques)
- Kernel density estimation
- Signal filtering
- Principle Component Analysis
- (maybe one other dimensionality reduction technique)
- Expectation maximization (for filling in missing values)
- Cross-validation

The third chapter will give one or two page summaries of the trade-offs and concepts that come up throughout the book. These will include:

- Parametric v.s. non-parametric learning
- Classification v.s. regression
- Bias v.s. variance
- Data sets
- Magic knobs

Part II – The Classroom

This part is the meat and potatoes of the book. It will describe each of the individual tools listed above, providing background material, derivations (where appropriate), and examples.

The point is to organize each chapter to be stand-alone, in that if the reader has read part one or is familiar with the ML landscape then they can read any chapter representing a particular technique to understand and leverage that tool. This ideal won't be completely achievable, but it is the goal.

Although most of the chapters in part two will follow the same structure (describe problem, then describe the solution), the "working with large data sets and large ML programs" chapter has some substructure. Reflecting on my experience with large data sets and programs, I made a list of the following hints, each of which will be discussed in sufficient detail.

- Theme: levels of abstractions, modularity, and orders of magnitudes
- When to use files versus databases
- Separating data from programs
- Automation and scripting
- Validate often
- Fail fast
- Interfacing to third party software
- Parsing and text operations
- Use stream operations

Part III – The Garage

Although I do not want to play favorites, part three will probably turn out to be the most fun to read; it is the dessert. It will contain case studies of ML applied to easy to understand and interesting problems. I want to keep this part limited to four of five case studies. The current forerunners are introduced below.

Each of these case studies will describe the problem, how it was approached, and how it was solved. All of the code generated in the process will be available for download from the book's website. The big win of this part is that these are real applications. They are meant to inspire readers to think creatively about how to apply ML to problems in the world. They also highlight the entire lifecycle of a project, including gathering data from the Internet and managing large amounts of processing steps.

1) *EBay Marketeer*

   The idea is to build a program that learns how to predict the sales price for sporting event and/or concert tickets. Given this, if a ticket is going for less than fair market value on EBay, then this program will suggest that the user buys it and explains how to sell it later to maximize returns.

   Intuition suggests that there are two cases in which this strategy would work. First, there might be something wrong with the auction; for example, auctions ending in the middle of the night probably tend to sell at lower prices. The second reason is that certain tickets might appreciate in value over time. The data might suggest that we should buy the ticket that's for a concert four months away, and sell the ticket three months later.

   EBay has an API that will allow us to collect the appropriate data, make bids, etc.

2) *Collaborative Antispam*

   Spam filtering today is done using naïve bayes classifiers on user-separated training data. The user marks some messages as spam, others as ham, and the algorithm tries to learn a classifier based on that data. The implementation of this approach is straightforward, especially after learning about naïve bayes in part two. I will suggest that the interested reader see Paul Grahams' essays on the subject of naïve bayes filtering applied to spam.

   This chapter will focus on a more sophisticated filter. The idea is that if you can look at everyone's inboxes then it would be straightforward to identify spam, using the following rule. If a very large number of people get the same (or similar) message but do not normally get the same ham (not spam) messages, then that message is probably spam.

   There are many implementation issues, including privacy problems, defining similarity, and getting enough people to use it.

   Because this idea will not be fully launched, the book will not be able to describe the empirical results of a full scale implementation. It might be possible, however, to test the idea on a user base of around 100 people, which should indicate how successful the idea is.

3) *Parking Spot Detection*

   I initiated a project at MIT called CarTel; the idea was to put computers in cars so that they could infer traffic conditions based on each cars' current position and speed. Shortly thereafter I left the project to do my own startup in this area. Although the startup later

failed, CarTel is still around. There is a high likelihood that I will work on the CarTel project for my Masters of Engineering thesis, trying to leverage the units to do parking spot detection.

The idea is that if you have cheap cameras looking out the front window of a car then you can look for open street side parking spots. The locations of legal parking spots can be inferred based on where you've seen people parked before. If a device sees an open spot then it sends a message (wirelessly) to a central computer where others can access the information.

If this works out then I will adapt my thesis to be this chapter in the book.

4) *Facebook Friend Predictor*

Facebook.com is a college social networking site where individuals maintain profiles and link in with their friends. I have crawled the web site for friendship data, and am building a friend predictor for my machine learning class at MIT. The friend predictor will indicate people who you are not yet friends with on the site, but with whom you are statistically likely to be friends with.

Writing Samples

I am including six writing samples.

- *Metrics for Counter-insurgencies* is a paper that describes how to measure the level of success (or lack thereof) of a counter-insurgency effort. It is an example of analyzing a real-world, complex situation using rigorous methods.
- *BED Device Data Bus* describes a data bus, similar in nature to USB. The Beta processor (a generic 32-bit CPU) is the target platform and BED (Beta Expansion Devices) are the plug-in hardware devices.
- *Tracking Moving Devices with the Cricket Location System* is a conference paper (Mobisys 2004) that describes my research in position estimation. Others contributed to the writing, although I am the lead author.
- *Distribution of Dynamic Binary Data Using ASP* is a short technical white paper I wrote three years ago.
- *Transit Advisor – Final Report* describes a system for finding the optimal transport route involving the Boston T, taxi rides, ZipCar, biking, and walking.
- *Swift Ride – Request for Proposal* is an RFP I wrote at my startup seeking an ODM (Original Design Manufacturer) for our hardware needs.

Tools

I will use MS Word, and miscellaneous tools to develop graphics (usually MS PowerPoint and Paint Shop Pro).

Who am I?

Although I am attaching my resume, it only speaks indirectly to my ML expertise.

Academically, I tested out of the undergraduate AI course at MIT. I have taken three graduate level AI classes: Knowledge Based Systems, Techniques in AI, and Machine Learning. I have a passion for ML outside of my classes, though. This summer I spent large amounts of time (outside of work) hacking on a game theoretic computer poker player. This project called for manipulation of gigabyte-sized files and over twelve thousand lines of code. Thus, I have been around academia to know what is going on in the world of ML, but I also have empathy for the real world mechanics of taking an idea and running with it.

Appendix – List of ML Related Books

Amazon Sales Ranking	Book Title
6566	Artificial Intelligence: A Modern Approach (2nd Edition) (Hardcover)
8793	Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)
14653	Pattern Classification (2nd Edition)
15707	Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)
16506	Statistical Inference (Casella)
20935	Information Theory, Inference & Learning Algorithms
24545	Machine Learning
30591	Bayesian Data Analysis, Second Edition
31025	Managing Gigabytes: Compressing and Indexing Documents and Images (The Morgan Kaufmann Series in Multimedia and Information Systems)
34291	Statistical Decision Theory and Bayesian Analysis (Springer Series in Statistics)
36563	An Introduction to Support Vector Machines and Other Kernel-based Learning Methods
41702	Kernel Methods for Pattern Analysis
46102	Bayesian Statistics: An Introduction (Arnold Publication)
48197	Introduction to Machine Learning (Adaptive Computation and Machine Learning)
49669	Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization (Natural Language Processing, 5)
55750	Introduction to Data Mining, (First Edition) (Hardcover)
58098	Neural Networks for Pattern Recognition
59402	Classification and Regression Trees
59503	Mining the Web: Analysis of Hypertext and Semi Structured Data (The Morgan Kaufmann Series in Data Management Systems)
76406	Learning Bayesian Networks (Hardcover)
90641	Data Analysis: A Bayesian Tutorial (Oxford Science Publications)
91240	Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (Hardcover)
97028	Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management (Paperback)
98393	Bayesian Theory (Wiley Series in Probability and Statistics)
102104	Survey of Text Mining : Clustering, Classification, and Retrieval
108077	Markov Chain Monte Carlo in Practice
112445	Data Preparation for Data Mining (The Morgan Kaufmann Series in Data Management Systems) (Paperback)
114077	Applied Data Mining : Statistical Methods for Business and Industry (Statistics in Practice) (Paperback)
136303	Mathematical Methods in Artificial Intelligence (Practitioners)
155004	Genetic Algorithms in Search, Optimization, and Machine Learning
159487	Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management (Paperback)
162037	Bayes and Empirical Bayes Methods for Data Analysis, Second Edition
165903	Learning Kernel Classifiers: Theory and Algorithms (Adaptive Computation and Machine Learning)
205762	Model Selection and Multi-Model Inference
215897	Fundamentals of Neural Networks
216844	Pattern Recognition and Neural Networks
219346	Pattern Recognition, Second Edition

222783	Business Modeling and Data Mining (The Morgan Kaufmann Series in Data Management Systems) (Paperback)
225222	Neural Networks: A Comprehensive Foundation (2nd Edition)
228089	Principles of Data Mining (Adaptive Computation and Machine Learning) (Hardcover)
235943	Bayesian Statistical Modelling (Wiley Series in Probability and Statistics - Applied Probability and Statistics Section)
304316	An Introduction to Computational Learning Theory
342312	Learning in Graphical Models (Adaptive Computation and Machine Learning) (Paperback)
362619	Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models (Complex Adaptive Systems)
414654	An Introduction to Bayesian Inference and Decision, Second Edition
471842	Applied Bayesian Modelling (Wiley Series in Probability and Statistics)
472049	Introduction to Statistical Pattern Recognition (Computer Science and Scientific Computing Series)
512833	Neural and Adaptive Systems: Fundamentals through Simulations
535677	Machine Learning : An Artificial Intelligence Approach (Volume I) (Machine Learning) (Hardcover)
555763	Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference
818207	Advanced Algorithms for Neural Networks: A C++ Sourcebook
989096	Pattern Recognition From Classical to Modern Approaches (Hardcover)
1500383	Feedforward Neural Network Methodology (Springer Series in Statistics)
2173972	Artificial Intelligence and Neural Networks: Steps Toward Principled Integration (Neural Networks, Foundations to Applications)
6566	Artificial Intelligence: A Modern Approach (2nd Edition) (Hardcover)
8793	Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)
14653	Pattern Classification (2nd Edition)