Some notes and solutions to Tom Mitchell's *Machine Learning* (McGraw Hill, 1997)

Peter Danenberg

13 October 2011

Contents

1 DONE 1.1

`CLOSED:` *2011-10-12 Wed 04:21*

Appropriate **animal languages** could craft appropriate responses and prompts, perhaps, though ignorant of the semantics.

fugues train on bach data, or buxtehude. performance measure? perfect authentic cadence, of course. ;) no, not that simple.

narratives learn the structure of narratives? performance measure is tricky here.

Not appropriate **comedy** requires a bizarre ex nihilo and sponteneity (distinguishable from three above?) in fact, the second and third above are inappropriate, rather? define "inappropriate": difficult? vague performance measure?

data representation and search or have meta-learning-problems been solved?

new science and mathematics can "creativity" be modelled?

So we can't, indeed, escape the question of modelling; once the mechanics of learning have been mastered, there lies the ex nihilo.

2 DONE 1.2

`CLOSED:` *2011-10-12 Wed 04:21*

Learning task: produces melodic answers to query phrases. Given a phrase that ends on a dominant, say, within a key; gives an appropriate response that ends on the tonic. Must follow a constrained set of progressions (subdominant to dominant, dominant to tonic, flat-six to neopolitan, etc.), and be of an appropriate length.

task T constructing answering phrases to musical prompts (chords)

performance measure P percent of answers that return to the dominant once at the end (given appropriate length and progression constraints)

training experience E expert (bach, chopin, beethoven) prompts and answers.

target function V : progression $\rightarrow \Re$; $V(b = \text{final tonic}) = 100$, $V(b = \text{final non-tonic}) = -100$.

target function representation $\hat{V}(b) = w_0 + w_1 x_1$, where $x_1 =$ length of prompt$-$ number of chords in answer

3 DONE 1.3

`CLOSED:` *2011-10-12 Wed 12:46*

Here's a solution for the trivial case in which $|\langle b, V_{\text{train}}(b)\rangle| = 1$ and the target function $\hat{V}$ consists of a single feature x and a single weight w:

$$\frac{\partial E}{\partial w} = \frac{\partial}{\partial w}(V_{\text{train}}(b) - \hat{V}(b))^2 \tag{1}$$

$$= 2(V_{\text{train}}(b) - \hat{V}(b))\frac{\partial}{\partial w}(V_{\text{train}}(b) - \hat{V}(b)) \tag{2}$$

$$= 2(V_{\text{train}}(b) - \hat{V}(b))(0 - x) \tag{3}$$

$$= -2(V_{\text{train}}(b) - \hat{V}(b))x \tag{4}$$

which gives:

$$w_{n+1} = w_n - \frac{\partial E}{\partial w} \tag{5}$$

$$\propto w_n + \eta(V_{\text{train}}(b) - \hat{V}(b))x \qquad \text{(by 4)} \tag{6}$$

It should be trivial to extend this to the case where $\mathbf{w}$ and $\mathbf{X}$ are vectors; the LMS training rule, furthermore, covers the summation.

3.1 Notes

From page 11: "The LMS training rule can be viewed as performing a stochastic gradient-descent search through the space of possible hypotheses (weight values) to minimize the squared error E."

3.1.1 Gradient descent

- Gradient descent is a first-order optimization algorithm. To find a local minimum of a function . . . one takes steps proportional to the negative of the gradient of the function at the current point.
  - If one takes steps proportional to the positive of the gradient, one approaches a local maximum: gradient ascent.
- Known as steepest descent.
- If $F(x)$ is defined and differentiable in a neighborhood of point a, $F(x)$ decreases fastest if one goes from a in the direction of the negative gradient of F at a, $-\bigtriangledown F(a)$.
- If $b = a - \gamma \bigtriangledown F(a)$ for $\gamma > 0$, then $F(a) \geq F(b)$.
- One starts with a guess x_0 for a local minimum of F, and considers the sequence $x_0, x_1, \ldots$ such that $x_{n+1} = x_n - \gamma_n \bigtriangledown F(x_n), n \geq 0$.
- We have $F(x_0) \geq F(x_1) \geq \cdots$.
- Gradient descent can be used to solve a system of linear equations, reformulated as a quadratic minimization problem, e.g., using linear least squares.
- Convergence can be made faster by using an adaptive step size.

4 DONE 1.4

`CLOSED:` *2011-10-12 Wed 18:21*

"Generating random legal board positions" is an interesting sampling strategy that might expose the performance system to improbable positions that

form the "long tail" of possible board positions; the resultant hypothesis might not be optimized for common positions, though.

"Generating a position by picking a previous game and applying one of the moves not executed" is a good exhaustive search strategy which may or may not be feasible, depending on the state-space complexity of the game.

One mechanism might be to generate positions from expert endgames in reverse: you're starting with a pruned search space such that, by the time you're generating from openings, you might already have a reasonably good hypothesis. If e.g. world championship games share characteristics with expert games, this may be a reasonable heuristic for competitions. On the other hand, the hypothesis would be vulnerable to paradigmatic shifts in expert play.

If an exhaustive search of the state-space in infeasible and training examples were held constant, I'd bet on reverse-search of an expert-corpus over random sampling; especially if, *vide supra*, championship and expert play share certain characteristics.

4.1 Notes

$$\begin{aligned}\text{Training} &\xrightarrow{\text{Games against self}} V \\ &\xrightarrow{\text{Board}\to\text{value}} \text{Representation} \\ &\xrightarrow{\text{Linear function}} \text{Algorithm} \\ &\xrightarrow{\text{Gradient descent}} \text{Design}\end{aligned}$$

Figure 1: Summary of design

Experiment generator Take as input the current hypothesis and output a new problem for the performance system to explore. Our experiment generator always proposes the same initial board game. More sophisticated strategies could involve creating board positions designed to explore particular regions of the state space.

5 TODO 1.5

5.1 Notes

"Non-operational" definition of $V(b)$:

$$V(b) = \begin{cases} 100 & b \text{ is a final winning board state} \\ -100 & b \text{ is a final losing board state} \\ 0 & b \text{ is a final drawing board state} \\ V(b') & \text{otherwise, where } b' \text{ is an optimal final board state} \end{cases} \tag{7}$$

of which the operational approximation is $\hat{V}(b)$.

Whereas for checkers:

x_1 black pieces

x_2 red pieces

x_3 black kings

x_4 red kings

x_5 black pieces threatened

x_6 red pieces threatened

For tic-tac-toe, maybe we can use the following features as a starting hypothesis:

x_1 number of Xs

x_2 number of Os

x_3 number of potential 3s for X

x_4 number of potential 3s for O

It covers forks, doesn't it? Or should we explicitly enumerate it?

Maybe, on the other hand, number of Xs and Os doesn't matter; since they increase perforce. A full enumeration of features will probably slow down analysis, won't it? Maybe that's the tradeoff: speed for precision.

x_1 X 3-in-a-row?

x_2 O 3-in-a-row?

x_3 X fork?

x_4 O fork?

x_5 X center?

x_6 O center?

x_7 X opposite corner?

x_8 O opposite corner?

x_9 X empty corner?

x_10 O empty corner?

x_11 X empty side?

x_12 O empty side?

Page 8: "In general, this choice of representation involves a crucial tradeoff. On one hand, we wish to pick a very expressive representation to allow representing as close an approximation as possible to the ideal target function V. On the other hand, the more expressive the representation, tho more training data the program will require in order to choose among the alternative hypotheses it can represent."

Here's a crazy thought: since the space-state complexity of tic-tac-toe is utterly tractable, let's have nine features: one corresponding to each of the squares.

How do we deal with training the opposite direction, by the way: invert the outcome of the training data?

I have no idea how much training data nine variables need: we'll have to plot it; interesting to compare a strategy containing e.g. forks and wins.

Is it interesting that each variable is binary?

Let's start with the generalizer and a catalog of games; in order to map the number of training-examples . . . Ah, I see: the second player has a fixed evaluation function. Can we abstract xkcd? Problem is, the space for O is much more complicated. Maybe we can abstract the Wikipedia strategy:

1. Win
2. Block
3. Fork
4. Block a fork
5. Center
6. Opposite corner
7. Empty side

(It looks like the Wikipedia strategy was abstracted from here, by the way; damn: it looks like there are separate X- and O-heuristics.)

Represent the board as a vector of nine values; can we set up abstractions for $< x, y >$ as well as `{map,reduce,for-each}-{row,column,diagonal,triplet}`?

Meh; maybe we can implement the X/O-agnostic heuristics.

6 Notes

6.1 1

- a computer program is said to learn from experienc E with respect to some class of tasks T and performance measure P, if its performanc at tasks in T, as measured by P, improves with experience E.
- neural network, hidden markov models, decision tree

artificial intelligence	symbolic representations of concepts
bayesian	estimating values of unobserved variables
statistics	characterization of errors, confidence intervals
statistics	attributes of training experience:

- type of training experience from which our system will learn
  - direct or indirect feedback

    direct individual checkers board states and the correct move for each

    indirect move sequences, final outcomes
    - credit assignment: game can be lost even when early moves are optimal
- degree to which learner controls sequence of training examples
- how well it represents the distribution of examples over which the final system performance P must be measured
  - mastery of one distribution of examples will not necessary (sic) lead to strong performance over some other distribution

statistics task T: playing checkers; performance measure P: percent of games won; training experience E: games played against itself.

statistics 1. the exactly type of knowledge to be learned; 2. a representation for this target knowledge; 3. a learning mechanism.

statistics program: generate legal moves: needs to learn how to choose the best move; some large search space

statistics class for which the legal moves that define some large search space are known a priori, but for which the best search strategy is not known

target function choosemove : B -> M (some B from legal board states to some M from legal moves)

- very difficult to learn given the kind of indirect training experience available
- alternative target function: assigns a numerical score to any given board state

alternative target function V : B -> R (V maps legal board state B to some real value)

- higher scores to better board states

alternative target function V(b = finally won) = 100

alternative target function V(b = finally lost) = -100

alternative target function V(b = finally drawn) = 0

alternative target function else V(b) = V(b') where b' is the best final board state starting from b and playing optimally until the end of the game (assuming the oppont plays optimally, as well).

– red black trees? greedy optimization?

alternative target function: this definition is not efficiently computable; requires searching ahead to end of game.

alternative target function: *nonoperational* definition

alternative target function: goal: *operational* definition

alternative target function: *function approximation*: $\hat{V}$ (distinguished from ideal target function V)

alternative target function: the more expressive the representation, the more training data program will require to choose among alternative hypotheses

alternative target function: $\hat{V}$ linear combination of following board features:

- $\mathbf{x}_1$ number of black pieces
- $\mathbf{x}_2$ number of red pieces
- $\mathbf{x}_3$ number of black kings
- $\mathbf{x}_4$ number of red kings
- $\mathbf{x}_5$ number of black pieces threatened by red
- $\mathbf{x}_6$ number of red pieces threatened by black

alternative target function: $\hat{V} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6$

alternative target function: $\mathrm{w}_0 \ldots \mathrm{w}_6$ are weights chosen by the learning algorithm

alternative target function: partial design, learning program:

- T playing checkers
- P percent games won
- E games played against self

target function: $\mathrm{V} : \mathrm{Board} \rightarrow \Re$

target function representation: $\hat{V} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6$

target function representation: last two: design choices

alternative target function: require set of training examples, describing board state b and training value $\mathrm{V}_{\mathrm{train}}(\mathrm{b})$ for b: ordered pair $\langle b, V_{train}(b) \rangle$: $\langle \langle x_1 = 3, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 0 \rangle, +100 \rangle$.

alternative target function: less obvious how ta assign training values to the more numerous intermediate board states

alternative target function: $V_{train}(b) \leftarrow \hat{V}(Successor(b))$

alternative target function: $Successor(b)$ denotes the next board state following b for which it is again the program's turn to move

– train separately red and black

alternative target function $\hat{V}$ tends to be more accurate forboard states closer to game's end

alternative target function best fit: define the best hypothesis, or set of weights, as that which minimizes the squared error E between the training values and the values predicted by the hypothesis $\hat{V}$

$E \equiv \Sigma_{\langle b, V_{train}(b)\rangle \in training\ examples}(V_train(b) - \hat{V}(b))^2$

in statistics and signal processing, a minimum mean square error (MMSE) estimator describes the approach which minimizes the mean square error (MSE), which is a common measure of estimator quality.

the term MMSE specifically refers to estimation in a bayesian setting, since in the alternative frequentist setting there does not exist a single estimator having minimal MSE.

let X be an unknown random variable, and let Y be a known random variable (the measurement). an estimator $\hat{X}(y)$ is any function of the measurement Y, and its MSE is given by

$MSE = E\left\{(\hat{X} - X)^2\right\}$

where the expectation is taken over both X and Y.

$cov(X) = E[XX^T]$

http://en.wikipedia.org/wiki/Minimum_mean_square_error

in statistics, the mean square error or MSE of an estimator is one of many ways to quantify the difference between an estimator and the true value of the quantity being estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. . . . the defference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

http://en.wikipedia.org/wiki/Mean_squared_error

alternative target function thus we seek the weights, or equivalently the $\hat{V}$, that minimize E for the observed training examples

– damn, statistics would make this all intuitive and clear

alternative target function several algorithms are known for finding weights of a linear function that minimize E; we require an algorithm that will incrementally refine the weights as new training examples become available and that will be robust to errors in these estimated training values.

alternative target function one such algorithm is called the least mean squares, or LMS training rule.

least mean squares (LMS) algorithms is a type of adaptive filter used to mimic a desired filter by finding the filter coefficients that relate to producing the least mean squares of the error signal (difference between the desired and the actual signal). it is a stochastic gradient descent method in that the filter is only adapted based on the error at the current time.

the diea behind LMS filters is to use steepest descent to find filter weight h(n) which minimize a cost function:

$C(N) = E\left\{|e(n)|^2\right\}$

where e(n) is the error at the current sample 'n' and E{.} denotes the expected value.

this cost function is the mean square error, and is minimized by the LMS.

applying steepest descent means to take the partial derivatives with respect to the individual entries of the filter coefficient (weight) vector, where $\bigtriangledown$ is the gradient operator:

$\hat{h}(n+') = \hat{h}(n) - \frac{\mu}{2} \bigtriangledown C(n) = \hat{h}(n) + \mu E\{x(n)e^*(n)\}$

where $\frac{mu}{2}$ is the step size. that means we have found a sequential update algorithm which minimizes the cost function. unfortunately, this algorithm is not realizable until we know $E\{x(n)e^*(n)\}$.

for most systems, the expectation function must be approximated. this can be done with the following unbiased estimator:

$\hat{E}\{x(n)e^*(n)\} = \frac{1}{N}\sum_{i=0}^{N-1} x(n-i)e^*(n-i)$

where N indicates the number of samples we use for that estimate.

the simplest case is $N = 1$:

$\hat{h}(n+1) = \hat{h}(n) + \mu x(n)e^*(n)$

http://en.wikipedia.org/wiki/Least_mean_squares_filter

in probability theory and statistics, the expected value (or expectation value, or mathematical expectation, or mean, or first moment) of a random variable is the integral of the random variable with respect to its probability measure.

for discrete random variables this is equivalent to the probability-weighted sum of the possible values.

for continuous random variables with a density function it is the probability density-weighted integral of the possible values.

it os often helpful to interpret the expected value of a random variable as the long-run average value of the variable over many independent repetitions of an experiment.

the expected value, when it exists, is almost surel the limit of the sample mean as sample size grows to infitiny.

http://en.wikipedia.org/wiki/Expected_value

- damn, everytime we encroach something interesting; find out why differential equations, linear algebra, probability and statistics are so important. that's like two years of fucking work, isn't it? or at least one? maybe it's worth it, if we can pull it

alternative target function LMS weight update rule: for each training example $\langle b, V_{train}(b)\rangle$:

- use the current weights to calculate $\hat{V}(b)$
- for each weight w_i, update it as: $w_i \leftarrow w_i + \eta(V_{train}(b) - \hat{V}(b))x_i$

alternative target function here η is a small constant (e.g., 0.1) that moderates the size of the weight update.

alternative target function notice that when the error $V_{train}(b) - \hat{V}(b)$ is zero, no weights are changed. when $V_{train}(b) - \hat{V}(b)$ is positive (i.e., when $\hat{V}(b)$ is too low), then each weight is increased in proportion to the value of its correpsonding feature. this will raise the value of $\hat{V}(b)$, reducing the error. notice that if the value of some feature x_i is zero, then its weight is not altered regardless of the error, so that the only weights updated are those whose features actually occur on the training example board.

- mastering these things takes practice; the practice, indeed, of mastering things; long haul, if crossfit, for instance, is to be believed; and raising kids
- don't forget: $V_{train}(b)$ (for intermediate values) is $\hat{V}(Successor(b))$, where $\hat{V}$ is the learner's current approimation to V and where $Successor(b)$ denotes the next board state following b for which it is again the program's turn to move

performance system solve the given performance task (e.g. playing checkers) by using the learned target function(s). it taks an instance of a new problem (game) as input and produces a trace of its solution (game history) as output (e.g. select its next move at each step by the learned $\hat{V}$ evaluation function). we expect its performance to improve as this evaluation function becomes increasingly accurate.

critic takes history or trace of the game produces as output set of training examples of the target function: $\{\langle b_1, V_{train}(b_1)\rangle, \ldots, \langle b_n, V_{train}(b_2)\rangle\}$.

generalizer takes as input training examples, produces an output hypothesis that is its estimate of the target function. it generalizes from the specific training examples, hypothesizing a general function that covers these examples and other cases beyond the training examples. generalize correpsonds to the LMS algorithm, and the output hypothesis is the function $\hat{V}$ described by the learned weight $w_0, \ldots, w_6$.

experiment generator takes as input the current hypothesis (currently learned function) and outputs a new problem (i.e. initial board state) for the performance system to explore. more sophisticated strategioes could involve creating board positions designed to explore particular regions of the state space.

experiment generator many machine learning systems can be usefully characterized in terms of these four generic modules.

```
digraph design {
  generator [label="Experiment Generator"]
  performer [label="Performance System"]
  critic [label=Critic]
  generalizer [label=Generalizer]
  performer -> critic [label="Solution trace"]
  critic -> generalizer [label="Training examples"]
  generalizer -> generator [label=Hypothesis]
  generator -> performer [label="New problem"]
}
```

experiment generator restricted type of knowledge to a single linear eval function; constrained eval function to depend on only six specific board features; if not, best we can hope for is that it will learn a good approximation.

experiment generator let us suppose a good approximation to V can be represented thus; question as to whether this learning technique is guaranteed to find one.

experiment generator linear function representation for $\hat{V}$ too simple to capture well the nuances of the game.

- program represents the learned eval function using an artifical neural network that considers the complete description of the board state rather than a subsect of board features.

nearest neighbor store training examples, try to find "closest" stored situation

genetic algorithm generate large number of candidate checkers programs allow them to play against each other, keeping only the most successful programs

explanation-based learning analyze reasons underlying specific successes and failures

explanation-based learning learning involves searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner.

explanation-based learning many chapters preset algorithms that search a hypothesis space defined by some underlying representation (linear functions, logical descriptions, decision trees, neural networks); for each of these hypotheses representations, the correpsponding learning algorithm takes advantage of a different underlying structure to organize the search through the hypothesis space.

explanation-based learning	. . . confidence we can have that a hypothesis consistent with the training data will correctly generalize to unseen examples
explanation-based learning	what algorithms exist?
explanation-based learning	how much training data?
explanation-based learning	prior knowledge?
explanation-based learning	choosing useful next training experience?
explanation-based learning	how to reduce the learning task to one of more function approximation problems?
explanation-based learning	learner alter its representation to improve ability to represent and learn the target function?
explanation-based learning	determine type of training experience (games against experts, games against self, table of correct moves, . . .); determine target function (board -> move, board -> value, . . .); determine representation of learned function (polynomial, linear function, neural network, . . .); determine learning algorithm (gradient descent, linear programming, . . .).