Foundations of AI

14. Machine Learning

Learning from Observations
Wolfram Burgard and Luc De Raedt and Bernhard Nebel

Learning

• What is learning?
  An agent learns when it can improve its capability to perform a task through experience.
  → e.g. Game programs
• Why learn?
  → Engineering, philosophy, cognitive science
  → Data Mining (discovery of new knowledge through data analysis)
  No intelligence without learning!

Contents (Preliminary)

• The learning agent
• Types of Learning
• Learning decision trees
• Why learning works
• Statistical learning methods

The Learning Agent

So far an agent’s percepts have only served to help the agent choose its actions. Now they will also serve to improve future behaviour.
Building Blocks of the Learning Agent

- **Performance element:** Processes percepts and chooses actions.
  - Corresponds to the agent model we have studied so far.

- **Learning element:** Carries out improvements
  - requires self knowledge and feedback on how the agent is doing in the environment.

- **Critic:** Evaluation of the agent’s behaviour based on a given external behavioral measure
  - feedback.

- **Problem generator:** Suggests explorative actions that lead the agent to new experiences.

The Learning Element

Its design is affected by three major issues:

1. Which **components** of the performance element are to be learned?
2. What **representation** should be chosen?
3. What form of **feedback** is available?
4. Which **prior information** is available?

Types of Feedback During Learning

The **type of feedback** available for learning is usually the **most important factor in determining the nature of the learning problem.**

- **Supervised learning:** Involves learning a function from examples of its inputs and outputs.

- **Unsupervised learning:** The agent has to **learn patterns** in the input when no specific output values are given.

- **Reinforcement learning:** The **most general form of learning** in which the agent is not told what to do by a teacher. Rather it must learn from a reinforcement or reward. It typically involves learning how the environment works.

Inductive Learning

An **example** is a pair (x, f(x)). The **complete set of examples** is called the **training set.**

**Pure inductive inference:** for a collection of examples for f, return a function h (hypothesis) that approximates f.

The function h typically is member of a **hypothesis space H.**

A **good hypothesis** should generalize the data well, i.e. will predict unseen examples correctly.

A **hypothesis is consistent** with the data set if it agrees with all the data.

How do we choose from among multiple consistent hypotheses?

**Ockham’s razor:** prefer the simplest hypothesis consistent with the data.
Example: Fitting a Function to a Data Set

- a) consistent hypothesis that agrees with all the data
- b) degree-7 polynomial that is also consistent with the data set
- c) data set that can be approximated consistently with a degree-6 polynomial
- d) sinusoidal exact fit to the same data

Decision Trees

**Input:** Description of an object or a situation through a set of attributes.

**Output:** a decision, that is the predicted output value for the input.

Both, input and output can be discrete or continuous.

Discrete-valued functions lead to classification problems.

Learning a continuous function is called regression.

Boolean Decision Tree

**Input:** set of vectors of input attributes X and a single Boolean output value y (goal predicate).

**Output:** Yes/No decision based on a goal predicate.

**Goal of the learning process:** Definition of the goal predicate in the form of a decision tree.

**Boolean decision trees** represent **Boolean functions**.

Properties of (Boolean) Decision Trees:
- An internal node of the decision tree represents a test of a property.
- Branches are labeled with the possible values of the test.
- Each leaf node specifies the Boolean value to be returned if that leaf is reached.

When to Wait for Available Seats at a Restaurant

**Goal predicate:** WillWait

**Test predicates:**
1. **Patrons:** How many guests are there? (none, some, full)
2. **WaitEstimate:** How long do we have to wait? (0-10, 10-30, 30-60, >60)
3. **Alternate:** Is there an alternative? (T/F)
4. **Hungry:** Am I hungry? (T/F)
5. **Reservation:** Have I made a reservation? (T/F)
6. **Bar:** Does the restaurant have a bar to wait in? (T/F)
7. **Fri/Sat:** Is it Friday or Saturday? (T/F)
8. **Raining:** Is it raining outside? (T/F)
9. **Price:** How expensive is the food? ($, $$, $$$)
10. **Type:** What kind of restaurant is it? (French, Italian, Thai, Burger)
Restaurant Example (Decision Tree)

Expressiveness of Decision Trees

Each decision tree hypothesis for the WillWait goal predicate can be seen as an assertion of the form

$$\forall s \text{ WillWait}(s) \iff (P_1(s) \lor P_2(s) \lor \ldots \lor P_n(s))$$

where each $P_i(s)$ is conjunction of tests along a path from the root of the tree to a leaf with a positive outcome.

Any Boolean function can be represented by a decision tree.

Limitation: All tests always involve only one object and the language of traditional decision trees is inherently propositional.

$$\exists r_2 \text{ Nearby}(r_2, r) \cdot \text{Price}(r, p) \cdot \text{Price}(r_2, p_2) \cdot \text{Cheaper}(p_2, p)$$

cannot be represented as a test.

We could always add another test called CheaperRestaurantNearby, but a decision tree with all such attributes would grow exponentially.

Extensions exist, e.g. (Blockeel and De Raedt, Artificial Intelligence, 1998).

Example: Mutagenicity

Active

Inactive

Structural alert:

Compact Representations

For every Boolean function we can construct a decision tree by translating every row of a truth table to a path in the tree.

This can lead to a tree whose size is exponential in the number of attributes.

Although decision trees can represent functions with smaller trees, there are functions that require an exponentially large decision tree:

Parity function:

$$p(x) = \begin{cases} 1 & \text{if } x \text{ is } \text{odd} \cr 0 & \text{otherwise} \end{cases}$$

Majority function:

$$m(x) = \begin{cases} 1 & \text{if } \frac{x}{2} \text{ are } 1 \cr 0 & \text{otherwise} \end{cases}$$

There is no consistent representation that is compact for all possible Boolean functions.
The Training Set of the Restaurant Example

Classification of an example = Value of the goal predicate

TRUE → positive example
FALSE → negative example

<table>
<thead>
<tr>
<th>Example</th>
<th>Attributes</th>
<th>Will Wait</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Yes No No Yes Some SSS No Yes French 0-10</td>
<td>Yes</td>
</tr>
<tr>
<td>X2</td>
<td>Yes No No Yes Full $ No No Thai 30-60</td>
<td>No</td>
</tr>
<tr>
<td>X3</td>
<td>No Yes No No Some $ No No Burger 0-10</td>
<td>Yes</td>
</tr>
<tr>
<td>X4</td>
<td>Yes No Yes Yes Full $ No No Thai 10-30</td>
<td>Yes</td>
</tr>
<tr>
<td>X5</td>
<td>Yes No Yes No Full SSS No Yes French &gt;60</td>
<td>No</td>
</tr>
<tr>
<td>X6</td>
<td>No Yes No Yes Some $S Yes Yes Italian 0-10</td>
<td>Yes</td>
</tr>
<tr>
<td>X7</td>
<td>No Yes No No None $ Yes No Burger 0-10</td>
<td>No</td>
</tr>
<tr>
<td>X8</td>
<td>No No No Yes Some SSS Yes Yes Thai 0-10</td>
<td>Yes</td>
</tr>
<tr>
<td>X9</td>
<td>No Yes Yes No Full $ Yes No Burger &gt;60</td>
<td>No</td>
</tr>
<tr>
<td>X10</td>
<td>Yes Yes No Yes Full SSS No Yes Italian 10-30</td>
<td>No</td>
</tr>
<tr>
<td>X11</td>
<td>No No No No None $ No No Thai 0-10</td>
<td>No</td>
</tr>
<tr>
<td>X12</td>
<td>Yes Yes Yes Yes Full $ No No Burger 30-60</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Inducing Decision Trees from Examples

• **Naïve solution**: we simply construct a tree with one path to a leaf for each example.
• In this case we test all the attributes along the path and attach the classification of the example to the leaf.
• Whereas the resulting tree will correctly classify all given examples, it will not say much about other cases.
• It just memorizes the observations and does not generalize.

Inducing Decision Trees from Examples

• **Smallest solution**: applying Ockham’s razor we should instead find the smallest decision tree that is consistent with the training set.
• Unfortunately, for any reasonable definition of smallest finding the smallest tree is intractable.
• Dilemma:
  
  smallest \[\rightarrow\] simplest
  
  intractable \[\rightarrow\] no learning

• We can give a decision tree learning algorithm that generates “smallish” trees.

Idea of Decision Tree Learning

Divide and Conquer approach:

• **Choose** an (or better: the best) attribute.
• **Split** the training set into subsets each corresponding to a particular value of that attribute.
• Now that we have divided the training set into several smaller training sets, we can recursively apply this process to the smaller training sets.
Splitting Examples (1)

- Type is a **poor attribute**, since it leaves us with four subsets each of them containing the same number of positive and negative examples.
- It does not reduce the problem complexity.

Splitting Examples (2)

- Patrons is a **better choice**, since if the value is None or Some, then we are **left with example sets for which we can answer definitely** (Yes or No).
- Only for the value Full we are left with a mixed set of examples.
- One potential next choice is Hungry.

Recursive Learning Process

In each recursive step there are **four cases to consider**:

1. **Positive and negative examples**: choose a new attribute.
2. **Only positive (or only negative) examples**: done (answer is Yes or No).
3. **No examples**: there was no example with the desired property. Answer Yes if the majority of the parent node’s examples is positive, otherwise No.
4. **No attributes** left, but there are still examples with different classifications: there were errors in the data (\(\rightarrow\) **NOISE**) or the attributes do not give sufficient information. **React as in the previous case**.

The Decision Tree Learning Algorithm

```plaintext
function DECISION-TREE-LEARNING(examples, attrs, default) returns a decision tree
inputs: examples, set of examples
        attrs, set of attributes
        default, default value for the goal predicate

if examples is empty then return default
else if all examples have the same classification then return the classification
else if attrs is empty then return MAJORITY-VALUE(examples)
else
    best ← CHOOSE-ATTRIBUTE(attrs, examples)
    tree ← a new decision tree with root best
    m ← MAJORITY-VALUE(examples)
    for each value \(v_i\) of best do
        examples_i ← {elements of examples with best \(= v_i\)}
        subtree ← DECISION-TREE-LEARNING(examples_i, attrs \(\setminus\) best, m)
        add a branch to tree with label \(v_i\) and subtree subtree
    return tree
```

21 22 23 24
Application to the Restaurant Data

Properties of the Resulting Tree

- The resulting tree is considerably simpler than the one originally given (and from which the training examples were generated).
- The learning algorithm outputs a tree that is consistent with all examples it has seen.
- The tree does not need to agree with the correct function.
- For example, it suggests not to wait if we are not hungry. If we are, there are cases in which it tells us to wait.
- Some tests (Raining, Reservation) are not included since the algorithm can classify the examples without them.

Choosing Attribute Tests

choosing-attribute\(\langle\text{attrs}, \text{examples}\rangle\)

- One goal of decision tree learning is to select attributes that minimize the depth of the final tree.
- The perfect attribute divides the examples into sets that are all positive or all negative.
- Patrons is not perfect but fairly good.
- Type is useless since the resulting proportion of positive and negative examples in the resulting sets are the same as in the original set.
- What is a formal measure of “fairly good” and “useless?”

Evaluation of Attributes

Tossing a coin: What value has prior information about the outcome of the toss when the stakes are $1 and the winnings $1?

- Rigged coin with 99% heads and 1% tails. (average winnings per toss = $0.98)
  → Worth of information about the outcome is less than $0.02.
- Fair coin
  → Value of information about the outcome is less than $1.
  → The less we know about the outcome, the more valuable the prior information.
Information Provided by an Attribute

- One suitable measure is the expected amount of information provided by the attribute.
- Information theory measures information content in bits.
- One bit is enough to answer a yes/no question about which one has no idea (fair coin flip).
- In general, if the possible answers \(v_i\) have probabilities \(P(v_i)\), the information content is given as

\[
I(P(v_1),...,P(v_n)) = \sum_{i=1}^{n} - P(v_i) \log_2(P(v_i))
\]

Examples

\[
I\left(\frac{1}{2}, \frac{1}{2}\right) = 1
\]

\[
I(1,0) =
\]

\[
I(0,1) =
\]

Attribute Selection (1)

Attribute A divides the example set into \(p\) positive and \(n\) negative examples:

\[
I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = \frac{p}{p+n} \log_2\left(\frac{p+n}{p}\right) + \frac{n}{p+n} \log_2\left(\frac{p+n}{n}\right)
\]

The value of A also depends on additional information that we still need to collect after we selected it.

Suppose A divides the training set \(E\) into subsets \(E_i, i = 1, ..., v\).

Every subset has

\[
I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)
\]

A random example has value \(i\) with probability \(\frac{p_i}{p+n}\)

Attribute Selection (2)

→ The average information content after choosing A is

\[
R(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)
\]

→ The information gain from choosing A is

\[
Gain(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - R(A)
\]

Heuristic in choose-attribute is to select the attribute with the largest gain.

Examples:

\[
Gain(Patrons) = 1 - \left[\frac{12}{12} I(0.1,0.1) + \frac{12}{12} I(1,0) + \frac{6}{12} I(\frac{3}{4}, \frac{1}{4})\right] = 0.541
\]

\[
Gain(Type) = 1 - \left[\frac{12}{12} I(\frac{1}{4}, \frac{3}{4}) + \frac{12}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{12}{12} I(\frac{3}{4}, \frac{1}{4}) + \frac{12}{12} I(\frac{3}{4}, \frac{1}{4})\right] = 0
\]
Assessing the Performance of the Learning Algorithm

Methodology for assessing the power of prediction:
- Collect a large number of examples.
- Divide it into two disjoint sets: the training set and the test set.
- Use the training set to generate $h$.
- Measure the percentage of examples of the test set that are correctly classified by $h$.
- Repeat the process for randomly-selected training sets of different sizes.

Learning Curve for the Restaurant Example

As the training set grows, the prediction quality increases.

Important Strategy for Designing Learning Algorithms

- The training and test sets must be kept separate.
- Common error: Changing the algorithm after running a test, and then testing it with training and test sets from the same basic set of examples. By doing this, knowledge about the test set gets stored in the algorithm, and the training and test sets are no longer independent.

Noise

- What is noise? Random errors in the learning data
- Effect: Larger trees make more mistakes on new data (overfitting)
- Avoidance of overfitting by means of a "validation set": The training set is divided into two groups; 70% of the training set is used to build the tree, and the remaining 30% to define an appropriate tree size ("pruning")
Post-pruning

• One way to deal with noise
  – Split training set into e.g.
    • 70% for learning
    • 30% for validation
  – First build the tree as usual on learning set
  – Then iterate as long as accuracy on validation set increases
    • Turn each subtree into leaf and measure accuracy on validation set
    • Select that new tree that increases the accuracy the most on the validation set

Summary: Decision Trees

• One possibility for representing (Boolean) functions.
• Decision trees can be exponential in the number of attributes.
• It is often too difficult to find the minimal DT.
• One method for generating DTs that are as flat as possible is based on weighing the attributes.
• The weights are computed based on the information gain.

Why Learning Works

How can we decide that $h$ is close to $f$ when $f$ is unknown?

→ Probably approximately correct

Stationarity as the basic assumption of PAC-Learning: training and test sets are selected from the same population of examples with the same probability distribution.

Key question: how many examples do we need?

$X$ Set of examples
$D$ Distribution from which the examples are drawn
$H$ Hypothesis space ($f \in H$)
$m$ Number of examples in the training set

$\text{error}(h) = P(h(x) \neq f(x) \mid x \text{ drawn from } D) \leq \varepsilon$

PAC-Learning

A hypothesis $h$ is approximately correct if $\text{error}(h) \leq \varepsilon$.

To show: After the training period with $m$ examples, with high probability, all consistent hypotheses are approximately correct.

How high is the probability that a wrong hypothesis $h_p \in H_{\text{bad}}$ is consistent with the first $m$ examples?
Sample Complexity

Assumption: \( \text{error}(h_0) > 0 \). \( \Rightarrow \)

\[
\begin{align*}
P(h_0 \text{ is consistent with 1 example}) & \leq (1 - \delta) \\
P(h_0 \text{ is consistent with } N \text{ examples}) & \leq (1 - \delta)^N \\
P(H_{rad} \text{ contains a consistent } h) & \leq |H_{rad}|(1 - \delta)^N \\
\text{Since } |H_{rad}| & \leq |H| \\
P(H_{rad} \text{ contains a consistent } h) & \leq |H|(1 - \delta)^N
\end{align*}
\]

We want to limit this probability by some small number \( \delta \):

\[|H|(1 - \delta)^N < \delta\]

Since \( \ln(1 - \delta) \leq e^{-\delta} \), we derive

\[N \geq \frac{1}{\delta} (\ln(1/\delta) + \ln|H|)\]

Sample Complexity: Number of required examples, as a function of \( \Box \) and \( \delta \).

---

Sample Complexity (2)

Example: Boolean functions

The number of Boolean functions over \( n \) attributes is \( |H| = 2^{2^n} \).

The sample complexity therefore grows as \( 2^n \).

Since the number of possible examples is also \( 2^n \), any learning algorithm for the space of all Boolean functions will do no better than a lookup table, if it merely returns a hypothesis that is consistent with all known examples.

---

Learning from Decision Lists

In comparison to decision trees:

- The overall structure is simpler
- The individual tests are more complex

\[
\begin{array}{cc}
\text{Patrons}(x, \text{Same}) & N \\
\downarrow \text{Y} & \text{No} \\
\end{array}
\]

\[
\begin{array}{cc}
\text{Patrons}(x, \text{Full}) \\
\land \text{Fri/Sat}(x) & N \\
\downarrow \text{Y} & \text{Yes} \\
\end{array}
\]

This represents the hypothesis

\( H_d: \forall x \in \text{WillWait}(x) \leftrightarrow \text{Patrons}(x, \text{same}) \land \text{Fri/Sat}(x) \)

If we allow tests of arbitrary size, then any Boolean function can be represented.

\( k\text{-DL}: \) Language with tests of length \( \leq k \).

---

Learnability of \( k\text{-DL} \)

\[
\text{function } \text{DECISION-LIST-LEARNING}(\text{examples}) \text{ returns a decision list, } \text{No } \text{or } \text{failure}
\]

\[
\text{if examples is empty then return the value } \text{No} \\
\text{let } t \leftarrow \text{a test that matches a nonempty subset examples, of examples} \\
\text{such that the members of examples are all positive or all negative} \\
\text{if there is no such } t \text{ then return failure} \\
\text{if the examples in examples are all positive then } a \leftarrow \text{Yes} \\
\text{else } a \leftarrow \text{No} \\
\text{return a decision list with initial test } t \text{ and outcome } a \text{ and remaining elements given by } \text{DECISION-LIST-LEARNING}(\text{examples} - \text{examples})
\]

\[
|k\text{-DL}(n)| \leq 3^{|\text{Conj}(n)|} \times |\text{Conj}(n, k)|! \quad (\text{Yes,No,no-Test,all permutations})
\]

\[
|\text{Conj}(n, k)| = \sum_{i=0}^{k} \binom{2^n}{i} \quad (\text{Combination without repeating pos/neg attributes})
\]

\[
O(n^k)
\]

\[
|k\text{-DL}(n)| = 2^{O(n^k \log(n^k))} \quad (\text{with Euler’s summation formula})
\]

\[
m \geq \frac{1}{\delta} \left( \ln(1/\delta) + O(n^k \log(n^k)) \right)
\]
Summary

Inductive learning as learning the representation of a function from example input/output pairs.

- **Decision trees** learn deterministic Boolean functions.
- **PAC learning** deals with the complexity of learning.
- **Decision lists** as functions that are easy to learn.

Statistical Learning Methods

- When discussing MDPs we saw that **probability and utility theory** allow agents to deal with uncertainty.
- To apply these techniques, however, the agents must first **learn** their **probabilistic theories of the world from experience**.
- We will discuss **statistical learning methods** as robust ways to **learn probabilistic models**.

An Example for Statistical Learning

- The **key concepts** are **data** (evidence) and hypotheses.
- A candy manufacturer sells five kinds of bags that are indistinguishable from the outside:
  - $h_1$: 100% cherry
  - $h_2$: 75% cherry and 25% lime
  - $h_3$: 50% cherry and 50% lime
  - $h_4$: 25% cherry and 75% lime
  - $h_5$: 100% lime
- Given a sequence $D_1, \ldots, D_N$ of candies inspected, what is the flavor of the next piece of candy?

An Example for Statistical Learning

- The **key concepts** are **data** (evidence) and hypotheses.
- A candy manufacturer sells five kinds of bags that are indistinguishable from the outside:
  - $h_1$: 100% cherry
  - $h_2$: 75% cherry and 25% lime
  - $h_3$: 50% cherry and 50% lime
  - $h_4$: 25% cherry and 75% lime
  - $h_5$: 100% lime
- Given a sequence $D_1, \ldots, D_N$ of candies inspected, what is the flavor of the next piece of candy?
Bayesian Learning

- **Calculates** the probability of each hypothesis, given the data.
- It then **makes predictions using all hypotheses weighted by their probabilities** (instead of a single best hypothesis).
- **Learning** is reduced to probabilistic inference.

Application of Bayes Rule

- Let $D$ represent all the data with observed value $d$.
- The probability of each hypothesis is obtained by Bayes rule:
  
  $$P(h_i \mid d) = \alpha \ P(d \mid h_i) \ P(h_i)$$

  - The manufacturer tells us that the **prior distribution** over $h_1$, ..., $h_5$ is given by $<0.1, 0.2, 0.4, 0.2, 0.1>$
  - We compute **likelihood of the data** under the assumption that the observations are independently, and identically distributed (i. i.d.):
    
    $$P(d \mid h_i) = \prod_j P(d_j \mid h_i)$$

How to Make Predictions?

- Suppose we want to **make predictions about an unknown quantity** $X$ given the data $d$.
  
  $$P(X \mid d) = \sum_i P(X \mid h_i, d) \ P(h_i \mid d)$$

  - Predictions are weighted averages over the predictions of the individual hypotheses.
  - The **key quantities** are the hypothesis prior $P(h_j)$ and the likelihood $P(d \mid h_j)$ of the data under each hypothesis.

Example

- Suppose the bag is an all-lime bag ($h_3$)
- The first 10 candies are all lime.
- Then $P(d \mid h_3)$ is 0.5$^{10}$ because half the candies in an $h_3$ bag are lime.
- Evolution of the five hypotheses given 10 lime candies were observed (the values start at the prior!).
Observations

- The **true hypothesis often dominates** the Bayesian prediction.
- For any **fixed prior** that does **not rule out the true hypothesis**, the **posterior of any false hypothesis** will eventually **vanish**.
- The **Bayesian prediction is optimal** and, given the hypothesis prior, any other prediction will be **correct less often**.
- It comes at a price that the **hypothesis space** can be **very large or infinite**.

Maximum a Posteriori (MAP)

- A common **approximation** is to **make predictions based on a single most probable hypothesis**.
- The **maximum a posteriori** (MAP) hypothesis is the one that **maximizes** $P(h|d)$.
  \[
  P(X \mid d) = P(X \mid h_{MAP})
  \]
- In the candy example, $h_{MAP} = h_5$ after three lime candies in a row.
- The MAP learner predicts that the fourth candy is lime with probability 1.0, whereas the Bayesian prediction is still 0.8.
- As more data arrive, MAP and Bayesian predictions become closer.
- Finding MAP hypotheses is often much easier than Bayesian learning.

Maximum-Likelihood Hypothesis (ML)

- A final **simplification** is to **assume a uniform prior** over the hypothesis space.
- In that case **MAP-learning reduces to choosing the hypothesis that maximizes** $P(d|h_i)$.
- This hypothesis is called the **maximum-likelihood hypothesis** (ML).
- ML-learning is a **good approximation** to MAP learning and Bayesian learning when there is a uniform prior and when the data set is large.

Summary

(Statistical Learning Methods)

- **Bayesian learning techniques** formulate learning as a form of **probabilistic inference**.
- **Maximum a posteriori** (MAP) learning selects the **most likely hypothesis given the data**.
- **Maximum likelihood learning** selects the hypothesis that **maximizes the likelihood of the data**.