Universität Freiburg
Lehrstuhl für Maschinelles Lernen und natürlichsprachliche Systeme

Machine Learning (SS2012)

Prof. Dr. M. Riedmiller, Manuel Blum

Exercise Sheet 1

mblum@informatik.uni-freiburg.de

Exercise 1.1: Introduction to Machine Learning

(a) Visit the following website:

`http://www.pacman-vs-ghosts.net`

Think of different aspects of the Pacman Task, that could be solved using Machine Learning algorithms.

(b) Given an appropriate dataset, the problems given below can be solved by Machine Learning algorithms. Which of the problems would you apply supervised learning to?

(c) There are two types of supervised learning tasks, classification and regression. Decide for the supervised learning problems given below, whether it is a classification problem or not.

Task	Supervised Learning	Classification
Predict tomorrow's price of a particular stock.	☒	☐
Discover whether there are different types of spam mail and what categories there are.	☐	☐
Predict your life expectancy.	☒	☐
Predict if it is going to rain tomorrow.	☒	☒
Learn to grasp an object by trial and error.	☐	☐

Table 1: Problems that can be solved using Machine Learning techniques.

Exercise 1.2: Version Spaces and Conjunctive Hypotheses

(a) What are the elements of the version space? How are they ordered? What can be said about the meaning and sizes of S and G?

- hypothesen
- $VS_{H,D} \subseteq H$ hypotheses with the training data D
- general-to-specific

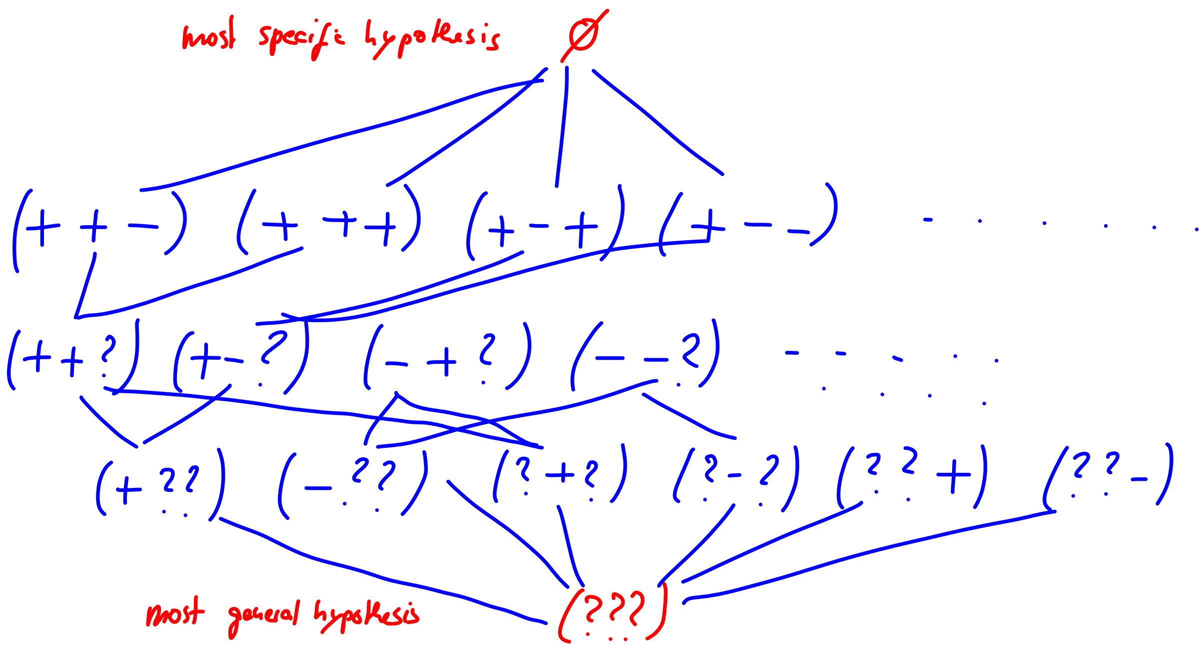
  $h_1 \leq_g h_2$ if $h_1(x) = 1$ implies $h_2(x) = 1$
- G: most general
- S: most specific

$\left.\begin{matrix} G \\ S \end{matrix}\right\}$ hypotheses consistent with the training data

$|S| = 1$ for consistent training data

(b) In the following, it is desired to describe whether a person is *ill*. We use a representation based on conjunctive constraints (three per subject) to describe individual person. These constraints are "running nose", "coughing", and "reddened skin", each of which can take the value true ('+') or false ('–'). We say that somebody is ill, if he is coughing and has a running nose. Each single symptom individually does not mean that the person is ill.

Specify the space of hypotheses that is being managed by the version space approach. To do so, arrange all hypotheses in a graph structure using the more-specific-than relation.



(c) Apply the candidate elimination (CE) algorithm to the sequence of training examples specified in Table 2 and name the contents of the sets S and G after each step.

Training	running nose	coughing	reddened skin	**Classification**
d_1	+	+	+	positive (ill)
d_2	+	+	−	positive (ill)
d_3	+	−	+	negative (healthy)
d_4	−	+	+	negative (healthy)
d_5	−	−	+	negative (healthy)
d_6	−	−	−	negative (healthy)

Table 2: List of training instances for the medical diagnosis task.

$G = \{\langle ???\rangle\}$ $S = \{\langle \emptyset \emptyset \emptyset\rangle\}$

$d_1 = [\langle +++\rangle, pos] \Rightarrow G = \{\langle ???\rangle\}$ $S = \{\langle +++\rangle\}$

$d_2 = [\langle ++-\rangle, pos] \Rightarrow G = \{\langle ???\rangle\}$ $S = \{\langle ++?\rangle\}$

$d_3 = [\langle +-+\rangle, neg] \Rightarrow S = \{\langle ++?\rangle\}$

$G = \{\cancel{\langle -??\rangle}\ \langle ?+?\rangle\ \cancel{\langle ??-\rangle}\}$

$S = \{[++?]\}$ $G = \{\langle ?+?\rangle\}$

$d_4 = [\langle -++\rangle, neg]$ $S = \{\langle ++?\rangle\}$

$G = \{\langle ++?\rangle, \cancel{\langle ?+-\rangle}\}$

$S = G$

$\leq_g$

(d) Does the order of presentation of the training examples (according to Table 2) to the learner affect the finally learned hypothesis?

No.

May influence the algorithms running time.

(e) Assume a domain with two attributes, i.e. any instance is described by two constraints. How many positive and negative training examples are minimally required by the candidate elimination algorithm in order to learn an arbitrary concept?

Learn concept: $S = G$

$S = \{\langle \emptyset \emptyset \rangle\}$ $G = \{\langle ?? \rangle\}$

Best case: all training samples can be used to adapt S or G

Neg.: G: $\langle ?? \rangle$ → $\langle v\,? \rangle$, $\langle ?, w \rangle$ → $\langle v, w \rangle$

Pos.: S: $\langle \emptyset \emptyset \rangle \rightarrow \langle v, w \rangle$ → $\langle ?, w \rangle$, $\langle v, ? \rangle$ → $\langle ?? \rangle$

(f) We are now extending the number of constraints used for describing training instances by one additional constraint named "fever". We say that somebody is ill, if he has a running nose and is coughing (as we did before), or if he has fever.

How does the version space approach using the CE algorithm perform now, given the training examples specified in Table 3? What happens, if the order of presentation of the training examples is altered?

Training	running nose	coughing	reddened skin	fever	**Classification**
d_1	+	+	+	−	positive (ill)
d_2	+	+	−	−	positive (ill)
d_3	−	−	+	+	positive (ill)
d_4	+	−	−	−	negative (healthy)
d_5	−	−	−	−	negative (healthy)
d_6	−	+	+	−	negative (healthy)

Table 3: List of training instances using the extended representation.

- Initially: $S = \{\langle \emptyset\emptyset\emptyset\emptyset \rangle\}$, $G = \{\langle * * ** \rangle\}$
- $d_1 = [\langle + + + - \rangle, pos] \Rightarrow S = \{\langle + + + - \rangle\}$, $G = \{\langle * * ** \rangle\}$
- $d_2 = [\langle + + - - \rangle, pos] \Rightarrow S = \{\langle + + * - \rangle\}$, $G = \{\langle * * ** \rangle\}$
- $d_3 = [\langle - - + + \rangle, pos] \Rightarrow S = \{\langle * * ** \rangle\}$, $G = \{\langle * * ** \rangle\}$

  $\rightarrow$ We already arrive at $S = G$.
- $d_4 = [\langle + - - - \rangle, neg] \Rightarrow S = \{\langle * * ** \rangle\}$, $G = \{\langle * * ** \rangle\}$
  - Now, S becomes empty since $\langle * * ** \rangle$ is inconsistent with d_4 and is removed from S.
  - G would be specialized to $\{\langle - * ** \rangle, \langle * + ** \rangle, \langle * * + * \rangle, \langle * * * + \rangle\}$. But it is required that at least one element from S must be more specific than any element from G.

    $\rightarrow$ This requirement cannot be fulfilled since $S = \emptyset$. $\Rightarrow G = \emptyset$

- Even a change in the order of presentation does not result in yielding a learning success (i.e. in $S = G \neq \emptyset$).
- When applying the CE algorithm, S and G become empty independent of the presentation order.
- Reason: The informally specified target concept of an "ill person" represents a disjunctive concept.
- The target concept is not an element of the hypothesis space H (which is made of conjunctive hypotheses).

Exercise 1.3: Decision Tree Learning with ID3

(a) Apply the ID3 algorithm to the training data provided in Table 4.

(b) Does the resulting decision tree provide a disjoint definition of the classes?

Training	fever	vomiting	diarrhea	shivering	**Classification**
d_1	no	no	no	no	healthy (H)
d_2	average	no	no	no	influenza (I)
d_3	high	no	no	yes	influenza (I)
d_4	high	yes	yes	no	salmonella poisoning (S)
d_5	average	no	yes	no	salmonella poisoning (S)
d_6	no	yes	yes	no	bowel inflammation (B)
d_7	average	yes	yes	no	bowel inflammation (B)

Table 4: Multi-class training examples.

Entropy: $E(S) = -\sum_{s \in S} p_s \cdot \log_2 p_s$

$= -\frac{1}{2}\log\frac{1}{2} - \frac{2}{7}\log\frac{2}{7} - \frac{2}{7}\log\frac{2}{7} - \frac{2}{7}\log\frac{2}{7} = 1{,}950$

$E(S|x) = \sum_{v \in Values(x)} \frac{|S_v|}{|S|} E(S_v)$

$$E(S|\text{fever}) = \frac{2}{7}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)$$
$$+ \frac{3}{7}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log 3 - \frac{1}{3}\log 3\right)$$
$$+ \frac{2}{7}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)$$
$$= 1{,}251$$

$E(S|\text{vomiting}) = 1{,}251$ $E(S|\text{Diarrhea}) = 0{,}965$ $E(S|\text{Shivering}) = 1{,}644$

Maximize information gain $G(S,x) = E(S) - E(S|x)$

$G(S|\text{fever}) = 1{,}95 - 1{,}251 = 0{,}699$

$G(S|\text{Vomiting}) = 1{,}95 - 1{,}251 = 0{,}699$

$G(S|\text{Diarrh}) = 1{,}95 - 0{,}965 = 0{,}958$

$G(S|\text{Shivering}) = 1{,}95 - 1{,}644 = 0{,}306$ disjoint = yes

(c) Consider the use of real-valued attributes, when learning decision trees, as described in the lecture. Table 5 shows the relationship between the body height and the gender of a group of persons (the records have been sorted with respect to the value of *height* in cm). Calculate the information gain for potential splitting thresholds and determine the best one.

Height	161	164	169	175	176	179	180	184	185
Gender	F	F	M	M	F	F	M	M	F
		C_1		C_2		C_3		C_4	

Table 5: Data on the correlation between body height and gender.

Discretize! Potential split points

$$E(S) = -\frac{5}{9}\log\frac{5}{9} - \frac{4}{9}\log\frac{4}{9} = 0.991$$

$$E(S|C_1) = \frac{2}{9}\cdot 0 - \frac{7}{9}\left(-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right) = \frac{7}{9}\cdot 0{,}985 = 0{,}766$$

$E(S|C_2) = 0{,}984$ $E(S,C_3) = 0{,}918$ $E(S|C_4) = 0{,}889$

C_1 is the best split