gender	age
M	48
M	67
F	53
M	49
F	34
M	21

Based on this data, can we infer anything regarding the patients entering the clinic?

1.6.3 Reinforcement learning

Reinforcement learning is the problem of getting an agent to act in the world so as to maximize its rewards.

A learner (the program) is not told what actions to take as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situations and, through that, all subsequent rewards.

For example, consider teaching a dog a new trick: we cannot tell it what to do, but we can reward/punish it if it does the right/wrong thing. It has to find out what it did that made it get the reward/punishment. We can use a similar method to train computers to do many tasks, such as playing backgammon or chess, scheduling jobs, and controlling robot limbs.

Reinforcement learning is different from supervised learning. Supervised learning is learning from examples provided by a knowledgeable expert.

1.7 Sample questions

(a) Short answer questions

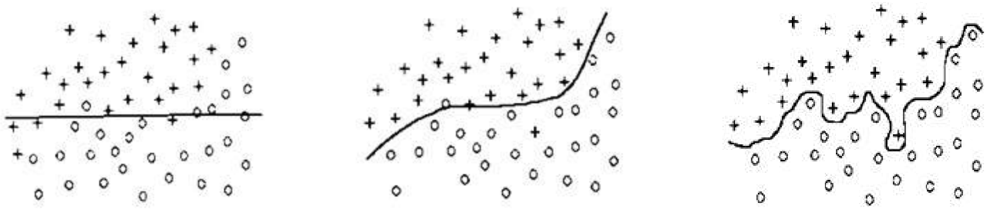
1. What is meant by "learning" in the context of machine learning?
2. List out the types of machine learning.
3. Distinguish between classification and regression.
4. What are the differences between supervised and unsupervised learning?
5. What is meant by supervised classification?
6. Explain supervised learning with an example.
7. What do you mean by reinforcement learning?
8. What is an association rule?
9. Explain the concept of Association rule learning. Give the names of two algorithms for generating association rules.
10. What is a classification problem in machine learning. Illustrate with an example.
11. Give three examples of classification problems from real life situations.
12. What is a discriminant in a classification problem?
13. List three machine learning algorithms for solving classification problems.
14. What is a binary classification problem? Explain with an example. Give also an example for a classification problem which is not binary.
15. What is regression problem. What are the different types of regression?

(b) Long answer questions

1. Give a definition of the term “machine learning”. Explain with an example the concept of learning in the context of machine learning.
2. Describe the basic components of the machine learning process.
3. Describe in detail applications of machine learning in any three different knowledge domains.
4. Describe with an example the concept of association rule learning. Explain how it is made use of in real life situations.
5. What is the classification problem in machine learning? Describe three real life situations in different domains where such problems arise.
6. What is meant by a discriminant of a classification problem? Illustrate the idea with examples.
7. Describe in detail with examples the different types of learning like the supervised learning, etc.

Consider a dataset shown in Figure 2.7(a). Let it be required to fit a regression model to the data. The graph of a model which looks "just right" is shown in Figure 2.7(b). In Figure 2.7(c)we have a linear regression model for the same dataset and this model does seem to capture the essential features of the dataset. So this model suffers from underfitting. In Figure 2.7(d) we have a regression model which corresponds too closely to the given dataset and hence it does not account for small random noises in the dataset. Hence it suffers from overfitting.

Example 2



(a) Underfitting (b) Right fitting (c) Overfitting

Figure 2.8: Fitting a classification boundary

Suppose we have to determine the classification boundary for a dataset two class labels. An example situation is shown in Figure 2.8 where the curved line is the classification boundary. The three figures illustrate the cases of underfitting, right fitting and overfitting.

2.8.1 Testing generalisation: Cross-validation

We can measure the generalization ability of a hypothesis, namely, the quality of its inductive bias, if we have access to data outside the training set. We simulate this by dividing the training set we have into two parts. We use one part for training (that is, to find a hypothesis), and the remaining part is called the *validation set* and is used to test the generalization ability. Assuming large enough training and validation sets, the hypothesis that is the most accurate on the validation set is the best one (the one that has the best inductive bias). This process is called *cross-validation*.

2.9 Sample questions

(a) Short answer questions

1. Explain the general-to-specific ordering of hypotheses.
2. In the context of classification problems explain with examples the following: (i) hypothesis (ii) hypothesis space.
3. Define the version space of a binary classification problem.
4. Explain the "one-against-all" method for learning multiple classes.
5. Describe the "one-against-one" method for learning multiple classes.
6. What is meant by inductive bias in machine learning? Give an example.
7. What is meant by overfitting of data? Explain with an example.
8. What is meant by overfitting and underfitting of data with examples.

(b) Long answer questions

1. Define version space and illustrate it with an example.

2. Given the following data

x	0	3	5	9	12	18	23
Label	0	0	0	1	1	1	1

and the hypothesis space

$$H = \{h_m \,|\, m \text{ a real number}\}$$

where h_m is defined by

$$\text{IF } x \leq m \text{ THEN 1 ELSE } 0,$$

find the version space the problem with respect to D and H.

3. What is meant by "noise" in data? What are its sources and how it is affecting results?

4. Consider the following data:

x	2	3	5	8	10	15	16	18	20
y	12	15	10	6	8	10	7	9	10
Class label	0	0	1	1	1	1	0	0	0

Determine the version space if the hypothesis space consists of all hypotheses of the form

$$\text{IF } (x_1 < x < x_2) \text{ AND } (y_1 < y < y_2) \text{ THEN "1" ELSE "0".}$$

5. For the date in problem 4, what would be the version space if the hypothesis space consists of all hypotheses of the form

$$\text{IF } (x - x_1)^2 + (y - y_1)^2 \leq r^2 \text{ THEN "1" ELSE "0".}$$

6. What issues are to be considered while selecting a model for applying machine learning in a given problem.

3. Gaussian (normal) density

A continuous random variable X has the Gaussian or normal distribution if its density function is

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Here μ and σ are the parameters.

Given a sample $x_1, x_2, \ldots, x_n$ from the distribution. the log likelihood function is

$$L(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left[(x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right].$$

Setting up the equations

$$\frac{dL}{d\mu} = 0, \quad \frac{dL}{d\sigma} = 0$$

and solving for μ and σ we get the maximum likelihood estimates of μ and σ as

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}(x_1 + \cdots + x_n)\\ \hat{\sigma}^2 &= \frac{1}{n}((x_1-\hat{\mu})^2 + \cdots + (x_n-\hat{\mu})^2)\end{aligned}$$

(We leave the details of the derivation as an exercise.)

6.6 Sample questions

(a) Short answer questions

1. What are the assumptions under the naive Bayes algorithm?
2. Why is naive Bayes algorithm “naive”?
3. Given an instance X of a feature vector and a class label c_k, explain how Bayes theorem is used to compute the probability $P(c_k \,|\, X)$.
4. What does a naive Bayes classifier do?
5. What is naive Bayes used for?
6. Is naive Bayes supervised or unsupervised? Why?
7. What is meant by the likelihood of a random sample taken from population?
8. How do we use numeric features in naive Bayes algorithm?

(b) Long answer questions

1. State Bayes theorem and illustrate it with an example.
2. Explain naive Bayes algorithm.
3. Use naive Bayes algorithm to determine whether a red domestic SUV car is a stolen car or not using the following data:

Example no.	Colour	Type	Origin	Whether stolen
1	red	sports	domestic	yes
2	red	sports	domestic	no
3	red	sports	domestic	yes
4	yellow	sports	domestic	no
5	yellow	sports	imported	yes
6	yellow	SUV	imported	no
7	yellow	SUV	imported	yes
8	yellow	SUV	domestic	no
9	red	SUV	imported	no
10	red	sports	imported	yes

4. Based on the following data determine the gender of a person having height 6 ft., weight 130 lbs. and foot size 8 in. (use naive Bayes algorithm).

person	height (feet)	weight (lbs)	foot size (inches)
male	6.00	180	10
male	6.00	180	10
male	5.50	170	8
male	6.00	170	10
female	5.00	130	8
female	5.50	150	6
female	5.00	130	6
female	6.00	150	8

5. Given the following data on a certain set of patients seen by a doctor, can the doctor conclude that a person having chills, fever, mild headache and without running nose has the flu?

chills	running nose	headache	fever	has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

6. Explain the general MLE method for estimating the parameters of a probability distribution.

7. Find the ML estimate for the parameter p in the binomial distribution whose probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

8. Compute the ML estimate for the parameter λ in the Poisson distribution whose probability function is

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Find the ML estimate of the parameter p in the geometric distribution defined by the probability mass function

$$f(x) = (1-p)p^x, \qquad x = 1, 2, 3, \ldots$$

$$
\begin{aligned}
X^T X &= \begin{bmatrix} 4 & 4 & 6 \\ 4 & 6 & 7 \\ 6 & 7 & 10 \end{bmatrix} \\
(X^T X)^{-1} &= \begin{bmatrix} \frac{11}{4} & \frac{1}{2} & -2 \\ \frac{1}{2} & 1 & -1 \\ -2 & -1 & 2 \end{bmatrix} \\
B &= (X^T X)^{-1} X^T Y \\
&= \begin{bmatrix} 2.0625 \\ -2.3750 \\ 3.2500 \end{bmatrix}
\end{aligned}
$$

The required model is

$$y = 2.0625 - 2.3750x_1 + 3.2500x_2.$$

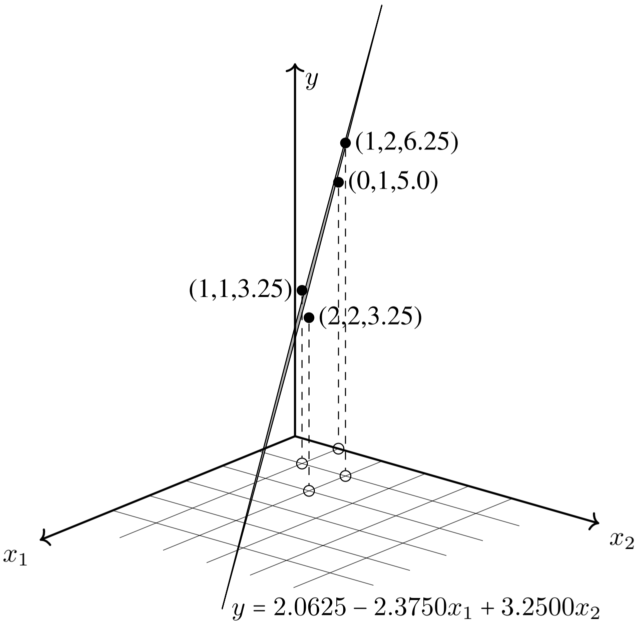


Figure 7.4: The regression plane for the data in Table 7.4

7.6 Sample questions

(a) Short answer questions

1. What are the different types of regression.
2. Is regression a supervised learning? Why?
3. Explain the ordinary least squares method for regression.
4. What are linear, multinomial and polynomial regressions.
5. If model used for regression is
$$y = a + b(x-1)^2,$$
is it a multinomial regression? If not, what type of regression is it?
6. What does the line of regression tell you?

(b) Long answer questions

1. Discuss linear regression with an example.

2. In the table below, the x_i row shows scores in an aptitude test. Similarly, the y_i row shows statistics grades. If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

Student i	1	2	3	4	5
x_i	95	85	80	70	60
y_i	85	95	70	65	70

3. Use the following data to construct a linear regression model for the auto insurance premium as a function of driving experience.

Driving experience (in years)	5	2	12	9	15	6	25	16
Monthly auto insurance premium ($)	64	87	50	71	44	56	42	60

4. Determine the regression equation by finding the regression slope coefficient and the intercept value using the following data.

x	55	60	65	70	80
y	52	54	56	58	62

5. The following table contains measurements of yield from an experiment done at five different temperature levels. The variables are y = yield and x = temperature in degrees Fahrenheit. Compute a second degree polynomial regression model to predict the yield given the temperature.

Temperature (x)	Yield (y)
50	3.0
70	2.7
80	2.6
90	2.9
100	3.3

6. An experiment was done to assess how moisture content and sweetness of a pastry product affect a tasterâĂŹs rating of the product. The following table summarises the findings.

Rating	Moisture	Sweetness
64	4	2
73	4	4
61	4	2
76	4	4
72	6	2
80	6	4
71	6	2
83	6	4
83	8	2
89	8	4
86	8	2
93	8	4
88	10	2
95	10	4
94	10	2
100	10	4

Compute a linear regression model to predict the rating of the pastry product.

7. The following data contains the Performance IQ scores (PIQ) (in appropriate scales), brain sizes (in standard units), heights (in inches) and weights (in pounds) of 15 American college students. Obtain a linear regression model to predict the PIQ given the values of the other features.

PIQ	Brain	Height	Weight
124	81.69	64.5	118
150	103.84	73.3	143
128	96.54	68.8	172
134	95.15	65.0	147
110	92.88	69.0	146
131	99.13	64.5	138
98	85.43	66.0	175
84	90.49	66.3	134
147	95.55	68.8	172
124	83.39	64.5	118
128	107.95	70.0	151
124	92.41	69.0	155
147	85.65	70.5	155
90	87.89	66.0	146
96	86.54	68.0	135

8. Use the following data to generate a linear regression model for annual salary as function of GPA and number of months worked.

Example no.	Annual salary ($)	GPA	Months worked
1	20000	2.8	48
2	24500	3.4	24
3	23000	3.2	24
4	25000	3.8	24
5	20000	3.2	48
6	22500	3.4	36
7	27500	4.0	24
8	19000	2.6	48
9	24000	3.2	36
10	28500	3.8	12

Chapter 8

Decision trees

“Decision tree learning is a method for approximating discrete valued target functions, in which the learned function is represented by a decision tree. Decision tree learning is one of the most widely used and practical methods for inductive inference.” ([4] p.52)

8.1 Decision tree: Example

Consider the following situation. Somebody is hunting for a job. At the very beginning, he decides that he will consider only those jobs for which the monthly salary is at least Rs.50,000. Our job hunter does not like spending much time traveling to place of work. He is comfortable only if the commuting time is less than one hour. Also, he expects the company to arrange for a free coffee every morning! The decisions to be made before deciding to accept or reject a job offer can be schematically represented as in Figure 8.6. This figure represents a *decision tree*[1].

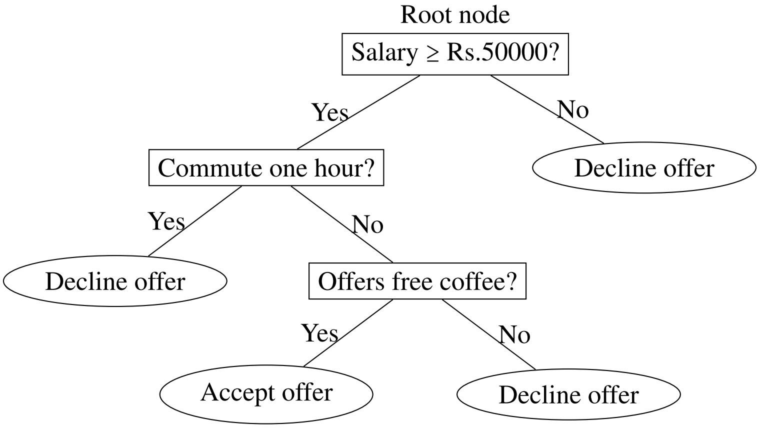


Figure 8.1: Example for a decision tree

Here, the term “tree” refers to the concept of a tree in graph theory in mathematics[2]. *In graph theory, a tree is defined as an undirected graph in which any two vertices are connected by exactly one path.* Using the conventions of graph theory, the decision tree shown in Figure 8.6 can be represented as a graph-theoretical tree as in Figure 8.2. Since a decision tree is a graph-theoretical tree, all terminology related to graph-theoretical trees can be applied to describe decision trees also. For example, in Figure 8.6, the nodes or vertices shown as ellipses are called the *leaf nodes*. All other nodes, except the root node, are called the *internal nodes*.

[1] In such diagrams, the “tree” is shown upside down with the root node at the top and all the leaves at the bottom.

[2] The term “tree” was coined in 1857 by the British mathematician Arthur Cayley (see Wikipedia).