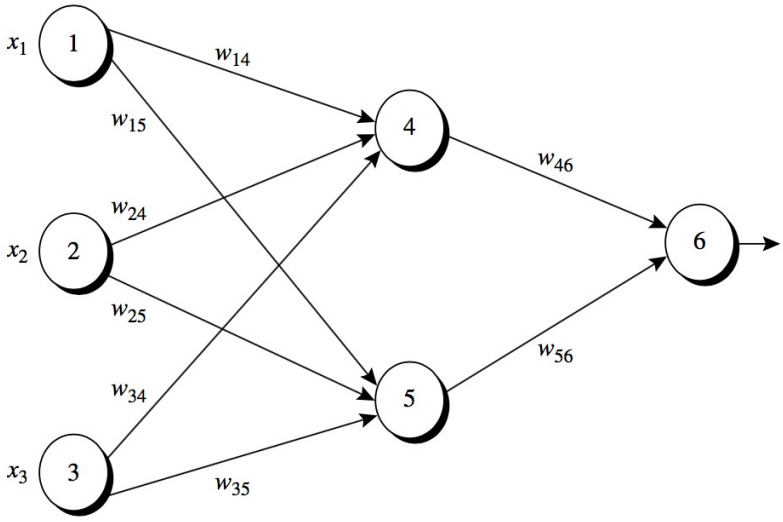
Chapter 2: Data Exploration
2.1. Given the following measurements for the variance age:

18, 22, 25, 42, 28, 43, 33, , 35, 56, 28

(a) Compute the mean and absolute deviation of age.
(b) Compute the z-score for the first four measurements

Chapter 6: Classification

6.1: Neural Net
Example 9.1 (DMCT)



Multi-layer feedforward neural network. Learning rate = 0.9; the first training tuple (1,0,1) with the class label of 1. For each node 4, 5, 6: input net = linear; activation function is the sigmoid.

Initial Input, Weight, and Bias Values

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

Calculate weights and biases updating.

6.2: Decision Tree & Naive Bayes:
The following table consists of training data from an employee database. The data has been generalized. For a given row, count represents the number of data tuples having the values for department, status, age and salary given in each row.

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let status be the class label attribute.
a) How would you modify the basic decision tree algorithm to take into account the consideration the count of each generalized data tuple?
b) Use the algorithm to construct a decision tree from the given data.
c. Given a data tuple having the values “systems”, “26---30” and “46---50K'. What would a naive Bayesian classification of the status for the tuple be?

6.3. Support Vector Machines
What is a margin in Support-Vector-Machine (SVM)? What are support vectors? The optimization problem of SVM is as follows:

$$\min_{\xi,W,b} \frac{1}{2} W^T W + C\left(\sum_{i=1}^{N} \xi_i\right)$$

Subject to:

$$Y_i(W^T X_i + b) \geq 1 - \xi_i, i = 1, \ldots, N$$
$$\xi_i > 0, i = 1, \ldots, N$$

What does the first term ½ W^T W minimize? What are ξ_i?

Chapter 7: Clustering

7.1. Agglomerative Clustering
Perform the agglomerative clustering of one-dimensional set of points 1,4, 9, 16, 25, 36 in the following cases:

a. Clusters are merged by single-linkage strategy.

b. Clusters are merged by complete-linkage strategy.

For each case, draw the corresponding dendrogram.

Chapter 13: Mining Streaming Data

13.1. Bloom Filter

a) What types of error is possible in Bloom Filter: FP, FN

c) Suppose we have an array of 2 billion bits, 5 hash functions, we insert 300 million elements, calculate the probability of error.

Chapter 14: Recommendation Systems

14.1: Locality Sensitive Hashing

Suppose that H is a (d1,d2, p1, p2)-sensitive family, construct H' consisting of b functions from H using OR construction, prove that H' is (d1, d2, 1-(1-p1)^b, 1-(1-p2)^b)-sensitive.

Solution:

6.1

Forward pass:

$$I_4 = w_{14}x_1 + w_{24}x_2 + w_{34}x_3 + \theta_4$$
$$O_4 = \sigma(I_4) = \frac{1}{1+e^{-I_4}}$$
$$I_5 = w_{15}x_1 + w_{25}x_2 + w_{35}x_5 + \theta_5$$
$$O_5 = \sigma(I_5)$$
$$I_6 = w_{46}O_4 + w_{56}O_5 + \theta_6$$
$$O_6 = \sigma(I_6)$$

Net error:

$$E = 1/2(O_6 - 1)^2$$

Backward pass: update parameters using gradient decent with error rate $\eta = 0.9$

Update w_14 and w_46

$$w_{46}^{(new)} = w_{46} - \eta \frac{\partial E}{\partial w_{46}}$$
$$w_{14}^{(new)} = w_{14} - \eta \frac{\partial E}{\partial w_{14}}$$

Here, we have:

$$\frac{\partial E}{\partial w_{46}} = \frac{\partial E}{\partial O_6}\frac{\partial O_6}{\partial w_{46}}$$
$$\frac{\partial E}{\partial w_{14}} = \frac{\partial E}{\partial O_6}\frac{\partial O_6}{\partial O_4}\frac{\partial O_4}{\partial w_{14}}$$

In a general neural net:

$$\frac{\partial E}{\partial w_{ij}} = [\sum_k \frac{\partial E}{\partial O_k}\frac{\partial O_k}{\partial O_j}]\frac{\partial O_j}{\partial w_{ij}}$$

where k runs over all the node k that there is a link between node j and node k.

If we set $grad[j] = \frac{\partial E}{\partial O_j}$, then $grad[j] = \sum_k grad[k]\frac{\partial O_k}{\partial O_j}$

The gradient is back-propagated through the backward pass.

Note: it is important to know the procedure how to update weights and biases in a general neural net. In general, we can have different network structure, activation functions and net error function. Examples are CBOW, Skip-gram.

6.2

(a) Two modifications:
- The count of each tuple must be integrated into the calculation of the attribute selection measure.
- Take the count into consideration to determine the most common class among the tuples.

(b) The resulting tree is:

(salary = 26K...30K: junior
= 31K...35K: junior
= 36K...40K: senior
= 41K...45K: junior
= 46K...50K (department = secretary: junior
= sales: senior
= systems:
= marketing: senior)
= 66K...70K: senior)

If we use salary as the root node, the only non-pure child node is 46K...50K (Senior 40, Junior 23). If we use age as the root node, the only non-pure child node is 31...35 (Senior 35, Junior 44). Note that pure nodes have zero entropy and zero GINI. Comparing InfoGain or GINI_split of salary and age turns into comparing Entropy or GINI of the two non-pure child nodes weighted by their sizes.

Another note is that after choosing salary as the root node, the next child node can be selected is department. There is one empty partition corresponding to department=secretary if we use department as the next splitting attribute. In this case, we apply the rule on the line 12 in the algorithm (slide 14, basic classification).
(c) P(X|senior) = 0; P(X|junior) = 0.018. Thus, a naive Bayesian classification predicts “junior”

Chapter 7: Clustering

7.1 (a)

The mean absolute deviation of *age* is 8.8, which is derived as follows.

$$
\begin{aligned}
m_f &= \frac{1}{n}(x_{1f} + \cdots + x_{nf}) = \frac{1}{10}(18 + 22 + 25 + 42 + 28 + 43 + 33 + 35 + 56 + 28) = 33 \\
s_f &= \frac{1}{n}(|x_{1f} - m_f| + \cdots + |x_{nf} - m_f|) \\
&= \frac{1}{10}(|18 - 33| + |22 - 33| + |25 - 33| + |42 - 33| + |28 - 33| + |43 - 33| + |33 - 33| + |35 - 33| \\
&\quad + |56 - 33| + |28 - 33|) \\
&= 8.8
\end{aligned}
$$

(b) Compute the z-score for the first four measurements.

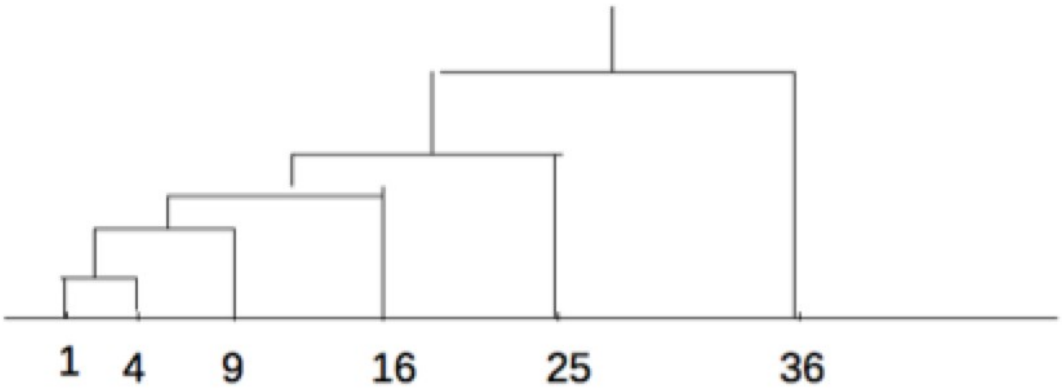
According to the z-score computation formula,

$$z_{if} = \frac{x_{if} - m_{if}}{s_f}.$$

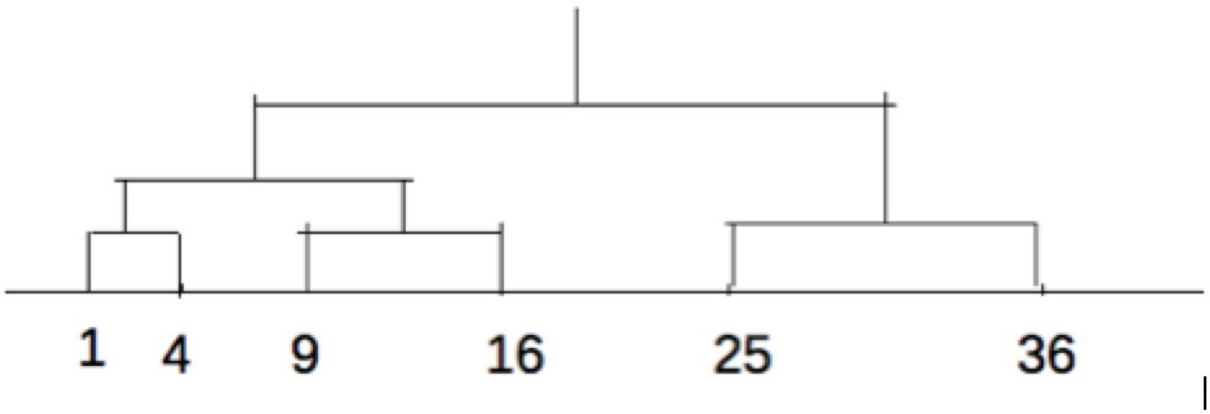
We have

$$
\begin{aligned}
z_{1f} &= \frac{18 - 33}{8.8} = -1.70 \\
z_{2f} &= \frac{22 - 33}{8.8} = -1.25 \\
z_{3f} &= \frac{25 - 33}{8.8} = -0.91 \\
z_{4f} &= \frac{42 - 33}{8.8} = 1.02
\end{aligned}
$$

7.2 Perform agglomerative clustering
Clusters are merged by single-linkage strategy:



Clusters are merged by complete-linkage strategy:


1
4
9
16
25
36