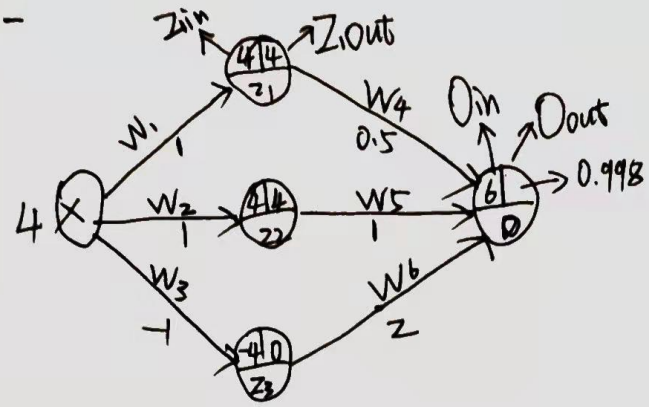
1.1

1.1-



1.1-a $\hat{y} = \frac{1}{1+e^{-O_{in}}}$ # $O_{in} = 4 \times 0.5 + 4 + 0 = 6$

$= \frac{1}{1+e^{-6}} = 0.998$

1.1-b $E(y, \hat{y}) = (y-\hat{y})^2 = (0-0.998)^2 = 0.996$

1.1.c $\frac{\partial E_{total}}{\partial O_{in}} = \frac{\partial E_{total}}{\partial O_{out}} \cdot \frac{\partial O_{out}}{\partial O_{in}} =$

$= -2(y - O_{out}) \times O_{out} \times (1 - O_{out})$

$= -2(0 - 0.998) \times 0.998 \times (1 - 0.998)$

$= 0.004$

$\frac{\partial E_{total}}{\partial W_4} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial W_4}$

$= 0.004 \cdot Z_1 O_{out} = 0.004 \times 4 = 0.016$

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial w_5} = 0.004 \times Z_{2out}$$

$$= 0.004 \times 4 = 0.016$$

$$\frac{\partial E_{total}}{\partial w_6} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial w_6} = 0.004 \times Z_{3out}$$

$$= 0.004 \times 0 = 0$$

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial Z_{1out}} \cdot \frac{\partial Z_{1out}}{\partial Z_{1in}} \cdot \frac{\partial Z_{1in}}{\partial w_1}$$

$$= 0.004 \times w_4 \times 1 \times X$$

$$= 0.004 \times 0.5 \times 1 \times 4$$

$$= 0.008$$

$$\frac{\partial E_{total}}{\partial w_2} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial Z_{2out}} \cdot \frac{\partial Z_{2out}}{\partial Z_{2in}} \cdot \frac{\partial Z_{2in}}{\partial w_2}$$

$$= 0.004 \times w_5 \times 1 \times X$$

$$= 0.004 \times 1 \times 1 \times 4$$

$$= 0.016$$

$$\frac{\partial E_{total}}{\partial w_3} = \frac{\partial E_{total}}{\partial O_{in}} \cdot \frac{\partial O_{in}}{\partial Z_{3out}} \cdot \frac{\partial Z_{3out}}{\partial Z_{3in}} \cdot \frac{\partial Z_{3in}}{\partial w_3}$$

$$= 0.004 \times w_6 \times 0 \times X$$

$$= 0.004 \times 2 \times 0 \times 4$$

$$= 0$$

$\therefore \Delta E_w = [0.008, 0.016, 0, 0.016, 0.016, 0]$

1.1-d

$W_{update} = [1, 1, -1, 0.5, 1, 2]^T - 1.0 \times [0.008, 0.016, 0, 0.016, 0.016, 0]$

$= [0.992, 0.984, -1, 0.484, 0.984, 2]$

$\hat{y} = \frac{1}{1+e^{-O_{in}}}$

$O_{in} = (4 \times 0.992) \times 0.484 + (4 \times 0.984) \times 0.984 + (1 \times 2) \times 2$

$= 3.968 \times 0.484 + 3.936 \times 0.984$

$= 1.920 + 3.876 = 5.796$

$\hat{y} = \frac{1}{1+e^{-5.796}} = 0.997$

$E(y, \hat{y}) = (0 - 0.997)^2 = 0.994$

i. The new predicted output is 0.997

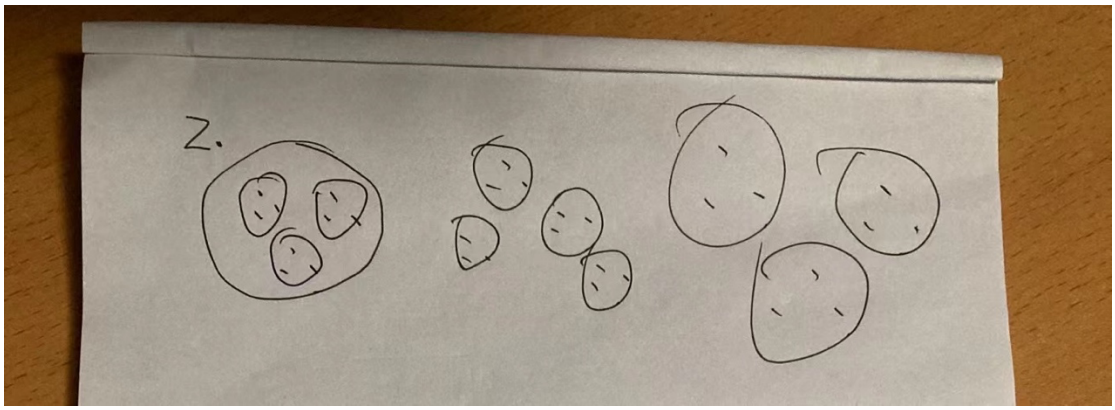
the loss is 0.994

1.1-e

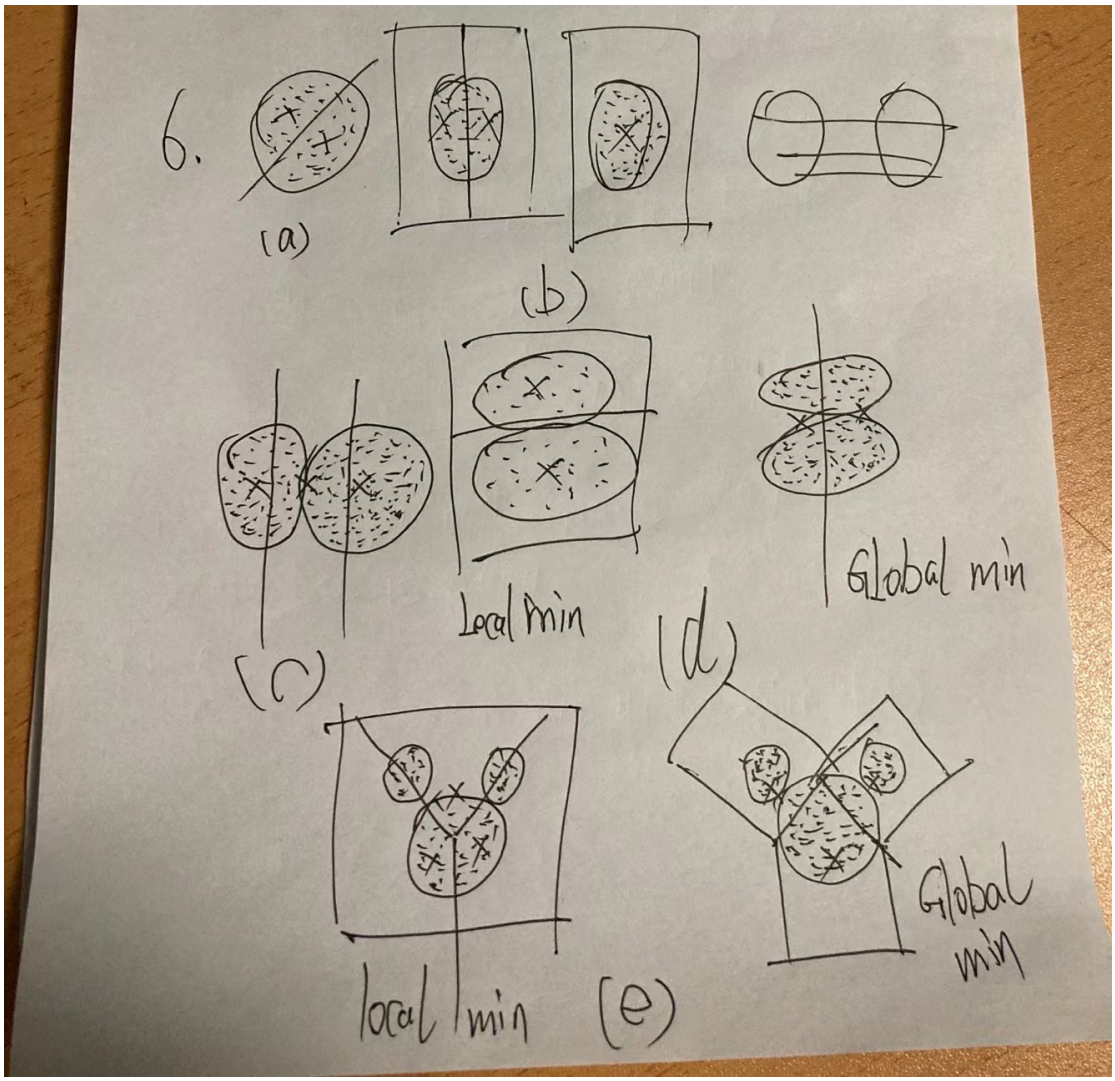
(d) is smaller than the loss value of (e)

1.2

2:



6:



6-a:

In theory, there are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle. This line can make any angle $0^\circ \leq \theta \leq 180^\circ$ with the x axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

6-b:

If you start with initial centroids that are real points, you will necessar- ily get this solution because of the restriction that the circles are more than one radius apart. Of course, the bisector could have any angle, as above, and it could be the other circle that is split. All these solutions have the same globally minimal error.

6-c:

The three boxes show the three clusters that will result in the realistic case that the initial centroids are actual data points.

6-d:

In both case, the rectangles show the clusters. In the first case, the two clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.

6-e:

For the solution shown in the top figure, the two top clusters are en- closed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle. (The two smaller clusters in the drawing are supposed to be symmetrical.) I believe that the second solution—suggested by a student—is also possible, although it is a lo- cal minimum and might rarely be seen in practice for this configuration of points. Note that while the two pie shaped cuts out of the larger cir- cle are shown as meeting at a point, this is not necessarily the case—it depends on the exact positions and sizes of the circles. There could be a gap between the two pie shaped cuts which is filled by the third (larger) cluster. (Imagine the small circles on opposite sides.) Or the boundary between the two pie shaped cuts could actually be a line segment.

11:

1. If the SSE of one attribute is low for all clusters, then the variable is essentially a constant and of little use in dividing the data into groups.
2. if the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster
3. If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise.
4. If the SSE of an attribute is relatively high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define the cluster. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes, but in any case, it means that this attribute does not help define the cluster.

5. The idea is to eliminate attributes that have poor distinguishing power between clusters, i.e., low or high SSE for all clusters, since they are useless for clustering. Note that attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes (perhaps because of their scale) since they introduce a lot of noise into the computation of the overall SSE

12-a:

The leader algorithm requires only a single scan of the data and is thus more computationally efficient since each object is compared to the final set of centroids at most once. Although the leader algorithm is order dependent, for a fixed ordering of the objects, it always produces the same set of clusters. However, unlike K-means, it is not possible to set the number of resulting clusters for the leader algorithm, except indirectly. Also, the K-means algorithm almost always produces better quality clusters as measured by SSE

12-b:

Use a sample to determine the distribution of distances between the points. The knowledge gained from this process can be used to more intelligently set the value of the threshold.

The leader algorithm could be modified to cluster for several thresholds during a single pass.