CISC 271 Class 20

Classification – Linear Separability

Text Correspondence: $\sim$

Main Concepts:

- *Binary classification: 2 classes of data vectors*
- *Linear separability: mathematical separation*
- *Hyperplane: implemented as* weight *vector and* bias *scalar*

Sample Problem, Machine Inference: How does k-means clustering separate data vectors into two classes?

For this class, and for the rest of the course, we will concentrate on the problem of supervised binary classification. By “supervised” we mean that there is a data attribute that tells us the label of a data vector. By “binary” we mean that each data vector belongs to exactly one of two classes, which are non-intersecting sets. By “classification” we mean that a definite answer is sought, which is that a given data vector belongs to either one class or the other.

For this lecture, in the context of binary classification, we will write the labels and classes as “+1” and “-1”. For subscripts, we will use 1 to denote class +1 and “2” to denote class -1.

20.1 Separation of Two Clusters

Suppose that, in the context of binary cluster analysis, we know the centroid of the cluster with label +1 as $\vec{g}_1$ and the centroid of the cluster with label -1 as $\vec{g}_2$. What is the geometry of the clustering?

By this, we mean what is the shape of the boundary between the clusters. We will explore this by considering a data vector $\vec{t}$ that is on the boundary, that is, a situation in which $\vec{t}$ could be classified as belonging to either cluster +1 or to cluster -1.

Using a slight modification of the definition in Definition 19.2, a data vector $\vec{t}$ would be assigned to cluster +1 if it is closer to centroid $\vec{g}_1$, so

$$\|\vec{t} - \vec{g}_1\|^2 \leq \|\vec{t} - \vec{g}_2\|^2 \tag{20.1}$$

The data vector $\vec{t}$ would be assigned to cluster -1 if it is closer to centroid $\vec{g}_2$, so

$$\|\vec{t} - \vec{g}_2\|^2 \leq \|\vec{t} - \vec{g}_1\|^2 \tag{20.2}$$

Combining Equation 20.1 and Equation 20.2, the data vector $\vec{t}$ is on the boundary between cluster +1 and cluster -1 if and only if it is equidistant from the respective centroids, which is the requirement that

$$\|\vec{t} - \vec{g}_2\|^2 = \|\vec{t} - \vec{g}_1\|^2 \tag{20.3}$$

Replacing the squared norm of a vector with the dot product, Equation 20.3 implies that

$$\begin{aligned}
[\vec{t} - \vec{g}_2] \cdot [\vec{t} - \vec{g}_2] &= [\vec{t} - \vec{g}_1] \cdot [\vec{t} - \vec{g}_1] \\
\Rightarrow\ \vec{t} \cdot \vec{t} + \vec{g}_2 \cdot \vec{g}_2 - 2\vec{g}_2 \cdot \vec{t} &= \vec{t} \cdot \vec{t} + \vec{g}_1 \cdot \vec{g}_1 - 2\vec{g}_1 \cdot \vec{t} \\
\Rightarrow\ \vec{g}_2 \cdot \vec{g}_2 - 2\vec{g}_2 \cdot \vec{t} &= \vec{g}_1 \cdot \vec{g}_1 - 2\vec{g}_1 \cdot \vec{t} \\
\Rightarrow\ 2\vec{g}_1 \cdot \vec{t} - 2\vec{g}_2 \cdot \vec{t} &= \vec{g}_1 \cdot \vec{g}_1 - \vec{g}_2 \cdot \vec{g}_2 \\
\Rightarrow\ \vec{g}_1 \cdot \vec{t} - \vec{g}_2 \cdot \vec{t} &= [\vec{g}_1 \cdot \vec{g}_1 - \vec{g}_2 \cdot \vec{g}_2]/2 \\
\Rightarrow\ [\vec{g}_1 - \vec{g}_2] \cdot \vec{t} &= [\vec{g}_1 \cdot \vec{g}_1 - \vec{g}_2 \cdot \vec{g}_2]/2 \\
\Rightarrow\ [\vec{g}_1 - \vec{g}_2] \cdot \vec{t} &= [\vec{g}_1 - \vec{g}_2] \cdot \frac{[\vec{g}_1 + \vec{g}_2]}{2}
\end{aligned} \tag{20.4}$$

We can abbreviate:

- the difference vector in Equation 20.4 as $\vec{m} \stackrel{\text{def}}{=} \vec{g}_1 - \vec{g}_2$
- the vector average, on the right-hand side of Equation 20.4, as $\vec{h} \stackrel{\text{def}}{=} \dfrac{[\vec{g}_1 + \vec{g}_2]}{2}$
- the scalar values β and b as $\vec{m} \cdot \vec{h} = \beta$, or as $-\vec{m} \cdot \vec{h} = b$

With these abbreviations, we can write Equation 20.4 as

$$\vec{m} \cdot \vec{t} = \vec{m} \cdot \vec{h} \quad \text{or as} \quad \vec{m}^T\vec{t} = \beta \quad \text{or as} \quad \vec{m}^T\vec{t} + b = 0 \tag{20.5}$$

From prerequisite material, we recognize Equation 20.5 as the implicit equation of a line in 2-space or a plane in 3-space. To avoid using terms specific to the size of the space, we will refer to Equation 20.5 as the implicit equation of a *hyperplane*.

A vector $\vec{t}$ is on a hyperplane if and only if it satisfies Equation 20.5. One such reference vector on the hyperplane is $\vec{h}$, which is the average of the centroids $\vec{g}_1$ and $\vec{g}_2$.

We can understand the way this hyperplane works by using a hypothetical example. Suppose that a new vector $\vec{x}$ is at least as close to the centroid $\vec{g}_1$ of cluster +1 as it is to the centroid $\vec{g}_2$ of cluster -1. This implies that the squared distance from $\vec{x}$ to $\vec{g}_2$ is greater than the squared distance from $\vec{x}$ to $\vec{g}_1$. Using this, and the fact that adding or subtracting a quantity from both sides of an inequality does not change the inequality, we find that

$$\begin{aligned}
\|\vec{x}-\vec{g}_2\|^2 &> \|\vec{x}-\vec{g}_1\|^2 \\
[\vec{x}-\vec{g}_2]\cdot[\vec{x}-\vec{g}_2] &\geq [\vec{x}-\vec{g}_1]\cdot[\vec{x}-\vec{g}_1] \\
\vec{x}\cdot\vec{x}+\vec{g}_2\cdot\vec{g}_2-2\vec{g}_2\cdot\vec{x} &\geq \vec{x}\cdot\vec{x}+\vec{g}_1\cdot\vec{g}_1-2\vec{g}_1\cdot\vec{x} \\
\vec{g}_2\cdot\vec{g}_2-2\vec{g}_2\cdot\vec{x} &\geq \vec{g}_1\cdot\vec{g}_1-2\vec{g}_1\cdot\vec{x} \\
2\vec{g}_1\cdot\vec{x}-2\vec{g}_2\cdot\vec{x} &\geq \vec{g}_1\cdot\vec{g}_1-\vec{g}_2\cdot\vec{g}_2 \\
[\vec{g}_1-\vec{g}_2]\cdot\vec{x} &\geq [\vec{g}_1\cdot\vec{g}_1-\vec{g}_2\cdot\vec{g}_2]/2 \\
\vec{m}\cdot\vec{x} &\geq \vec{m}\cdot\vec{h} \\
\vec{m}^T\vec{x} &\geq \beta \\
\vec{m}^T\vec{x}+b &\geq 0 \qquad (20.6)
\end{aligned}$$

One common interpretation of Equation 20.6 is that there is an *orientation* of the hyperplane. Any vector on the “positive” side has a dot product with the normal vector $\vec{m}$ that is greater than the hyperplane scalar β, and any vector on the “negative” side has a dot product with the normal vector $\vec{m}$ that is less than the hyperplane scalar β. This means that Equation 20.6 can be used to determine whether a new vector is in cluster +1 or in cluster -1, using only a dot product and either a comparison of two scalar values, or a subtraction and comparison to zero. Using this, we can perform binary classification of a vector $\vec{x}$ by linear separation as

$$\begin{aligned}
\vec{x}\in\text{class +1} &\equiv \vec{m}^T\vec{x}+b\geq 0 \\
\vec{x}\in\text{class –1} &\equiv \vec{m}^T\vec{x}+b< 0
\end{aligned} \qquad (20.7)$$

We can summarize our mathematical findings as:

- Binary clustering uses a hyperplane to separate the clusters
- One reference vector on the hyperplane is the mid-point of the centroids of the clusters
- One normal vector for the hyperplane is the vector difference between the centroids
- The class of any vector is determined by which “side” of the hyperplane the vector lies

We can now re-visit the Iris data set with this new information. Figure 20.1 shows the separating hyperplane from the k-means clustering. One anomalous data vector is on the class -1 side of the hyperplane, but was labeled by a human expert as corresponding to a type P Iris plant.

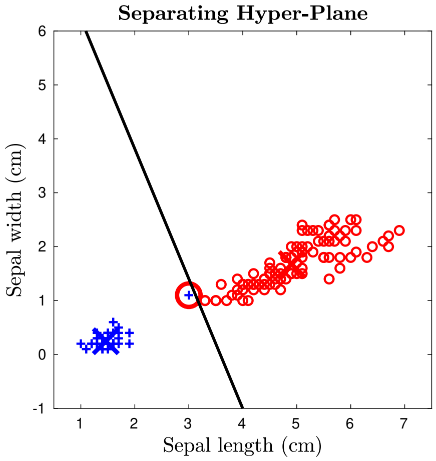

Figure 20.1: Fisher's Iris data set, expertly labeled. The separating hyperplane from k-means clustering is shown as a line and the vector mis-classified by k-means clustering is circled.

We can see that k-means clustering uses a separating hyperplane to perform classification. For the Iris data set, it is visually apparent that there are many better choices of a hyperplane that completely separates the two clusters of data vectors.

20.2 Separation of Three or More Clusters

The k-means algorithm is intended to partition data into k clusters, where $k \geq 2$. We can understand the geometry by combining the definition of a cluster, from Definition 19.2, with the observation that arises from Equation 20.6.

Suppose that data vectors are separated into three clusters – sets S_1, S_2, and S_3 – that have corresponding centroids $\vec{g}_1$, $\vec{g}_2$, and $\vec{g}_3$. A new vector $\vec{x}$ is in S_1 if it is closer to $\vec{g}_1$ than to $\vec{g}_2$, and also closer to $\vec{g}_1$ than to $\vec{g}_3$. An example of three centroids in the plane is provided in Figure 20.2(A).

Cluster S_1 is separated from cluster S_2 by a hyperplane; let us refer to this as hyperplane H_{12}. There is also a hyperplane that separates S_1 from S_3 which we can refer to as H_{13}, and a hyperplane that separates S_2 from S_3 which we can refer to as H_{23}. Three centroids, and these three hyperplanes, are illustrated in Figure 20.2.

Figure 20.2 show additional information. A new data vector $\vec{x}$ is classified into set S_1 if it is closer to $\vec{g}_1$ than it is to $\vec{g}_2$ or to $\vec{g}_3$; this implies that it is on one side of both H_{12} and H_{13}, which is

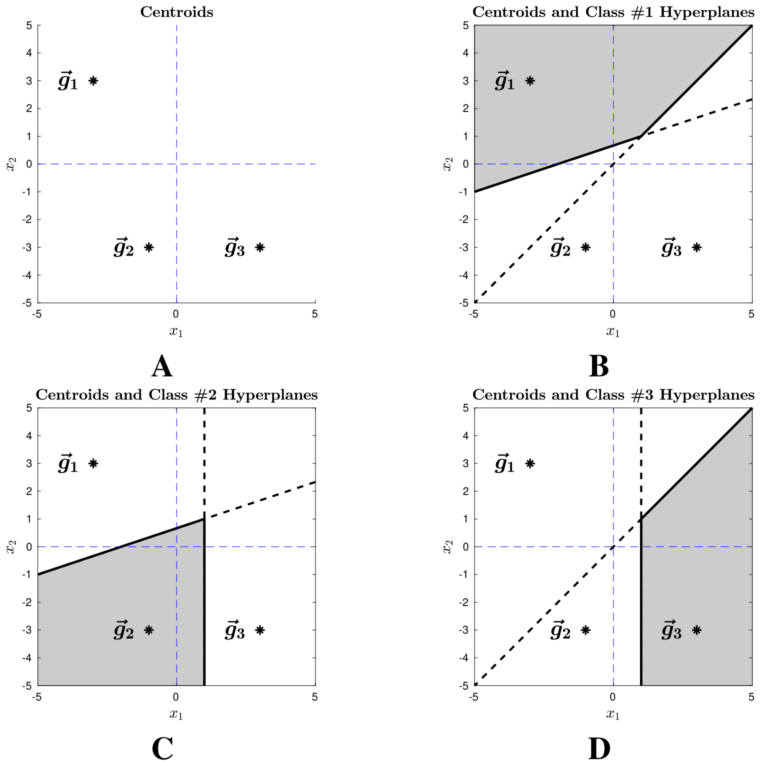

Figure 20.2: Three sets defined by k-means clustering, with centroids and separating hyperplanes. (A) The centroids $\vec{g}_1$, $\vec{g}_2$, and $\vec{g}_3$. (B) Shaded region indicates vectors in cluster S_1; solid lines show the set boundary and dashed lines show the infinite extension of the hyperplanes. (C) Cluster set S_2, its boundary, and extended hyperplanes. (D) Cluster set S_3, its boundary, and extended hyperplanes.

shown as the shaded region in Figure 20.2(B). Using similar reasoning, the region for classifying a vector in set S_2 is shown as the shaded region in Figure 20.2(C), and for S_3 in Figure 20.2(D).

We now understand that k-means clustering divides a data space into “cells” with linear edges, which are the separating hyperplanes of the clusters. How can we easily compute the assignment of a vector $\vec{x}$ into exactly one of the sets S_1, S_2, and S_3? We can begin by reasoning how to determine whether $\vec{x}$ is in set S_1.

We originally defined a vector $\vec{x}$ as being clustered into set S_1 if and only if the Euclidean distance from $\vec{x}$ to the centroid $\vec{g}_1$ is less than or equal to the Euclidean distance of $\vec{x}$ to every other set. Mathematically, this definition is

$$\vec{x} \in S_1 \overset{\text{def}}{=} (\|\vec{x}-\vec{g}_1\| \leq \|\vec{x}-\vec{g}_2\|) \wedge (\|\vec{x}-\vec{g}_1\| \leq \|\vec{x}-\vec{g}_3\|) \tag{20.8}$$

We can, in a way that is equivalent to Equation 20.8, define a vector $\vec{x}$ as being clustered into set S_1 if and only if it is on the "positive" side of the hyperplane that separates S_1 and S_2, and if it is also on the "positive" side of the hyperplane that separates S_1 and S_3. To make this definition easier, let us introduce some abbreviations.

Let us define the separating hyperplane H_{ij}, which separates set S_i from S_j, as the ordered pair of the weight vector $\vec{m}_{ij}$ and the bias scalar b_{ij}:

$$H_{ij} \overset{\text{def}}{=} (\vec{m}_{ij}, b_{ij}) \tag{20.9}$$

Using Equation 20.9, we would define $\vec{x}$ as being clustered into set S_1 if and only if both of these conditions is true:

$$\vec{x} \in S_1 \overset{\text{def}}{=} \big((\vec{m}_{12}^T\vec{x} + b_{12} \geq 0) \wedge (\vec{m}_{13}^T\vec{x} + b_{13} \geq 0)\big) \tag{20.10}$$

We can gather the terms of Equation 20.10 as a matrix M_1 and a bias vector $\vec{b}_1$, which gives us

$$\begin{aligned} M_1 &\overset{\text{def}}{=} \begin{bmatrix} \vec{m}_{12}^T \\ \vec{m}_{13}^T \end{bmatrix} \\ \vec{b}_1 &\overset{\text{def}}{=} \begin{bmatrix} b_{12} \\ b_{13} \end{bmatrix} \end{aligned} \tag{20.11}$$

Using the definition of Equation 20.11, we can write Equation 20.10 as

$$\vec{x} \in S_1 \equiv \left(M_1\vec{x} + \vec{b}_1 \geq \vec{0}\right) \tag{20.12}$$

The operator "$\geq$" in Equation 20.12 is generally understood to evaluate to true if and only if the inequality is true for every entry of the vectors that are compared.

We can similarly formulate the clustering of a vector $\vec{x}$ into the set S_2 by using abbreviations and the vector inequality operator. Because of the change of sign for the hyperplane H_{12} that separates set S_1 from set S_2, we can write the clustering for set S_2 as

$$\begin{aligned} M_2 &\overset{\text{def}}{=} \begin{bmatrix} -\vec{m}_{12}^T \\ \vec{m}_{23}^T \end{bmatrix} \\ \vec{b}_2 &\overset{\text{def}}{=} \begin{bmatrix} -b_{12} \\ b_{23} \end{bmatrix} \\ \vec{x} \in S_2 &\equiv \left(M_2\vec{x} + \vec{b}_2 \geq \vec{0}\right) \end{aligned} \tag{20.13}$$

The clustering for set S_2 can be defined in an analogous fashion. These inequalities in linear algebra provide us with a fast and concise way to write and think about k-means clustering.

20.3 Hyperplane Specification With a Unit Normal Vector

Sometimes we will want to know the *distance* of a data vector $\vec{x}$ to a hyperplane $\mathbb{H}$. We can do this computation by transforming the non-zero normal vector $\vec{m}$ to a unit vector $\vec{n}$ and transforming the bias scalar b to a scalar a. Recall that a hyperplane can be specified using a non-zero normal vector $\vec{m}$ and a vector $\vec{h}$ that is on the hyperplane. From Equation 20.6, the condition for a vector $\vec{t}$ to be on $\mathbb{H}$ was the equation $(\vec{m} \cdot \vec{t}) = (\vec{m} \cdot \vec{h})$ or $(\vec{m} \cdot \vec{t} - \vec{m} \cdot \vec{h}) = 0$. A unit vector $\vec{n}$ that is in the same direction as $\vec{m}$ can be written as

$$\vec{n} = \frac{\vec{m}}{\|\vec{m}\|} \tag{20.14}$$

Dividing the governing equation for $\vec{t}$ by $\|\vec{m}\|$, and using Equation 20.14, gives

$$\begin{array}{rrcl}
 & (\vec{m} \cdot \vec{t} - \vec{m} \cdot \vec{h}) & = & 0 \\
\equiv & (\vec{m} \cdot \vec{t} - \vec{m} \cdot \vec{h})/\|\vec{m}\| & = & 0/\|\vec{m}\| \\
\equiv & \dfrac{(\vec{m} \cdot \vec{t})}{\|\vec{m}\|} - \dfrac{(\vec{m} \cdot \vec{h})}{\|\vec{m}\|} & = & 0 \\
\equiv & \dfrac{\vec{m}}{\|\vec{m}\|} \cdot \vec{t} - \dfrac{(\vec{m} \cdot \vec{h})}{\|\vec{m}\|} & = & 0 \\
\equiv & \dfrac{\vec{m}}{\|\vec{m}\|} \cdot \vec{t} - \dfrac{\beta}{\|\vec{m}\|} & = & 0 \\
\equiv & \vec{n} \cdot \vec{t} + c & = & 0 \\
\text{where} & c & = & -\dfrac{\beta}{\|\vec{m}\|}
\end{array} \tag{20.15}$$

We can combine Equation 20.14 and Equation 20.15 to define a hyperplane $\mathbb{H}$ using a unit normal vector $\vec{n}$ and a bias scalar c. The transformations are

$$\begin{array}{rcl}
\vec{n} & = & \dfrac{\vec{m}}{\|\vec{m}\|} \\
c & = & \dfrac{b}{\|\vec{m}\|}
\end{array} \tag{20.16}$$