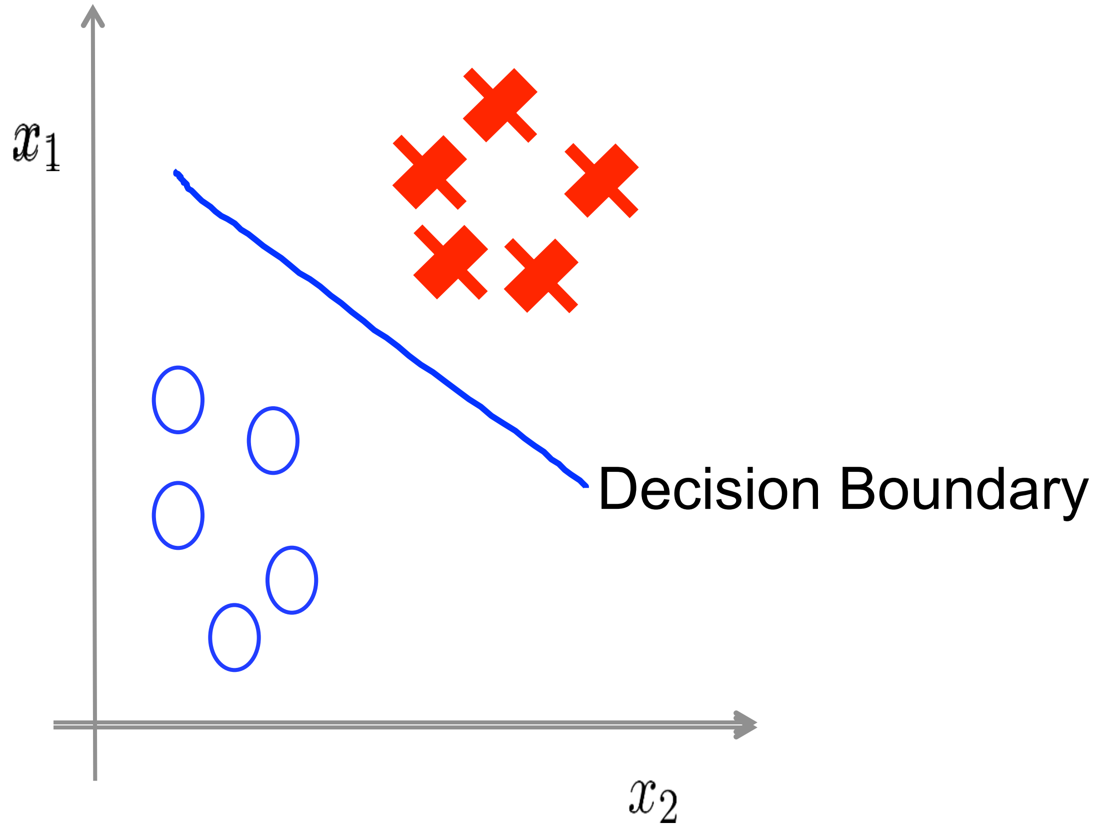
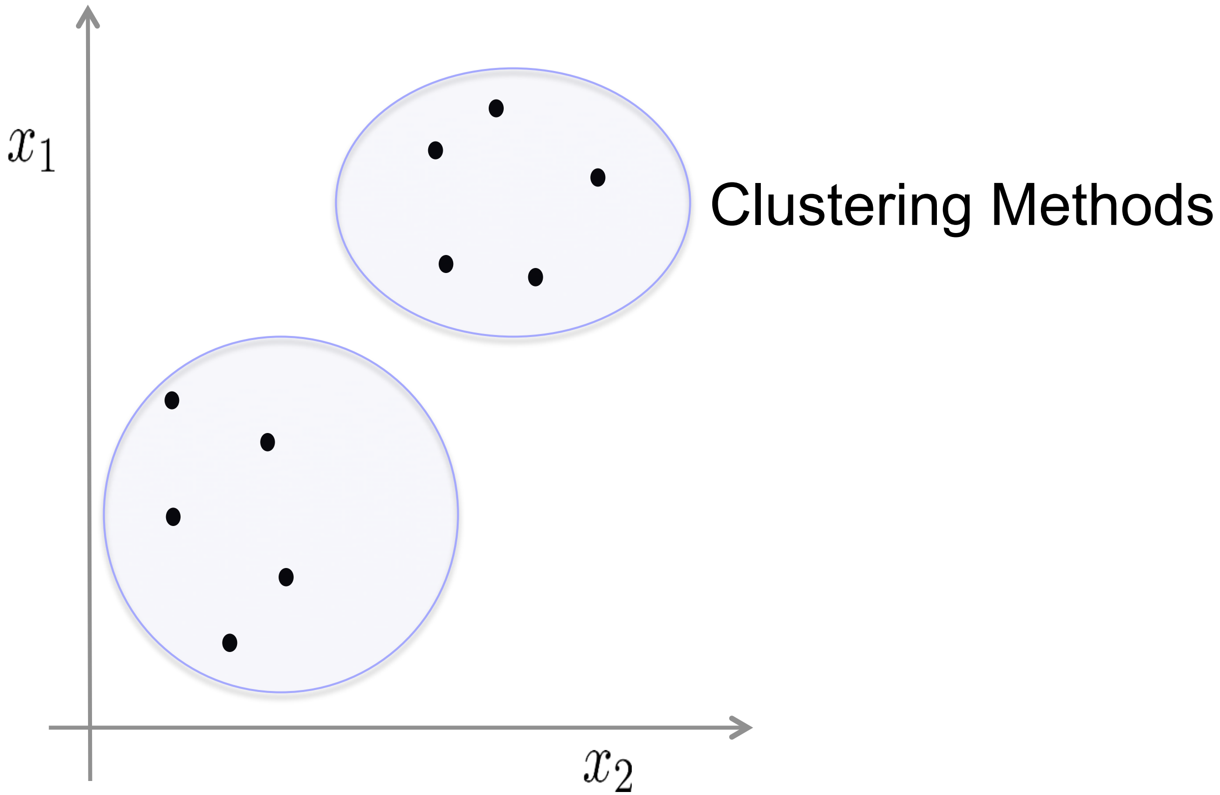
Chapter 13:
K-Means Clustering

Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Unsupervised learning



Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

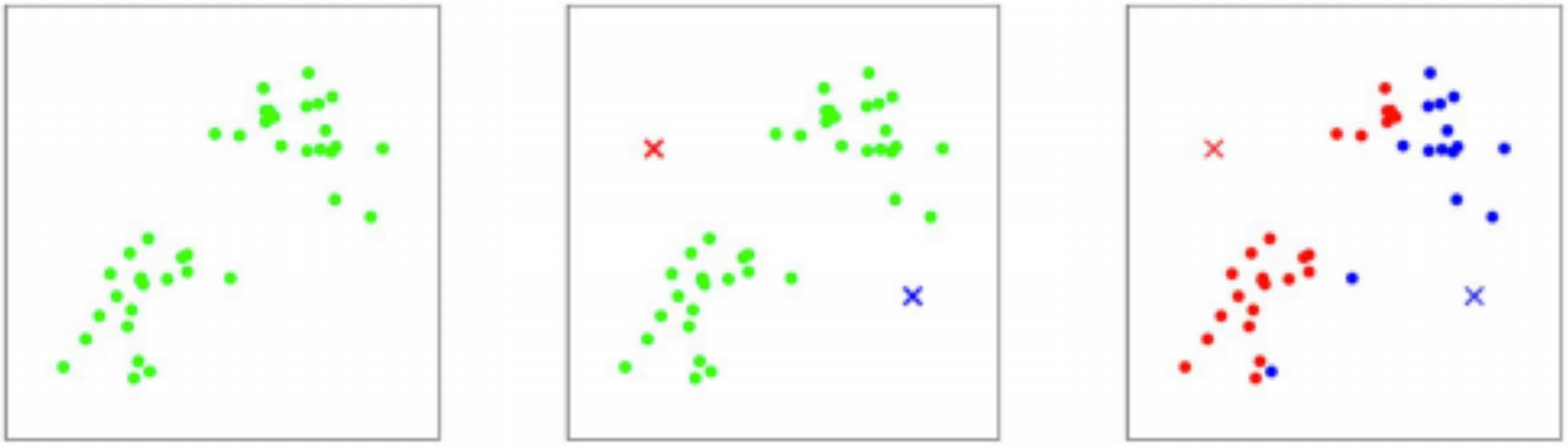
K-Means Clustering

- Say you are given a data set where each observed example has a set of features, but has no labels.
- Labels are an essential ingredient to a supervised algorithm like Support Vector Machines, which learns a hypothesis function to predict labels given features.
- So we can't run supervised learning. What can we do?
- One of the most straightforward tasks we can perform on a data set without labels is to find groups of data in our dataset which are similar to one another -- what we call clusters.

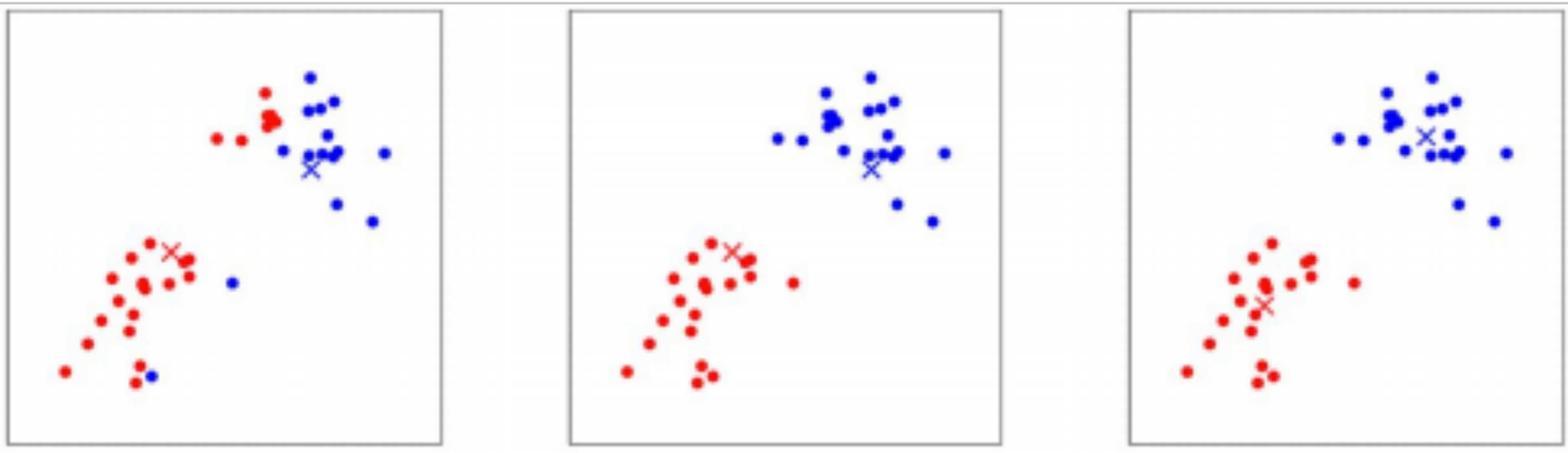
K-Means Clustering

- K-Means is one of the most popular "clustering" algorithms.
- K-means stores k centroids that it uses to define clusters.
- A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.
- How does K-means work?

K-Means Clustering



K-Means Clustering



K-Means Algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

K-Means Algorithm

- In the clustering problem, we are given a training set $x^{(1)},...,x^{(m)}$ and want to group the data into a few "clusters."
- Here, we are given feature vectors for each data point $x^{(i)}$ as usual; but no labels $y^{(i)}$.
- Our goal is to predict a label $c^{(i)}$ for each data point. The k-means clustering algorithm is as follows:

K-Means Algorithm

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

   For every i, set
$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

K-Means Algorithm

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

}

```
X = [randn(100,2)+ones(100,2);...
     randn(100,2)-ones(100,2)];
figure(1)
plot(X(:,1),X(:,2),'r+');

[CENTS, DAL] = k_means(X, 2);

figure(2)
hold on;
K=2;
for i = 1:K
PT = X(DAL(:,K+1) == i,:);                          % Find points of each
cluster
if i==1
    a='r+';
else
    a='b+';
end
plot(PT(:,1),PT(:,2),a,'LineWidth',2);
plot(CENTS(:,1),CENTS(:,2),'*k','LineWidth',7);
end
```

```
function [CENTS, DAL] = k_means(F, K)

CENTS = F( ceil(rand(K,1)*size(F,1)) ,:);        % Cluster Centers
DAL   = zeros(size(F,1),K+2);                    % Distances and Labels
hold on; plot(CENTS(:,1),CENTS(:,2),'*k','LineWidth',7);
for n = 1:100

   for i = 1:size(F,1)
      for j = 1:K
         DAL(i,j) = norm(F(i,:) - CENTS(j,:));
      end
      [Distance, CN] = min(DAL(i,1:K));          % Distance from Cluster Centers
      DAL(i,K+1) = CN;                           % K+1 is Cluster Label
      DAL(i,K+2) = Distance;                     % K+2 is Minimum Distance
   end

for i = 1:K
      A = (DAL(:,K+1) == i);                      % Cluster K Points
      CENTS(i,:) = mean(F(A,:));                  % New Cluster Centers
end
end
end
```