The target features are not given in the training examples.

The aim is to construct a natural classification that can be used to predict features of the data.

The examples are partitioned into clusters or classes. Each class predicts feature values for the examples in the class.

▶ In hard clustering each example is placed definitively in a class.
▶ In soft clustering each example has a probability distribution over its class.

Each clustering has a prediction error on the examples. The best clustering is the one that minimizes the error.
The *k*-means algorithm is used for hard clustering.

**Inputs:**
- training examples
- the number of classes, *k*

**Outputs:**
- a prediction of a value for each feature for each class
- an assignment of examples to classes
$k$-means algorithm formalized

- $E$ is the set of all examples
- the input features are $X_1, \ldots, X_n$
  $X_j(e)$ is the value of feature $X_j$ for example $e$.
- there is a class for each integer $i \in \{1, \ldots, k\}$.

The $k$-means algorithm outputs

- function $\text{class} : E \rightarrow \{1, \ldots, k\}$.
  $\text{class}(e) = i$ means $e$ is in class $i$.
- prediction $\hat{X}_j(i)$ for each feature $X_j$ and class $i$.

The sum-of-squares error for $\text{class}$ and $\hat{X}_j(i)$ is

$$\sum_{e \in E} \sum_{j=1}^{n} \left( \hat{X}_j(\text{class}(e)) - X_j(e) \right)^2.$$ 

Aim: find $\text{class}$ and prediction function that minimize sum-of-squares error.
The sum-of-squares error for class and \( \hat{X}_j(i) \) is

\[
\sum_{e \in E} \sum_{j=1}^{n} \left( \hat{X}_j(\text{class}(e)) - X_j(e) \right)^2.
\]

- Given \textit{class}, the \( \hat{X}_j \) that minimizes the sum-of-squares error is the mean value of \( X_j \) for that class.
- Given \( \hat{X}_j \) for each \( j \), each example can be assigned to the class that minimizes the error for that example.
The $k$-means algorithm

Initially, randomly assign the examples to the classes. Repeat the following two steps:

- For each class $i$ and feature $X_j$, let
  \[
  \hat{X}_j(i) \leftarrow \frac{\sum_{e: \text{class}(e) = i} X_j(e)}{|\{e : \text{class}(e) = i\}|}
  \]
  (the prediction of class $i$ on feature $X_j$)

- For each example $e$, assign $e$ to the class $i$ that minimizes
  \[
  \sum_{j=1}^{n} \left( \hat{X}_j(i) - X_j(e) \right)^2.
  \]

until the second step does not change the assignment of any example.
**k-means algorithm**

Sufficient statistics:
- $cc[c]$ is the number of examples in class $c$,
- $fs[j, c]$ is the sum of the values for $X_j(e)$ for examples in class $c$.

Then define $pn(j, c)$, current estimate of $\hat{X}_j(c)$

$$pn(j, c) = \frac{fs[j, c]}{cc[c]}$$

$\text{class}(e) = \arg \min_c \sum_{j=1}^n (pn(j, c) - X_j(e))^2$

These can be updated in one pass through the training data.
1: procedure \textit{k-means}(Xs, Es, k) \\
2: Initialize \(fs\) and \(cc\) randomly (based on data) \\
3: \textbf{def} \(pn(j, c) = fs[j, c]/cc[c]\) \\
4: \textbf{def} \(\text{class}(e) = \arg\min_c \sum_{j=1}^n (pn(j, c) - X_j(e))^2\) \\
5: \textbf{repeat} \\
6: \(fsn\) and \(ccn\) initialized to be all zero \\
7: \textbf{for each} example \(e \in Es \text{ do}\) \\
8: \(c := \text{class}(e)\) \\
9: \(ccn[c] + = 1\) \\
10: \textbf{for each} feature \(X_j \in Xs \text{ do}\) \\
11: \(fsn[j, c] + = X_j(e)\) \\
12: \(stable := (fsn=fs) \text{ and } (ccn=cc)\) \\
13: \(fs := fsn\) \\
14: \(cc := ccn\) \\
15: \textbf{until} stable \\
16: \textbf{return} \(\text{class}, pn\)
Example Data
Random Assignment to Classes

⊕ is mean of + and ⊗ is mean of ∗
Assign Each Example to Closest Mean
Resign Each Example to Closest Mean
Properties of $k$-means

- An assignment of examples to classes is **stable** if running both the $M$ step and the $E$ step does not change the assignment.

- This algorithm will eventually converge to a stable local minimum.

- Any permutation of the labels of a stable assignment is also a stable assignment.

- It is not guaranteed to converge to a global minimum.

- It is sensitive to the relative scale of the dimensions.

- Increasing $k$ can always decrease error (but does not always) until $k$ is the number of different examples.

How? Given an assignment with $k$ classes, for $k + 1$ classes start with the same assignment, but with the point most distant from its class center in its own new cluster.
EM Algorithm

- Used for soft clustering — examples are probabilistically in classes.
- $k$-valued random variable $C$

Model \[ \begin{array}{c}
C
\end{array} \]
Data \[ \begin{array}{cccc}
X_1 & X_2 & X_3 & X_4 \\
\hline
t & f & t & t \\
f & t & t & f \\
f & f & t & t \\
\ldots
\end{array} \]
Probabilities

\[ \begin{align*}
P(C) \\
P(X_1|C) \\
P(X_2|C) \\
P(X_3|C) \\
P(X_4|C)
\end{align*} \]
EM Algorithm

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$C$</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$t$</td>
<td>$f$</td>
<td>$t$</td>
<td>$t$</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>$t$</td>
<td>$f$</td>
<td>$t$</td>
<td>$t$</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>$t$</td>
<td>$f$</td>
<td>$t$</td>
<td>$t$</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

$P(C)$
$P(X_1|C)$
$P(X_2|C)$
$P(X_3|C)$
$P(X_4|C)$

M-step

E-step
Repeat the following two steps:

- **E-step** give the expected number of data points for the unobserved variables based on the given probability distribution.
- **M-step** infer the (maximum likelihood or maximum a posteriori probability) probabilities from the data.

Start either with made-up data or made-up probabilities.

EM will converge to a local maxima.
Suppose $k = 3$, and $\text{dom}(C) = \{1, 2, 3\}$.

$P(C = 1|X_1 = t, X_2 = f, X_3 = t, X_4 = t) = 0.407$

$P(C = 2|X_1 = t, X_2 = f, X_3 = t, X_4 = t) = 0.121$

$P(C = 3|X_1 = t, X_2 = f, X_3 = t, X_4 = t) = 0.472$: 

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>:</td>
</tr>
<tr>
<td>$t$</td>
<td>$f$</td>
<td>$t$</td>
<td>$t$</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>:</td>
</tr>
</tbody>
</table>
### M step

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$C$</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$1$</td>
<td>$40.7$</td>
</tr>
<tr>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$2$</td>
<td>$12.1$</td>
</tr>
<tr>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$t$</td>
<td>$3$</td>
<td>$47.2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>

\[
P(C=c)
\]

\[
P(X_i = v | C=c)
\]
EM sufficient statistics

- \( cc \), a \( k \)-valued array, \( cc[c] \) is the sum of the counts for \( \text{class}=c \).

- \( fc \), a 3-dimensional array such that \( fc[i, v, c] \), is the sum of the counts of the augmented examples \( t \) with \( X_i(t) = \text{val} \) and \( \text{class}(t) = c \).

- The probabilities can be computed by:

\[
P(C=c) = \frac{cc[c]}{|Es|}
\]

\[
P(X_i = v | C=c) = \frac{fc[i, v, c]}{cc[c]}
\]
procedure EM(Xs, Es, k)
cc[c] := 0; fc[i, v, c] := 0
repeat
cc_new[c] := 0; fc_new[i, v, c] := 0
for each example \( \langle v_1, \ldots, v_n \rangle \in Es \) do 
for each \( c \in [1, k] \) do 
\( dc := P(C = c \mid X_1 = v_1, \ldots, X_n = v_n) \)
cc_new[c] := cc_new[c] + dc
for each \( i \in [1, n] \) do 
fc_new[i, v_i, c] := fc_new[i, v_i, c] + dc
stable := (cc ≈ cc_new) and (fc ≈ fc_new)
cc := cc_new
fc := fc_new
until stable
return cc, fc