Machine Learning, Tutorial 9
Universität Bern

Simon Jenni (simon.jenni@inf.unibe.ch)

K-Means Clustering

1. Given a set of data-points $\boldsymbol{X} = \left\{x^{(1)}, \ldots, x^{(m)}\right\}$ and a number k of cluster centres $\{\mu_i\}_{i=1,\ldots,k}$, the k-means objective can be written as:

$$J(\boldsymbol{C}, \boldsymbol{\mu}) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 . \tag{1}$$

- What are the constraints on $\boldsymbol{C} = \{C_1, \ldots, C_k\}$?
  Hint: Here C_i is a point set, not an index.
  Solution.
  The constraints are that $\boldsymbol{C}$ is a partition of $\boldsymbol{X}$.
  Concretely, $\bigcup_{i=1}^{k} C_i = \boldsymbol{X}$ and $C_i \cap C_j = \emptyset$ for $i \neq j$.

- Alan wants to find the best value for k. He suggests to train k-means with different values of k and then choose the k which achieves the lowest value for $J(\boldsymbol{C}, \boldsymbol{\mu})$. Do you think that this is a good idea? Justify your answer.
  Solution.
  The value of $J(\boldsymbol{C}, \boldsymbol{\mu})$ will always decrease with an increase in k. This is therefore not a good idea.

- Alice suggests to modify the objective by adding a penalty on the ℓ_2-norm of the cluster means:

$$J^*(\boldsymbol{C}, \boldsymbol{\mu}) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 + \lambda \sum_{i=1}^{k} \|\mu_i\|^2 . \tag{2}$$

  Derive the update rule for the cluster means μ_i in this case.

  Solution.

  $J^*(\boldsymbol{C}, \boldsymbol{\mu})$ is a convex function wrt. μ_i. We can therefore set the gradient wrt. μ_i to 0 and solve for μ_i:

$$\nabla_{\mu_i} J^*(\boldsymbol{C}, \boldsymbol{\mu}) = 2\lambda\mu_i - 2\sum_{x \in C_i} (x - \mu_i) \tag{3}$$

  Setting this to 0 and solving for μ_i gives:

$$\mu_i = \frac{1}{|C_i| + \lambda} \sum_{x \in C_i} x \tag{4}$$

2. Consider the following data points:
$x^{(1)} = (1, 1)^T$,
$x^{(2)} = (1, 3)^T$,
$x^{(3)} = (7, 1)^T$,
$x^{(4)} = (7, 3)^T$.

   - Apply the k-means clustering algorithm, when $k = 2$, and the initial centres are $c_1 = (10, 4)^T$ and $c_2 = (0, 2)^T$.

     Solution. Let us first compute the squared distances between the data points and cluster centers.

$dist^2$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
c_1	90	82	18	10
c_2	2	2	50	50

     You can see that $x^{(1)}$ and $x^{(2)}$ are closer to c_2 and $x^{(3)}$ and $x^{(4)}$ are closer to c_1. The cluster assignment is therefore:

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$assign$	2	2	1	1

     The new cluster centres are $c_1 = (x^{(3)} + x^{(4)})/2 = (7, 2)^T$ and $c_2 = (x^{(1)} + x^{(2)})/2 = (1, 2)^T$. Let us iterate this one more time with the new cluster centres.

$dist^2$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
c_1	37	37	1	1
c_2	1	1	37	37

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$assign$	2	2	1	1

     The cluster assignments of the data points are the same as in the previous iteration, therefore the algorithm has converged.

- Apply the k-means clustering algorithm with a different initialisation. The number of clusters is $k = 2$, and the initial centres are $c_1 = (4, 4)^T$ and $c_2 = (4, 0)^T$.

  Solution. The results of the first iteration:

$dist^2$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
c_1	18	10	18	10
c_2	10	18	10	18

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$assign$	2	1	2	1

  $c_1 = (x^{(2)} + x^{(4)})/2 = (4, 3)^T$ and $c_2 = (x^{(1)} + x^{(3)})/2 = (4, 1)^T$

  The next iteration:

$dist^2$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
c_1	13	9	13	9
c_2	9	13	9	13

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$assign$	2	1	2	1

  The algorithm has converged since the assignments did not change.

- Compare the results of the two runs of the k-means algorithm above.

  Solution. We started from two different initializations. In both cases the k-means algorithm converged. The two solutions correspond to two different local optima. The first solution has better cost, since $J_1 = \sum_{i=1}^{n} \|x^{(i)} - c(assign(x^{(i)}))\|^2 = (1{+}1{+}1{+}1)/2 = 2$ and $J_2 = (9{+}9{+}9{+}9)/2 = 18$.