Problem Set 8

Problem Sets due Jul 26, 2022 23:00 PKT Completed

Problems 1-5 correspond to "Clustering with the k-means algorithm I"

Problem 1

4/4 points (graded)

Consider the following data set consisting of five points in $\mathbb{R}^1$:

$$-10, -8, 0, 8, 10.$$

We would like to cluster these points into $k = 3$ groups. Determine the optimal k-means solution.

a) What is the location of the leftmost center?

-9

-9

b) What is the location of the middle center?

0

0

c) What is the location of the rightmost center?

9

9

d) What is the k-means cost of this optimal solution?

4

4

Submit

Problem 2

4/4 points (graded)

Recall that the previous problem dealt with a data set of five points in $\mathbb{R}^1$,

$$-10, -8, 0, 8, 10,$$

and for $k = 3$ we figured out the optimal k-means solution for this data. Unfortunately, Lloyd's algorithm does not always return the optimal solution; in fact, its output depends, in part, on the initialization.

In this problem, suppose we run Lloyd's algorithm, starting with the initialization $\mu_1 = -10, \mu_2 = -8, \mu_3 = 0$. Determine the final set of cluster centers obtained by the algorithm.

a) What is the location of the leftmost center?

-10

-10

b) What is the location of the middle center?

-8

-8

c) What is the location of the rightmost center?

6

6

d) What is the k-means cost of this particular solution?

56

56

Submit

Problem 3

1/1 point (graded)

Suppose we have n data points in $\mathbb{R}^d$, and want to find the k-means clustering of this data. What is the running time of a *single iteration* of Lloyd's algorithm?

- ( ) nd
- ( ) kd
- ( ) nk
- (•) nkd

Submit

Problem 4

1/1 point (graded)

Which of the following properties are true of the $\mathbf{k\text{-}means++}$ algorithm, when run on its own? Select all that apply.

- [ ] If called multiple times, it always returns the same k centers.
- [x] The centers it returns are always actual data points.
- [ ] It is guaranteed to return the optimal k-means solution.

It selects a set of k centers, then spends some time iteratively improving them.

Submit

Problem 5

1/1 point (graded)

Suppose we have n data points in $\mathbb{R}^d$. What is the running time of the $\mathbf{k\text{-}means{+}{+}}$ algorithm?

- n^2kd
- n^2k
- nkd
- nk^2d

Answer

Correct:

At any given time, the algorithm maintains a table T of size n with the squared distance from each data point to the closest of the centers chosen so far. On each iteration: (1) A new center is chosen at random, picking point x_i with probability proportional to $T[i]$; this takes time $O(n)$. (2) The table T is updated: this requires computing the distance from each data point to the newly chosen center, and since there are n data points, and each distance computation takes time $O(d)$, the total time for this step is $O(nd)$. Since there are k iterations, the overall running time is $O(nkd)$.

Submit

Problems 6-8 correspond to "Clustering with the k-means algorithm II"

Problem 6

1/1 point (graded)

In lecture, we described two general uses of clustering: for vector quantization and for finding natural groups in data. Which of these two uses is most directly captured by the k-means cost function?

- Vector quantization
- Finding natural groups in data

Submit

Problem 7

1/1 point (graded)

In the *streaming model of computation*, we need to cluster data but are unable to hold the entire data set in memory. Suppose we only have room for M data points, while our full data set has size $n \gg M$. Here's one suggestion for a streaming algorithm for k-means clustering (assuming $k \ll M$):

- Pick M of the data points at random and store them in memory.
- Run k-means clustering on these M points, to get centers $\mu_1, \ldots, \mu_k$.
- For each point in the full training set: assign it to the cluster with the closest μ_j.

Algorithms like this, based on random sampling, often work reasonably well. In which of the following situations might it fail badly, though? Select all that apply.

- [x] When there are important clusters in the full data set that contain very few points.
- [ ] When there are duplicate copies of some of the data points.
- [ ] When the optimal cluster centers are far away from each other.

✔

Submit

Problem 8

1/1 point (graded)

The *sequential* k*-means algorithm* presented in lecture is able to handle an infinite stream of data, arriving one point at a time. It always maintains a set of k centers that are based on the data seen so far. If the data is d-dimensional, how long does this algorithm take to process the nth data point that arrives?

- ○ k
- ◉ kd
- ○ nd
- ○ nkd

✔

Submit

Problems 9-12 correspond to "Clustering with mixtures of Gaussians"

Problem 9

1/1 point (graded)

A *soft clustering* of a data set into k clusters is a clustering in which each data point isn't necessarily assigned to a single cluster, but rather is allowed to be fractionally assigned to all k clusters. We can represent a soft clustering of n data points by an $n \times k$ matrix P, where

$$P_{ij} = \text{fraction of the ith data point that is assigned to cluster j}$$

and $P_{i1} + \cdots + P_{ik} = 1$ for all $i = 1, \ldots, n$. A more traditional clustering, in which each point chooses a single cluster, is sometimes called a *hard clustering*. If matrix P corresponds to a hard clustering, which of the following conditions must hold? Select all that apply.

- [x] Each entry of P is either 0 or 1.
- [x] Each row of P has a single nonzero entry.
- [ ] Each column of P has a single nonzero entry.
- [ ] No two rows of P are identical.

✔

Submit

Problem 10

1/1 point (graded)

Which of the following is true of the EM algorithm for fitting a mixture of Gaussians to data? Select all that apply.

- [ ] It finds the optimal maximum-likelihood solution.
- [x] Different initializations could lead to different running times.
- [x] Each iteration consists of a soft clustering of the data, followed by an update of the mixture parameters.
- [ ] The algorithm eventually converges to a hard clustering of the data.

✔

Submit

Problem 11

1/1 point (graded)

Which of the following describe ways in which the mixture-of-Gaussians clustering method tries to improve upon k-means?

- [x] It can accommodate clusters of more general shapes and sizes.
- [ ] It is faster.
- [x] The final model doesn't just cluster the data, but also gives probabilities for cluster labels.
- [ ] It is easier to implement for massive data sets.

✔

Submit

Problem 12

Problem 12

0 points possible (ungraded)

For you to think about: Although we talked specifically about mixtures of Gaussians, it is also common to look at mixtures of other kinds of distributions. And indeed, the EM algorithm can quite easily be adapted to these other cases as well. Can you flesh out how this might be done? (No need to enter any answer.)

Submit

Problems 13-15 correspond to "Hierarchical clustering"

Problem 13

1/1 point (graded)

Which of the following are reasons for which hierarchical clustering might be preferred to flat clustering? Select all that apply.

- [x] It does not require the number of clusters to be specified.
- [x] It captures the structure of the data at multiple scales.
- [ ] It is computationally simpler.
- [ ] It has an intuitive cost function for which an optimal solution can efficiently be obtained.

✔

Submit

Problem 14

1/1 point (graded)

Bottom-up linkage" methods for hierarchical clustering work by (i) initially putting each data point in its own cluster and then (ii) successively merging two existing clusters. How many merge steps are needed when there are n data points? Your answer should be a function of n.

[Your answer will not be correctly parsed if you include spaces in it.]

n-1 ✔

Submit

Problem 15

1/1 point (graded)

Some hierarchical clustering algorithms can immediately be applied to data in any metric space. Which of the following algorithms have this property? Select all that apply.

- [x] Single linkage
- [x] Complete linkage

- [x] Average linkage based on average pairwise distance between clusters
- [ ] Ward's method

✓

Submit

‹ Previous Next ›