ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today's Lecture

Recap:

- Supervised Learning

New topics:

- Unsupervised learning
- K-means algorithm

Today's Lecture

Recap:

- Supervised Learning

New topics:

- Unsupervised learning
- K-means algorithm

Recap: Supervised Learning

- In supervised learning, data is labeled at training time:

- The goal is to perform classification or regression

- What are some of the algorithms that you know ?

Today’s Lecture

Recap:

- Supervised Learning

New topics:

- Unsupervised learning
- K-means algorithm

Unsupervised learning

- Another set of problems involves having unlabeled data.
- Clustering:
- Dimensionality reduction:
- Density estimation:

Unsupervised learning

- Clustering: K-means

- Dimensionality reduction: PCA

Today’s Lecture

Recap:

- Supervised Learning

New topics:

- Unsupervised learning
- K-means algorithm

K-means algorithm for clustering

Picture:

K-means is an instance of an iterative algorithm for clustering that iterates between two steps:

1. **Assignment**: for each data point assign the label of the closest prototype.
2. **Refitting**: move each cluster center (prototype) to its center of gravity.

Mathematical set-up

- Suppose we have N data points:

- Define an indicator variable:

Mathematical set-up

- Distortion measure

- Here μ_k is the prototype of class k.

- We want to minimize the distortion.
- We need to find both indicators as well as protypes. How ?

Procedure – overview

Step 0. Initialization.

- Start with some (random) initialization of the prototypes.

Step 1. Assignment.

- For each data point, find the prototype that is closest to it.
- Mathematically:

Procedure – overview

Step 3. Refitting.

- Suppose indicators are given (or known).
- Distortion is a quadratic function of prototypes, given fixed indicators.
- How to find prototypes to minimize the distortion?
- Take a derivative:

Procedure -- overview

- One evaluation of step 2 plus one evaluation of step 3 constitute one iteration.
- Iterate until a stopping criterion is satisfied.

Interpretation of the prototypes

- Sample mean of the points associated with this cluster

- Hence the name **K-means.**
- Where else did we see this before ?

Practical considerations

- How we start/initialize can affect the segments and prototypes that are found.

- Two examples:

- Remedy: perform averaging.

When to stop iterating ?

- When the distortion stops decreasing.

- This will typically be when the assignments stop changing although in some peculiar cases you may see oscillations.

More on initialization

One strategy:

- Pick as the first prototype one of the data points. Then, pick as the second prototype the data point furthest from the first; pick as the third prototype the data point furthest from the first two, and so on.

More on the distortion minimization

- Note that K-means will always converge but not necessarily to a global optimum; it will converge to a local optimum.
- Sum of quadratic (convex) functions is not convex.

- Why doesn’t distortion increase with iterations ?

More on the cluster assignments

- The preceding analysis was for the “hard” 0/1 cluster assignment.
- There is also a “soft” cluster assignment.
- Math:

More on cluster assignments

- Examples of “soft” assignments:

More on cluster assignments

- Examples of “soft” assignments:

More on cluster assignments

- This particular mathematical expression didn’t come out of nowhere. It is arises in Expectation-Maximization of Gaussian Mixture Models.

- Iterative clustering with partial cluster membership and Gaussian clusters.

- Can you recall when we studied Gaussian fitting?

More on convergence

- Note that for K clusters and N data points, there are K^N possible cluster membership assignments.
- We do not check for all of them!
- Recall distortion:

- In the first step, update indicators to lower the current distortion i.e., re-assign to clusters with a lower cost.
- In the second step, minimize the total within-cluster squared distance and output the prototype (cluster center) that gives this minimum.

What about the choice of K ?

- The number of clusters K is a hyperparameter here.
- On one hand, can have K = N. Issue ?
- On the other hand, can have K = 1. Issue ?

- Add the penalty term to the minimization (form of regularization).

What about the choice of K ?

- Another strategy is to start with 1 cluster, and then keep splitting as long as the overall distortion is being reduced.

- Or, start with N clusters, and then keep merging as long as the overall distortion is being reduced.

Other forms of the distance measure

- The preceding analysis was for the squared Euclidean distance (L2 norm).
- When we took the derivative, we got means for the prototypes.

- Other types of distances that are of interest are Manhattan distance (L1 norm); Hamming distance, etc.
- Depending on the choice of the distance measure, we arrive at a different formula for the prototypes (it’s not mean for L1!)

Stochastic update rule

- Recall stochastic gradient descent.

- We have a version of that here, too.

Data compression

- What the preceding method did was to compress the data set of size N into K representatives.
- This is more generally known as **lossy compression**.
- There is also **lossless compression**.

Lossless compression

- Example:

- Recall entropy (when did we study it?)

Lossless compression

- Example, ctd:

Lossless compression

- Example, ctd:

- Average length:

Lossless compression

- Theoretical result for Huffman coding