Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

The k-means formulations

In this section, we'll introduce a common way to do clustering called the k**-means approach**.

Defining a cluster: The mean

We need to define what makes a "cluster" in order to move **forward**.

We want the points within a cluster to be as **close together** as possible. So, you might measure the **distance** from one point to all the others.

So, it would make sense to **average** them out. And we need to average every pair of points. That's a lot of work: can we **simplify** it?

Well, if we're trying to **average** the result of many data points, it would make sense to use the **mean**!

That's how we'll **define** our cluster: as the **mean**, the point that is the **average** of all the other points in the cluster.

Definition 1

We want to represent our **cluster** using its **mean**: the **average** of all of the data points in that **cluster**.

Our goal is for the **cluster mean** to have the **minimum average distance** possible to all of our data points: it's as **close** to our points as we can get.

Example: We describe the "male lifespan" using **life expectancy**: the **average** time a male human lives for. Same for women as well.

k-means

Now, we've created **one** cluster. To extend this to **many** clusters, we just need each cluster to have its **own** mean.

There are k of these clusters: this is why we call this the **k-means formulation**.

How do we decide which point goes in which **cluster**? Well, we want our points to be close. So, we'll assign it to the **closest** one.

Concept 2

A **point** is assigned to the **closest cluster mean**.

For a point $x^{(i)}$, the **output** is which **cluster** ("new class") it has been assigned to: $y^{(i)}$.

Once we've successfully clustered using our **algorithm** below, we will find that both of these goals are met:

- Our points are **assigned** to the **closest** cluster mean.
  - This separates **different** clusters of points from each other.
- The cluster mean is the **average** of all of our points: the **minimum distance** to them.
  - This makes sure our cluster is made up of points that are **similar** to each other.
  - If our point is close to the **mean**, it's probably close to the **other** points in the cluster.

k-means loss

Now, we know what we want out of our **clusters**. But, the problem is, we don't know **which** points will give us our nice clusters.

So, first, we will have to **assign** our initial "cluster means": often, we **randomly** select some points from our dataset.

Concept 3

We **initialize** our clustering by **randomly** selecting one point to **represent** each cluster, which we call the **cluster mean**.

At first, each point is assigned to the **closest** cluster mean.

But as you'll notice, these points are **not** the cluster means we're looking for! They're just a random **initialization**. So, we have to **optimize**.

Clarification 4

Notice that, when we **first** select our "cluster means", we don't get them by **averaging** any points: we choose them **randomly**.

That means, at first, is our cluster mean **isn't a true mean**!

Our k-means algorithm is designed to **fix** this problem.

In order to **improve** our clustering, it helps to have a way to measure the **quality** of a clustering: we need a **loss function**.

One-cluster loss

Let's start with just one cluster: what do we want to **minimize**?

Well, we want the points within a cluster to be as **close together** as possible. So, we want to minimize the **distance** to the mean, μ.

To make our function smooth, we'll use **squared distance** instead.

Concept 5

In **k-means loss**, we want to minimize the **square distance** from each point $x^{(i)}$ to the **cluster mean** μ.

$$D_i = \left\| x^{(i)} - \mu \right\|^2 \tag{1}$$

We'll add this up for each of the n data points in our cluster.

$$\mathcal{L} = \sum_{i=1}^{n} \left\| x^{(i)} - \mu \right\|^2 \tag{2}$$

Building up to k clusters

So, what do we do for each of our k clusters? Well, we can just **add** up the **loss** for them.

We'll use $j \in \{1, 2, 3, ...k\}$ to represent our j^{th} cluster. Each cluster has a mean $\mu^{(j)}$.

$$\mathcal{L}_j = \sum_{i=1}^{n} \left\| x^{(i)} - \mu^{(j)} \right\|^2 \tag{3}$$

Problem is, we're including **every** point $x^{(i)}$ in **every** cluster! We want a way to filter by **cluster**.

Remember that we **label** clusters the same way we labeled **classes** before:

Notation 6

For a **data point** $x^{(i)}$, its **cluster** is given by

$$y^{(i)} \in \{1, 2, ...k\}$$

Where j represents the j^{th} cluster.

Cluster mean $\mu^{(j)}$ is the j^{th} cluster mean: it only counts for points in c_j. So, we **only** want to add up the loss when

$$y^{(i)} = j \tag{4}$$

We'll do this using the following helpful **function**:

Notation 7

The **indicator function** $\mathbb{1}$ tells you whether a statement p is true:

$$\mathbb{1}(p) = \begin{cases} 1 & \text{if } p = \text{True} \\ 0 & \text{otherwise (if } p = \text{False)} \end{cases}$$

Combined with our **condition** of matching clusters, this can be useful:

$$\mathbb{1}(y^{(i)} = j) \tag{5}$$

If we **multiply** this by our loss, it'll **only** appear if the clusters **match**! We can **eliminate** data points in a different cluster.

k-mean loss: final form

So, we can **filter** by the data points in our cluster:

$$\mathcal{L}_j = \sum_{i=1}^{n} \overbrace{\mathbb{1}(y^{(i)} = j)}^{\text{Check cluster}} \overbrace{\left\| x^{(i)} - \mu^{(j)} \right\|^2}^{\text{Dist from mean}} \tag{6}$$

And finally, we add up over many clusters:

$$\mathcal{L} = \sum_{j=1}^{k} \mathcal{L}_j \tag{7}$$

Using our equation, we get:

$$\mathcal{L} = \overbrace{\sum_{j=1}^{k}}^{\text{clusters}} \overbrace{\sum_{i=1}^{n}}^{\text{data points}} \overbrace{\mathbb{1}(y^{(i)} = j)}^{\text{Check cluster}} \overbrace{\left\| x^{(i)} - \mu^{(j)} \right\|^2}^{\text{Dist from mean}}$$

Let's clean that up:

Key Equation 8

The k**-means loss** is given as:

$$\mathcal{L} = \sum_{j=1}^{k} \sum_{i=1}^{n} \mathbb{1}(y^{(i)} = j) \left\| x^{(i)} - \mu^{(j)} \right\|^2$$

Where:

- μ_j is the **cluster mean**: the **average** of the points in the j^{th} cluster.
- $\mathbb{1}(y^{(i)} = j)$ is the **indicator function**: meaning that we only **include** terms where the data point and mean are in the **same cluster**.