Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Solutions: Validation

Now, we start trying to answer the **question**: how do we **check** whether we have a **good** clustering?

Well, first, we can check for a **poor fit** (or overfitting) using new, **held-out** testing data: do we get **low loss** on that testing data?

If we **don't**, then our clusters definitely aren't **representative** of the overall **dataset**: they don't **generalize** to new data.

> **Concept 1**
>
> If our clusters give **large testing loss**, then they aren't **generalizing** well, and are probably **not representative** of the overall distribution.
>
> So, we already know our clusters **don't fit the distribution**.

Solutions: Consistency

But, just like for classification/regression **validation**, we don't only run our algorithm **one time**: we'll run it **many** times, with different training and testing sets.

We can't **just** use the loss, though: having **more** clusters could make our error lower, without making a better clustering, for example.

Another thought: we're trying to find some patterns **inherent** in the data. The idea is: if the pattern we're finding is **real**, we should find a similar pattern **each time**!

So, we look to see if our clusters are **consistent** when we generate them using different training data: if they **aren't**, then it's possible we're not finding the "**real**" patterns in the data.

Different training data from the same distribution, of course.

> **Concept 2**
>
> If our **clusters** accurately **reflect** the underlying classes of data, then we should expect some **consistency** of which clusters we **generate** by running k-means many times.
>
> If our clusters aren't **consistent**, then we might doubt if any of them especially reflect the **distribution**, rather than **noise**.

If our clusters are **consistent**, then we're probably seeing something about the **real** dataset.

If it was based on random noise, then the odds of getting matching results would be really low!

Solutions: Ground Truth

But, even if we're getting something **consistent**, that doesn't mean we're seeing the patterns that **matter**.

One way to **check** this is, if we have some idea of what the "**true**" clustering looks like for just a few data points, we can compare those results to ours.

We call this "real" clustering the "ground truth".

Definition 3

In machine learning, the **ground truth** is what we know about the "real world".

In general, we want our models to be able to **reproduce** this reality: it is the data that we tend to **trust** the most, if it is gathered correctly.

That way, we can use a very **small** amount of **supervision** to get an idea of whether our clustering is on the **right track**.