Florian Rösel, Frauke Liers Winter semester 2021/2022

Mathematics of Learning – Various Questions 1

Explain the differences between clustering and categorization.
Well..
Clustering: Belongs to unsupervised learning. You are confronted with a set of data vectors, and you know originally nothing about it. You try (often for different k you guess, since you often do not know better), what is the best way to form valid groups of your data vectors.
Categorization: Belongs to supervised learning. Often, you are confronted with a set of data vectors, but you know already to which categories you want to assign them (this also implies that you often know the number of categories, k). For categorization, you would use different techniques, which we deal with later in the lecture.

Is semi-supervised learning applicable in real world data? What is a good example to know about it?

We come to semi-supervised learning later in the lecture; let's come back to this question then.

Why do we not find the second differential to find if the gradient we obtained is the minima or the maxima? As it is an optimization problem, we should aim to find the minima for it.

Two reasons: First, k-means-clustering is an algorithm consisting of very simple operations - leading to relatively fast calculation times. Calculating the second derivative is an expensive task, so it would ruin our algorithmic performance, even if we would not have our second reason: It is just not necessary, since we know that once we have assigned our data points to clusters, the hessian matrix (i.e. the second derivative) is positive in any case - so we indeed calculate some minima, anyways.

Is it possible that even though a certain number of clusters is "optimal" but not useful for the problem at hand? If so, do we deviate from the optimal solution?

Yes, it is possible that the cluster is not "optimal" for the problem, but is optimally solved by the algorithm. If this is the case, we need to look into the other ways of solving the problem, or maybe more features. - Addition: Just try what happens if you apply k-means-clustering to some random data or some spherical data or to some clustered data, but with a wrong value of k. The algorithm may find the minimum for k-clusters, but it might be more or less meaningless :)

If a point is equidistant from two or more clusters, then does it matter which cluster it is considered in? If so, then how do we determine which cluster to put it in?

It does not matter too much if a point is equidistant from 2 (or more) clusters, but even then there exist many tie-breaking methods, one of them being assigning randomly,

CAREFUL: assigning both clusters to the point is **NOT SUITABLE** (the rule is "assign **one** cluster'),

but assigning e.g. the smallest cluster is possible.

I forgot your question? Remember me: florian.roesel@fau.de