CFCS Tutorial Two: Solutions

Miles Osborne

January 30, 2008

This tutorial deals with simple vector-based machine learning. Suppose you are given the following sets of reading about *dogs*:

1	0.75	0.25
2	0.75	0.5
3	0.5	0.5

Also, you have the following readings about *cats*:

1	0.5	0.6
2	0.25	0.5
3	0.25	0.75

1. Using a *Knn* classifier, work-out whether the following examples are *cats* or *dogs*:

1	0.6	0.6
2	0.75	0.25
3	0.25	0.25

Answer: The way to answer this question is for each example, compute the distance between each *cat* vector and each *dog* vector. Them you sort all the distances, smallest first. Using the first example, compared against the cats, we would have:

1	0.5	0.6	0.1
2	0.25	0.5	0.4
3	0.25	0.75	0.4

(The number in the last column is the absolute distance).

And against the *dogs*, we have:

1	0.75	0.25	0.4
2	0.75	0.5	0.2
3	0.5	0.5	0.1

For $k = 1$, the single closest example (0.1) is a cat (or a dog, depending on how we break the tie). For $k = 3$, we have a dog, a cat and another dog (0.2), so we would say that it was a dog (there are two dog votes). Etc.

2. Now, using some paper, create a two-d graph and mark the various points on it. Do your previous results agree with a visual inspection?

3. What happens as you vary k? Answer: As k varies, we become more robust to errors in the training set (for example, mis-readings or saying a cat was a dog). If k is too high, we average over too many instances and so we over-generalise.

4. If you saw 10 more of the type two *cat* reading, what would happen to your results? Answer: This would act as a kind of attractor and in effect would have the influence of 10 cats (any example close to these 10 cats would much more likely be classified as a cat than before).

5. Suppose a *cat* reading is the same as a *dog* reading. What would happen? Answer: Nothing would happen, since both points would have the same label: we learn nothing from this example.

6. Our set of *cat* and *dog* readings very usefully told us which kind of animal went with which kind of reading. This is called *supervised machine learning. Unsupervised machine learning* deals with examples that have no explicit label –we do not know whether a set of readings came from a *dog* or a *cat.* How could you assign labels to such readings? Answer: One approach might be to see which label it would be assigned by the k-nn classifier and use that for it.

   You should think about what would happen if you added these newly labelled examples to your initial set of labelled examples. Answer: We would learn more about the examples, since we have more of them. If those unlabelled examples were in very clearly defined regions (for example, right in the middle of a set of cat examples) then it would be safe to conclude that they were really cat examples. This would then make us more confident about future examples, since we now would have more evidence of what it means to be a cat example.