Theory

One strategy to levy the flexibility and power of neural networks when faced with a classification problem and small data sets, is to make a so called siamese network. Such networks have shown good results when employed in image recognition(ref Hadsell, FaceNet). The basic premise of these networks is that the feature space is mapped to some metric space where the images of input vectors can be can be compared. Vectors whos images are “close” in the given metric can be thought of as similar. The goal then, is to train the network to map the different categories to different “areas” of the metric space and, in a sense, learn the concept of sameness.

The name siamese network stems from the structuring of the network. It consists of a base network, wich can have any structure, that acts as a function mapping input vectors into an N dimensional space, where N is the number of output nodes for the base network. For a twin network, two (three for triplets) instances of this base network are created with shared weights and biases. The network can thus take two (or three) different vectors as input and the conjoined base networks will output a corresponding number of N dimensional representations. Then follows a merging layer where the distances between the images of the input vectors are calculated. This output is then fed to a layer that outputs a similarity score between each vector.

One method to train a twin network is to first pick one representative, referred to as an anchor, from each class. Then, for each anchor a\_i and each datapoint x\_j pair (a\_i,x\_j) is created. If the pair belong to the same category they are labeled 1 and if they belong to different categegories the pair receives the label 0. These pairs, along with their labels, will then serve as training data for the network. The network is then updated according to a loss function that pushes inputs from different classes apart while clustering similar inputs. Hadsell et al proposes the contrastive loss function:

CONTRASTIVE LOSS y(dist(a,x)^2+(1-y)(max(m-dist(a,x),0))^2

This is basicaly two different loss functions in one. If the the label y = 1, then the distance is minimized, if y=0, it is maximized up to a margin m.

The method is somewhat similar for triplet networks. For each data point x\_i, pick one representative from the same category, a\_s, and one representative from a different category, a\_d. Update the shared network according to the triplet loss function from FaceNet:

Triplet loss max(dist(a\_s,x\_i)^2 – dist(a\_d,x\_i)^2 +m , 0)

Again, m is a margin beyond wich the network will stop updating, a maximum of “differentness” if you will.

Perhaps the main advantage to the siamese approach, though of no value to the tasks proposed in this report, but interesting in its own right, is the ability to learn new classes without retraining the network. For instance, when trained to differentiate faces, these networks are able to tell whether or not two pictures are of the same person, even if it has never “seen” this person before.

Results:

For the siamese network approach, the results were, on the whole, disappointing. While the networks certainly showed signs of learning, getting good results on the training data, they displayed little to no ability to generalize. The results from a typical run can be seen in figure (accuracy) and figure (loss).

Figure : Accuracy scores for neural networks predicting high or low total phosphorus content by looking at community composition. Top: Simple siames network with 2 hidden layers of 50 neurons. Middle: Complex siamese network with 3 hidden layers of 100 neurons. Bottom: Triplet network using the complex siamese base network.

Attempts to alleviate overfitting by adding dropout layers or adjusting hyper parameters, did not lead to improvements on predictive ability much beyond the level of random guessing.

With our choice of data set, it is not obvious whether this failure is due to the methods employed or a lack of significant correlation between the microbial community composition, and the various target variables, such as phosphate concentrations or temperature at the sample sites. Another possible reason for these results might be the choice of classification method. The choice to classify sites by whether they fell above or below the median of the data set was not inspired by any particular biological theory beyond different environments breeding different biota. The environmental variables wich the sites were classified by might have critical points at wich the community distribution significantly changes, but there is no a priori reason to assume this would be the median.

Siames implementation

To explore classification with random forests and siamese networks, the different sites were grouped into to classes

based on whether the local measurement was above or below the meadian for of all 72 lakes. It is not unreasonable to expect certain variables, such as temperature, pH or phosphate content to correlate with the distribution of microbial species. As such, attemps were made to predict, for one variable at a time, whether or not sites were above or below the median for the given variable by looking at the species distributions.

Only species present in all lakes were used as predictors and target variables chosen were total phosphorus(TP) and temperature. Three different approaches to classify the lakes into high/low categories were tested. Two siamese network varaints and one using random forest. For the siamese networks, anchors, i.e prototypical representatives for the given class, were chosen by inspection. For the twin siamese network, pairs consisting of one of the two anchors and a data point to be evaluated, were created for all combinations of data points and anchors, yielding a total of 144 pairs. For the triplet network, each data point was paired with both anchors, giving a total of 72 triplets. 80 percent of the pairs/triplets were used for training and the remaining 20 percent were used to asess predictive ability.

addendum to introduction (4) CLassify lakes

addendum research questions calssify lakes

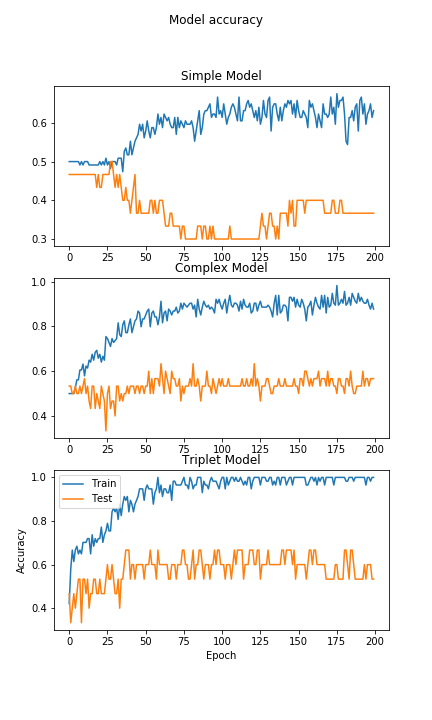


Figure : Loss values for neural networks predicting high or low total phosphorus content by looking at community composition. Top: Simple siames network with 2 hidden layers of 50 neurons. Middle: Complex siamese network with 3 hidden layers of 100 neurons.

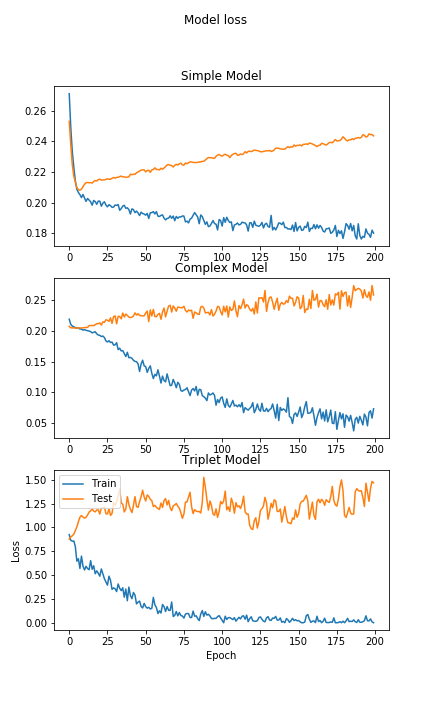


Figure :

Sources:

Hadsell et al 2006

http://yann.lecun.com/exdb/publis/pdf/hadsell-chopra-lecun-06.pdf

FaceNet

https://arxiv.org/abs/1503.03832v3