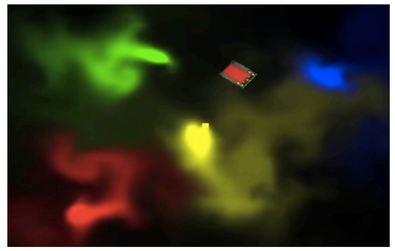
MACHINE HOUND: Creating an artificial nose

Jacques de Chalendar, Marguerite Graveleau, Clément Renault

CS 229: Machine Learning, STANFORD UNIVERSITY

Introduction



GOAL

- Recognize odors, and their concentrations, from a set of sensors on a robot evolving in space and time

MOTIVATION

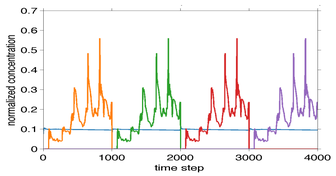
- An interesting application of this project would be the application of sensory networks trained by this data to enable neural sensorimotor control of the agent in the virtual worlds
- This task is analogous to the capability of biological systems to learn the smell of an object in one environment, and recognize it in multiple different contexts

Database description

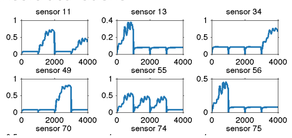
The training datasets are comprised of values representing sensor activation (input features) and odorant concentration (output) at 4000 different timepoints. The test is repeated for two machines (T20 and T40), with the same odor environment but with different sensors. During the training recording, only one odor is present at a time, with a noisy background. The response of the sensors to a time-varying concentration signal is recorded for each odor. The testing set however, includes all 4 odors.

TRAIN SETS

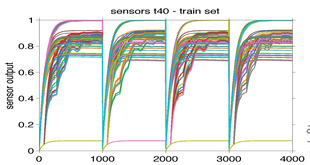
Concentrations



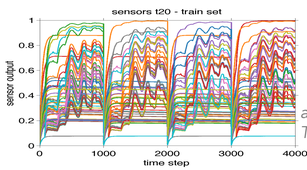
Sensors activations



typical single sensor response



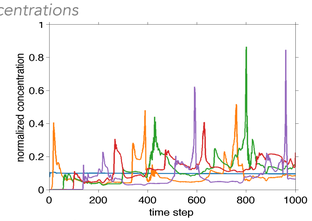
all sensors T40



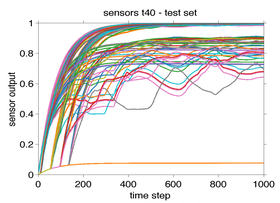
all sensors T20

TEST SETS

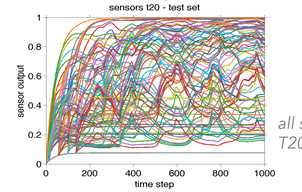
Concentrations



Sensors activations



all sensors T40

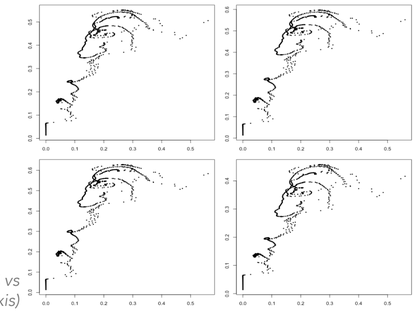


all sensors T20

One of the difficulties is that the testing set is more messy than the training set (in particular several odors are present at a time)

Correlation of sensor readings with odor concentration

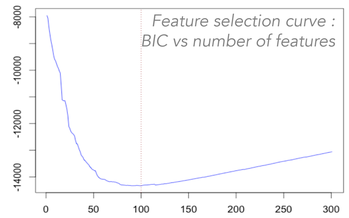
We plot here the 4 sensors the most correlated with Odor 1 concentration. Not only are they highly representative of the presence of Odor 1, but they are also correlated together (greater than 0.98). It seems that correlations are associative in this dataset. (we observe the same behavior for the 3 remaining odors).

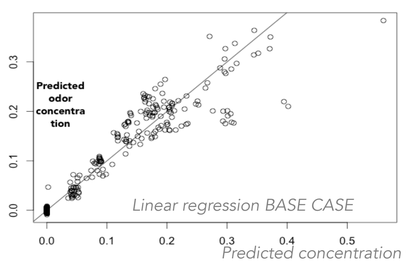


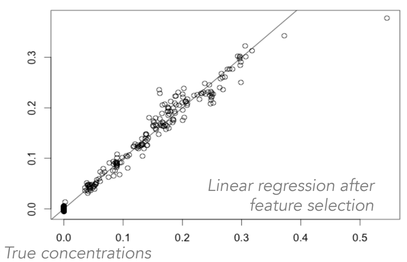
sensors activation (y axis) vs Odor 1 concentration (x-axis)

Linear regression models

We first performed a simple linear regression using all the sensors as predictors (BASE CASE plot). We then included all interaction terms of the sensor readings and performed forward feature selection using the BIC. We reduced the error by almost 2 after the feature selection (from MAE=0.0079 to MAE=0.004)





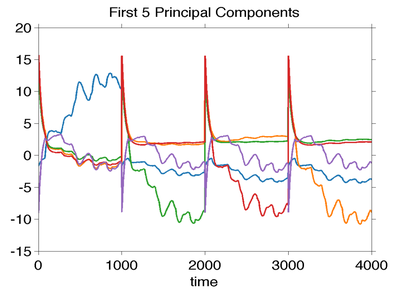


Predicted concentration vs True concentrations

We notice that out of the 100 first features that are selected only 5 are sensor signals, all the others being interaction terms. This makes sense if we consider the odor from a chemical perspective. The chemicals activate the sensors, and looking at combinations of sensors tells us which chemicals were present

Principal Component Analysis

The importance of interaction terms lead us to try out a Principal Component Analysis. We observe that indeed only the first 5 pc vectors explain more than 95% of the variance.

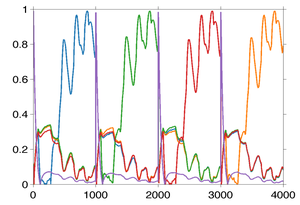


- In the plot we observe that one pc is specific of the background, the 4 others are specific of one odor.
- Running a linear regression only on these 5 pc reduces the dimension of our input by a factor of 20, and only increases the mean squared error from 0.79% to 0.84%.

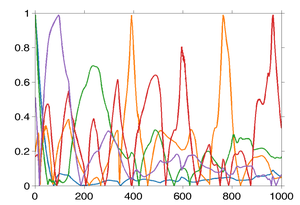
Independent Component Analysis

This problem is analogous to invariant object recognition in clutter and background, and called for an independent component analysis with the 5 first principal components as inputs (we normalize our data).

PREDICTION ON TRAIN SET



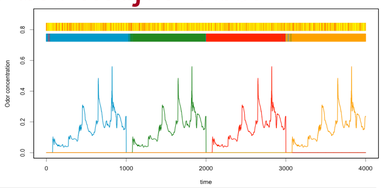
PREDICTION ON TEST SET



- Results are clearer on the training set then on the test set. Major peaks and evolution in time are respected, but the comparative concentrations are wrong (maybe due to normalization, but results are not conclusive without normalization)
- Two shortcomings of ICA are that we cannot recover scaling and that we cannot determine which odor is which (ICA is invariant to a permutation of the sources)

Predicting the class of the major odor

- We train a logistic regression model on 80% of the training set and validate it on the remainder.
- Less than 2% classification error is observed.



Conclusion

- This problem is made difficult by the great mix of information in a real environment (test set) where several odors occur at a time
- Simple linear regression gives fair results for a start
- Dimension reduction, through feature selection or PCA, is well adapted to our data where sensors share a lot of information and are correlated by nature

Future work

- So far the prediction of the class (predominant odor) works well, but a precise prediction of the amount od each odor is still our goal. (On the other hand, a prediction that quantifies the evolution of each odorant suffices to build a hypothetic machine hound)
- The data being time-dependent, an approach through Markov-chains is worth considering
- Seeing the shape of the activation of each sensor, being activated or not by some odors, we would expect a neural networks approach to give good results. This dataset actually comes from a challenge aiming to compare the performance of ML algorithms against neural networks.

References

[1] website: http://tunedit.org/challenge/artificialOlfaction
[2] video demonstration: https://www.youtube.com/watch?v=kCCe88OMpA8