## Real World Data Mining Task: Nuclear Magnetic Resonance Metabolite Prediction

Edited 3/9/2015 - Paul Anderson

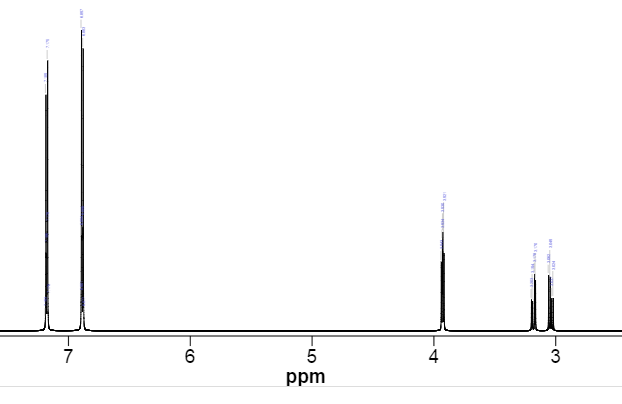
So far in class we have focused on learning the fundamentals of core machine learning and data mining algorithms, such as neural networks. We’ve focused on writing everything ourselves to get a firsthand feel for how the algorithms work. While we continue to do this we will focus our efforts on real world problem that people are actively trying to solve. We are going to work on developing an algorithm to predict whether or not a metabolite is present in a spectrum or not.

**Dataset:** 1H Urine NMR Human Fatigue Study. Specifically, my research collaborators and I have been able to show that we can predict whether someone will be fatigue resistant or susceptible based on their molecular profiles. In case you are interested in what a NMR machine is…



It is basically a huge magnet. Similar to those used in MRI machines, but we are using to measure the concentration of metabolites in biological samples.

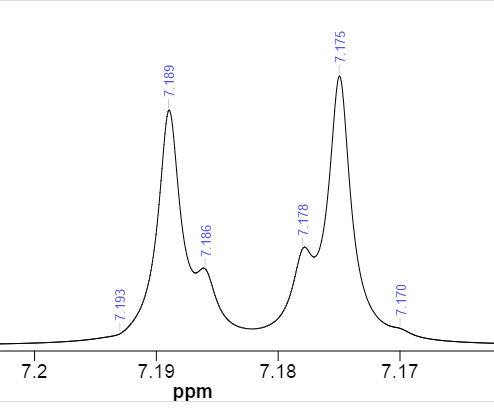
**Our task:** It is a difficult problem to identify metabolites from a noisy biological signal. You’ll see this very quickly. Our task is simplified but still difficult. In fact, I have no idea how you will perform. I am hoping you’ll surprise and amaze me. Specifically, we will work on predicting the presence or absence of a specific metabolite: Tyrosine. Here is a completely clean version of a Tyrosine spectrum:



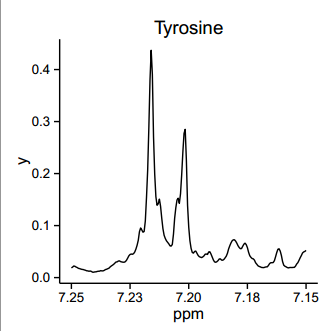
Should be easy right? Well, our real signal is not so clean. In fact, it can be horrendously complex. This is both the power and weakness of 1H NMR metabolomics, which is the field we are applying our data mining skills. OK. Now that our overall task is defined, here is what you have at your disposal.

**Data:** Everything is available on a public github repository that I’ve created for this assignment: <https://github.com/Anderson-Lab/Metabolite-Prediction>.

I’ve done the prep work of processing the original biology signals. The directory that you will want to focus on is Tyrosine. This contains the training data files that will need in addition to a bunch of images to help you along the way. The first images to examine are entire\_tyrosine, region1\_tyrosine, region2\_trysine, region3\_tyrosine, region4\_tyrosine, and region5\_tyrosine. It shows that there are actually 5 distinct regions that this signal appears. This redundancy is our friend and ally. It gives us multiple places to confirm and look for the existence of this metabolite. Here is region 1:



This shows you that around 7.19 to 7.17 in the clean/pure spectrum there is the above pattern. This is something that you can roughly see in one of the real spectra:



Notice that there is a similar shape but it is shifted to the left. It is also not as symmetrical as the “clean” image. These are the problems that you are going to have to deal with. Here are the problems we will have to deal with:

1. Noise from other signals
2. Shift of the signal left or right. You can assume that within a region the peaks will not shift relative to each other, but they will move as a unit.

One thing that will help you is that the intensities of the regions are relative to each other. Specifically, the relative intensities of the regions are:

Table of Peaks

No. (ppm) Height

1 7.193 -0.0004

2 7.189 0.0263

3 7.186 0.0058

4 7.178 0.0079

5 7.175 0.0300

6 7.170 0.0006

7 6.901 0.0010

8 6.897 0.0335

9 6.894 0.0070

10 6.886 0.0083

11 6.883 0.0313

12 6.877 0.0003

13 3.943 0.0077

14 3.934 0.0096

15 3.930 0.0106

16 3.921 0.0087

17 3.203 0.0035

18 3.194 0.0033

19 3.178 0.0064

20 3.170 0.0057

21 3.062 0.0063

22 3.049 0.0061

23 3.037 0.0036

24 3.024 0.0037

Table of Multiplets

No. Shift1 (ppm) Hs Type Atom1 Multiplet1 (ppm)

1 3.024 0 dd 8 M05 3.014 .. 3.072

2 3.170 0 dd 8 M04 3.161 .. 3.213

3 3.921 0 dd 9 M03 3.909 .. 3.954

4 6.877 2 m 2 6 M02 6.868 .. 6.909

5 7.170 2 m 3 5 M01 7.161 .. 7.201

Table of Assignments

No. Atom Exp. Shift (ppm) Multiplet

1 8 3.024 M05

2 8 3.170 M04

3 9 3.921 M03

4 2 6.877 M02

5 6 6.877 M02

6 3 7.170 M01

7 5 7.170 M01

We’ll go over this in class, so that it makes more sense. There are a bunch of precomputed images for you to examine and get a feel for the data, but the meat of the dataset is contained in two data files:

1. positive\_train.csv – this contains examples of data that contain tyrosine. The first column is the X values (ppm). The rest of the columns correspond to a sample. There are roughly 90 samples in this dataset.
2. negative\_train.csv – this contains examples of data that do not contain tyrosine. It is set up so that the X values (ppm) are shift exactly 1 ppm. But there will be the same number of ppm values in both the negative and positive files.

Now your task can be explicitly stated. You are to come up with an algorithm, where I could give you a previously unknown sample from either positive or negative and your code will be able to look at it and tell me whether there is tyrosine in that specific example. That’s it!

**Deliverables, Points, and Due Date**

So this is a very open ended and difficult problem. Like I said companies are designing software and have been for years in the hopes of doing this very thing. The good news is you may use anything you want that you find in either Python or R. This problem is worth 50 points or the equivalent of 5 Learn2Mine Assignments. We’ll still continue with the Learn2Mine assignments. This is in addition to those assignments and not a replacement.

**Due:** It is due on the day of the final exam. There is Dropbox on OAKS where you can upload your technical report and code.

**Deliverables:** You are responsible for writing up your methods and your findings. You will not be graded on whether or not you reach a given performance. But it is expected that you develop a thoughtful approach and testing methodology to evaluate your approach. This is critical and this is where most of the points will be assigned. I want to teach you how to approach a difficult problem. How to go about setting up a testable framework. How you can go about tuning parameters to see if they make any performance gain. And documenting the results. It is very much a lab science in this way. You’ll need to write up a technical report that states what methodology you’ve tried and what worked and also just as importantly what didn’t work. I want you to treat this as your first assignment as a data scientist/data miner. Someone has come to you with this problem, and in 5 weeks time, they will want to see thoughtful documented progress. You can discuss what you’ve learned what you’ve tried. But it is NOT a narrative or a blog. This is technical writing. Facts, facts, facts. Think hypotheses. Think scientifically. What might work? How do you know if it works? What are some of the challenges with the data? You should have an introduction, methods, results, discussion, and conclusion. It should be at least several pages. It is 5 weeks of work after all. I do NOT want to see code in the technical write-up. That you can upload, but it is separate. I’m looking forward to what you find and what you try. As an added incentive, if anyone manages to be very successful, they will be able to publish their findings in scientific and computer science journals!