SPLEX TME 12

Sequences and Structured Output Prediction
Dynamic Data

This TME consists of two parts. In the first part, the goal is to learn how to estimate a model which takes structured data into consideration, and how to do structured output prediction. The second part is devoted to (basic) treatment of kinetic data.

Data (both data sets are provided)

- For structured output prediction ("ProteinSecondStr_Train" and "ProteinSecondStr_Test")
- For analysis of dynamic data (data provided during the TME)

Analysis: Part 1 (Structured Output Prediction)

We will use `hmmlearn` library (already installed). It performs unsupervised learning.
Here you can find its tutorial
`http://hmmlearn.readthedocs.io/en/latest/tutorial.html`

1. Generate simulated data as it is done in the following example

   `http://hmmlearn.readthedocs.io/en/latest/auto_examples/plot_hmm_sampling.html#sphx-glr-auto-examples-plot-hmm-sampling-py`

2. Data are produced as follows, X are the observations, and Z are the states.

   ```
   X, Z = model.sample(500)
   ```

3. Train the model

   ```
   remodel = hmm.GaussianHMM(n_components=4, covariance_type=''full'', n_iter=100)
   remodel.fit(X)
   Z_pred = remodel.predict(X)
   ```

4. `hmmlearn` performs unsupervised segmentation. It is quite challenging to evaluate using accuracy or error rate. Plot the prediction and the original structure to see whether the prediction is accurate

   ```
   plt.plot(Z)
   plt.plot(Z_pred)
   Z_pred = remodel.predict(X)
   ```

5. Run the analysis on the protein structure prediction task. The first column contains observations, and the second one contains the states. Ignore the second column for the moment.

Analysis: Part 2 (Analysis of Dynamic Data)

Download the data

- foodT0 and foodT6 – nutritional patterns at two different time points
- glucoseT0 and glucoseT6 – glucose values at two different time points

$$\Delta = X_{\text{time = T+1}} - X_{\text{time = T}} \tag{1}$$

1. Perform canonical correlation analysis on the deltas to explain temporal changes between the nutritional habits and the glucose variables (look into the TME 8)
2. Construct a Bayesian network from the delta data to explore relations between changes in nutrition and glucose (look into the TME 10)

Bonus (Structured Output Prediction with Conditional Random Fields)

- Install `python-crfsuite`
- Perform supervised sequence labelling on the protein structure data using the CRF.

  You can use this tutorial to run the experiments

  `http://www.albertauyeung.com/post/python-sequence-labelling-with-crf/`
- What is the prediction error?