SNLP exercise set 02: mean, variance, correlation, ...

Ç. Çöltekin — SfS / University of Tübingen

May 12, 2017

This set of exercises intended for getting you up to speed with `numpy` with some simple exercises on probability / statistics and a bit of vector operations.

In this exercise set we will use the data from last exercise set. If you have completed the exercises 2 and 3, you already have most of the data ready. We need the following variables (lists of numbers) from the last exercise.

`age` Child's age (e.g., in days, but feel free to use any other unit of time).
`chittr` Child's type/token ratio (TTR)
`motttr` Mother's TTR
`chimlu` Child's average utterance length (or mean length of utterance, MLU)
`motmlu` Mother's MLU

Note that the accurate calculation of age requires paying attention to `@Date` header of the CHAT files. But an approximation will do fine for our purposes.

In case you haven't finished the first set of exercises (yet), You can get them in a tab-separated file from the course web page.

In addition to above data, we will create some more data by counting / string processing first.

For these exercises, using `numpy` arrays for the data will be more convenient. Alternatively, you may consider using `DataFrame` from the `pandas` package, which is similar to the data frames you may be familiar from R.

Except for the first exercise, you should be fine with working on the python command line.

Exercises

Exercise 1. The '`%mor:`' lines in CHAT files include morphological information, where the first part (separated with vertical bar, '`|`') contains the POS tags. For this exercise we also want to exclude sub-tags of POS tags, which may be prefixed after a colon '`:`'.

Calculate the number of occurrences of each POS tag in file `adam01.cha` separately for the mother and the child. Store your results such a way that the same indices on both arrays (or lists) correspond to the frequency of the same POS tag.

We will refer to the POS tag counts you calculated in this exercise as `chipos` and `motpos`.

Exercise 2. Calculate the *mean*, *variance*, *standard deviation* of both `chimlu` and `motmlu`.

Do not use `numpy` for this exercise, but compare your results against values you obtain with `numpy` array methods.

Exercise 3. Calculate the correlation coefficient for the following quantities.

`np.corrcoef()`

- `age` and `chimlu`
- `chittr` and `chimlu`
- `chimlu` and `motmlu`
- `age` and `motmlu`

You should try to interpret your results. For example, you should ask yourself, which one of the correlation coefficients indicate a strong relationship or when you observe a correlation, which one of the variables is like to 'cause' the other?

Exercise 4. Calculate the *dot product* and *cosine similarity* between `chimlu`–`motmlu`, and `age`–`motmlu`.

`np.dot()` takes dot product of two vectors, `scipy` and `sklearn` packages have functions for cosine similarity, but you are recommend it to write it yourself.

Which measure is a more appropriate measure of similarity in this case?

Exercise 5. If you subtract mean of a sample from every value of the sample, the resulting data is said to be *centered*.

Center the variable `chimlu`, divide the result to the number of elements in the array.

Does the result look familiar?

Tip: check your results from Exercise 2. Can you calculate covariance using the same trick?

Exercise 6. Scale the variables `chipos` and `motpos` calculated in Exercise 1 so that they form proper probability distributions.

- Calculate entropies of both distributions.
- Calculate cross entropy between the two distributions (in both directions).

Does mutual information makes sense in this exercise?