Lecture reviews — Week 07
with solutions

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Week 7 keypoints

- supervised/unsupervised
- preprocessing is key
- baseline methods:
  - classification: Naive Bayes, (Logistic regression,) KNN
  - clustering: K-means, dendrograms
  - dim. reduction: PCA, UMAP
- don't forget evaluation keypoints (see lesson 2)

J.-C. Chappelier & M. Rajman

Week 7 – study case

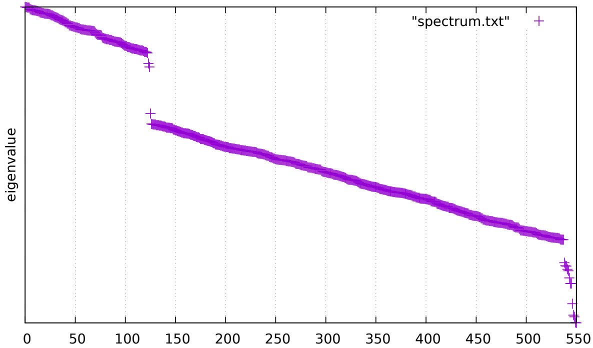
Some financial company offers you to work on
"*fraud detection using Natural Language Technology applied to client documents*".

① Some preliminary work has already been performed by a former intern who created document vectors based on an indexing set of 6'324 terms and reduced them to vectors of size 100 using PCA.

Reviewing his/her work and report, you found a graph related to the corresponding singular values.
Next slide shows a (rescaled) zoom on the first 550 left-most points in that graph.

J.-C. Chappelier & M. Rajman

Week 7 – study case



a) What is the abscissa (x-value, horizontal axis) of the right-most point in the original complete graph (not reported here)? 6'324

b) What do you think about the intern's methodology for selecting the dimension of the vector space? Would you have performed differently? If yes, how?

Week 7 – study case (solution)

The general idea is good (reducing dimension keeping most data variance), however the concreate approach is not really sound as 100 seems like a random choice. $\simeq 125$, or if compatible with other external constraints, $\simeq 540$ are more appropriate since big gap in intertia.

[Reporting the percentage of total inertia would also help in such a context.]

J.-C. Chappelier & M. Rajman

Week 7 – study case

② Before considering more sophisticated Deep-Learning methods, you wisely decide to start with a simple baseline, namely a Naive Bayes model (on the former representation).

a) Based on your former answer, what is the input of the Naive Bayes module?
What is the output?
What are the parameters?
What is needed for training such a model?

b) Concretely, what probability should be computed as an output from the (very simple excerpt of) client document:

My salary is about 10’000 CHF and I don’t pay any tax.

J.-C. Chappelier & M. Rajman

Week 7 – study case (solution)

a)

input: "*document* vector" i.e. document PCA representation as done in previous question

output: most proable class (fraud/non-fraud);

parameters: $P(\text{class})$ for both classes and $P(\text{feature}|\text{class})$ for each "feature" (PCA dimension) resulting from previous question

needed for training: supervised (fraud/non-fraud) corpus of typical documents

b) $P(\text{class}) \times \prod_{i=1}^{n} P(f_i|\text{class})$

where n is either 125 or 540 from former answer,
f_i are the coordinates of the PCA representation of the above document,
and "*class*" is either fraud or non-fraud.
[Sure, the difficult part is to properly model $P(f_i|\text{class})$, which is a continuous probability distribution!!]

J.-C. Chappelier & M. Rajman
EPFL

Week 7 – study case

③ From your first analysis of the baseline results, you realize that single tokens do not adequately capture dependencies that clearly appear at the syntactic level (for instance the one between “*don’t*” and “*pay*” in the former example).
Using some syntactic parser, you are able to transform the former example sentence

My salary is about 10’000 CHF and I don’t pay any tax.

into:

SALARY-10K-RANGE not_pay tax

a) What probability would then be computed as the resulting output by the Naive Bayes model in such a case?

b) Compared to former Naive Bayes model, what is <u>the</u> main fundamental reason why you can reasonably expect the results to be better?

J.-C. Chappelier & M. Rajman

Week 7 – study case (solution)

a) The same kind of formula as above except that now n is the number of remaining indexing tokens and f_i are those remaining indexing tokens

b) It increases features independence (Naive Bayes key assumption)
and certainly better task-oriented features (filtering)

J.-C. Chappelier & M. Rajman