Foundations of Machine Learning II

Bastien BRIER
bastien.brier@student.ecp.fr

February 27, 2017

TP5: Compression, Prediction, Generation, Text Entropy

1 Questions

1.1 Interpret the time invariant assumption associated to our Markov chains for text compression.

This assumption means that the probability of a character appearing after another one is the same wherever we are in the text. A n has the same probability of appearing after an a in the i^{th} position as in the 1^{st} position.

1.2 How can we rewrite a Markov chain of higher order as a Markov chain of order 1?

We can rewrite this Markov chain of higher order as a product of conditional probabilities.

1.3 Given a probability distribution over symbols, how to use it for generating sentences?

We use the state that we are in (the symbol or group of previous symbols) and, thanks to the distribution we know, generate the new symbol according to the probability distribution.

1.4 As for supervised learning, a model can overfit the training set. Propose a simple approach (cost function) for measuring the goodness of the model.

The KL divergence is a good way to measure the goodness of a model. We heavily penalize the fact that something exists in the test and has not been treated in the train : the higher the KL, the worst the generalization of the model.

2 Implementation

2.1 For different orders of dependencies, train the model on a novel and compute the associated entropy. What do you observe as the order increases? Explain your observations.

We compute the entropy of different order of dependencies in the two models, using the novel Dostoevsky.txt.

Order	1	2	3	4	5	6
IID	4.43	7.91	10.60	12.65	14.30	15.67
Markov	4.43	3.48	2.69	2.05	1.66	1.37

We can observe major differences. For the IID model, the entropy increases as the order increases. It can be explained by the fact that the symbols are larger and more diverse, thus we need more bits to compress and keep all the information.
On the contrary, for the Markov model, the entropy decreases as the order increases. In fact, as the order increases, we also increase the information fed to our probabilistic model. Thus the probability become higher and we then need less information to deduce probabilistically the end of the text.

2.2 Use the other novels as test sets and compute the cross-entropy for each model trained previously. How to handle symbols (or sequences of symbols) not seen in the training set?

We computed the cross-entropy for both models in different orders. Our reference text was still Dostoevsky.txt.

IID Model

Order	1	2	3	4	5	6
Goethe	5.56	10.42	15.90	21.28	24.66	26.08
Alighieri	5.82	10.42	15.53	20.75	24.40	25.93
Hamlet	5.09	9.84	15.27	20.69	24.26	25.90

Markov Model

Order	1	2	3	4	5	6
Goethe	5.56	5.86	9.02	15.54	21.62	25.10
Alighieri	5.82	6.39	8.93	14.74	21.13	24.73
Hamlet	5.09	5.34	8.30	14.42	20.57	24.54

For symbols not seen in the previous set, we penalised heavily by applying a value of 0.00000001 in P(train). As it is in the denominator in the KL divergence formula, it increases the cross-entropy.

2.3 For each order of dependencies, compare the cross-entropy with the entropy. Explain and interpret the differences.

For the IID model, the entropy and the cross-entropy both increase, the cross-entropy faster than the former. The symbols are more complex and more diverse which explains the large increase.
On the contrary, for the Markov model, the entropy decreases whereas the cross-entropy increases. It must be a sign of overfitting. In fact, we heavily penalize the fact that something exists in the test and not in the train : its associated probability is really really small, and thus needs a lot of bits to be represented.

2.4 Choose the order of dependencies with the lowest cross-entropy and generate some sentences.

For a better visibility, we instead selected generating sentences of order 4.
IID Model :

```
swofromem.  youxistt of stoturb andy. Otle ew abut stras cortangones,
hy heatch us.wyerd pa am all  aftto hateseeliat ee has, Il thdal
I dhat is h_thae. Ifee.uld na I pla and my shudradi
```

Markov Model :

```
plet of it's not she stancy, busive eight all in his eversburg, when.
"Stay, off. He look at that nican she still take then idescrupted
at him is, you art." He and there times... Dounia! You no once
```

2.5 Train one model per novel and use the KL divergence in order to cluster the novels.

We trained a Markov model of order 2 and computed the different KL divergences.

Train / Test	Dostoevsky	Goethe	Alighieri	Hamlet
Dostoevsky	0	2.52	3.20	1.97
Goethe	2.47	0	2.84	**0.33**
Alighieri	4.84	4.49	0	4.26
Hamlet	2.05	1.24	4.31	0

We see that the KL between Goethe and Hamlet is much lower than the other ones. We can then regroup these in the same cluster.
When we looked at the texts, it seems logical since these texts are in German, Dostoevsky in English and Alighieri in Italian.