Machine Learning: Natural Language Processing (NLP) through Optimal Transport and Statistics

CHHAIBI Reda

22-26 of July 2019

Abstract

These 4 lessons give a modern application of optimal transport and statistics: Natural Language Processing (NLP).

The general purpose of NLP is to compare sentences and giving a measure of proximity between them. This is necessary for both classification and random generation.

1 Lesson 1: Word prediction via Markov chains

As a warm-up and in order to contrast with modern techniques, we start by using a very classical tool, that is Markov chains. If the theoretical background is very well established, we shall see that the scope is limited.

2 Lesson 2: Word embeddings

It is in this lesson where the mathematically minded students have to show an open mind. We will present and use a rather miraculous technique called "Word Embeddings".

If the theoretical background is still at its infancy, we shall see that this technique is extremely promising in machine learning.

3 Lesson 3: Numerical aspects of optimal transport

The goal of this lesson is to give you the tools to actually compute *numerically* the transport maps between probability measures. We shall focus on techniques which work for discrete empirical measures - as this is the most common case in statistics.

4 Lesson 4: Optimal transport of language and document classification

By combining the powerful techniques taught in the two previous lessons, we give a taste of state-of-the-art document classification.