Time-series information and unsupervised representation learning

Daniil Ryabko

INRIA Lille, SequeL

COMPLACS/MLCOGS 2013

Given a highly dependent time-series

$$X = X_1, X_2, \ldots, X_n, \ldots,$$

where X_i are from some high-dimensional space $\mathcal{X}$, find a (representation) sequence

$$Y = Y_1, Y_2, \ldots, Y_n, \ldots,$$

where Y_i are from some small (in this talk, *finite*) space $\mathcal{Y}$ such that

Y preserves as mush as possible of the information about time-series dependence in X.

Why would we need such representations Y_i? hoping to solve easier some (semi-) supervised learning problems later

The problem can be thought of as dimensionality reduction with the focus on time-series dependence

Time-series information

Under some conditions, given a certain a set of functions $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$, the function that preserves the most of the time-series information is the one that maximizes

$$I_\infty(f) := h_0(f(X)) - h_\infty(f(X)) = I(f(X_0); f(X_{-1}), f(X_{-2}), \dots).$$

Theorem

Let $f : \mathcal{X} \to \mathcal{Y}$ be such that $(X_i)_{i \in \mathbb{N}}$ are conditionally independent given $(f(X_i))_{i \in \mathbb{N}}$. Then for any $g : \mathcal{X} \to \mathcal{Y}$ we have

$$I_\infty(f) \geq I_\infty(g),$$

with equality if and only if $(X_i)_{i \in \mathbb{N}}$ are conditionally independent given $(g(X_i))_{i \in \mathbb{N}}$.

The time-series information $h_0(f) - h_\infty(f)$ does not involve the (distribution of) X_i directly — only through $f(X_i)$. This means that for estimation we only need to look at the representations; modelling or estimating the distribution of X_i is never needed.

Learning-theory-style results

To find a function $f \in \mathcal{F}$ that maximizes $I(f)$, maximize empirical $I(f)$; for finite sets $\mathcal{F}$ of functions f this works for any stationary ergodic time series; for infinite $\mathcal{F}$ one needs conditions on mixing times and on the VC dimension of $\mathcal{F}$.

Active case: MDPs

To find a function that maximizes time-series information for an MDP, it is enough to execute a random policy and find the function that maximizes empirical time-series information for that policy. It will be good for any stationary policy!