Using a signature-based machine learning model to analyse a psychiatric stream of data

IMANOL PEREZ
(JOINT WORK WITH T. LYONS AND K. SAUNDERS)

Mathematical Institute
University of Oxford

7th Berlin-Oxford Meeting

UNIVERSITY OF OXFORD
Mathematical Institute

Oxford Mathematics

Signature of a path

Continuous paths with finite p-variation

- Given $p \geq 1$ and $X \in \mathcal{C}([s,t],\mathbb{R}^d)$ with $s < t$ we define

$$\|X\|_{p,[s,t]} := \sup_{\{t_i\}_i \subset [s,t]} \left(\sum_i \|X_{t_i} - X_{t_{i-1}}\|^p \right)^{1/p} .$$

- Given $p \geq 1$ and $X \in \mathcal{C}([s,t],\mathbb{R}^d)$ with $s < t$ we define

$$\|X\|_{p,[s,t]} := \sup_{\{t_i\}_i \subset [s,t]} \left(\sum_i \|X_{t_i} - X_{t_{i-1}}\|^p \right)^{1/p}.$$

- $\mathcal{V}^p([s,t],\mathbb{R}^d) := \{X \in \mathcal{C}([s,t],\mathbb{R}^d) : \|X\|_{p,[s,t]} < \infty\}.$

University of Oxford
Mathematical Institute

Definition (Signature of a continuous path)

Let $X \in \mathcal{V}^p([0,T],\mathbb{R}^d)$ so that the following integration makes sense. The signature of X is defined as

$$S(X) = (1, X^1, X^2, \dots) \in \bigoplus_{n=0}^{\infty} (\mathbb{R}^d)^{\otimes n}$$

where

$$X^n = \int \dots \int_{0<u_1<u_2<\dots<u_n<T} dX_{u_1} \otimes \dots \otimes dX_{u_n} \quad \forall n \geq 1.$$

Definition (Truncated signature of a continuous path)

Similarly, we define, for $n \geq 0$,

$$S^n(X) := (1, X^1, X^2, \ldots, X^n).$$

- We have two data sets: a known set of known input-output pairs (the *training set*), $\{X_i, Y_i\}_i$, which is used to train the model, and a set of inputs that is used for testing (the *out-of-sample set*).

- We have two data sets: a known set of known input-output pairs (the *training set*), $\{X_i, Y_i\}_i$, which is used to train the model, and a set of inputs that is used for testing (the *out-of-sample set*).
- Features play an important role in machine learning.

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem (B. Hambly, T. Lyons)

The signature of a path with bounded variation is unique up to tree-like equivalence.

Theorem

Let $X \in \mathcal{V}^1([0,T], \mathbb{R}^d)$ be a path with bounded variation. Then, given $1 \leq i_1, i_2, \ldots, i_n \leq d$ we have

$$\left\| \int \cdots \int_{0<u_1<u_2<\ldots<u_n<T} dX^{i_1}_{u_1} \ldots dX^{i_n}_{u_n} \right\| \leq \frac{\|X\|^n_{1,[0,T]}}{n!}.$$

- Given a training set $\{R_i, Y_i\}_{i=0}^{N}$, of input-output pairs, where $R_i = \{t_{ij}, r_{ij}\}_j$ is a stream of data, construct a new set $\{X_i, Y_i\}_{i=0}^{N}$ with $X_i \in \mathcal{V}^1$.

UNIVERSITY OF OXFORD
Mathematical Institute

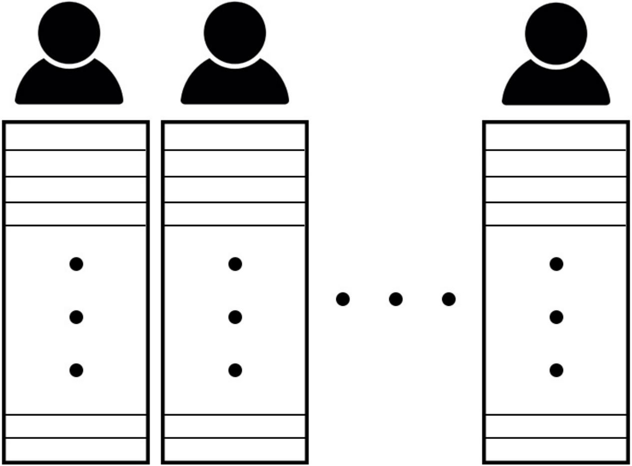
- Given a training set $\{R_i, Y_i\}_{i=0}^{N}$, of input-output pairs, where $R_i = \{t_{ij}, r_{ij}\}_j$ is a stream of data, construct a new set $\{X_i, Y_i\}_{i=0}^{N}$ with $X_i \in \mathcal{V}^1$.
- Compute $\{S^n(X_i), Y_i\}_{i=0}^{N}$ for some $n \in \mathbb{N}$.

UNIVERSITY OF OXFORD
Mathematical Institute

- Given a training set $\{R_i, Y_i\}_{i=0}^N$, of input-output pairs, where $R_i = \{t_{ij}, r_{ij}\}_j$ is a stream of data, construct a new set $\{X_i, Y_i\}_{i=0}^N$ with $X_i \in \mathcal{V}^1$.
- Compute $\{S^n(X_i), Y_i\}_{i=0}^N$ for some $n \in \mathbb{N}$.
- Apply regression against the truncated signature.

Application to psychiatric data

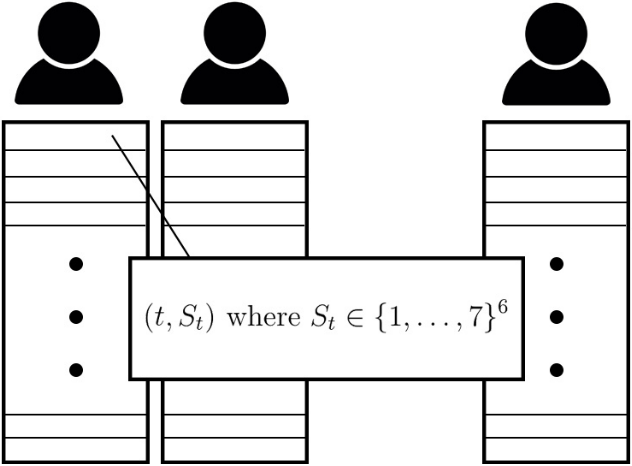
The problem



Application to psychiatric data

The problem

UNIVERSITY OF OXFORD
Mathematical Institute



- Given some information about a participant, can we tell if he or she was diagnosed to have bipolar disorder, borderline personality disorder or to be healthy?

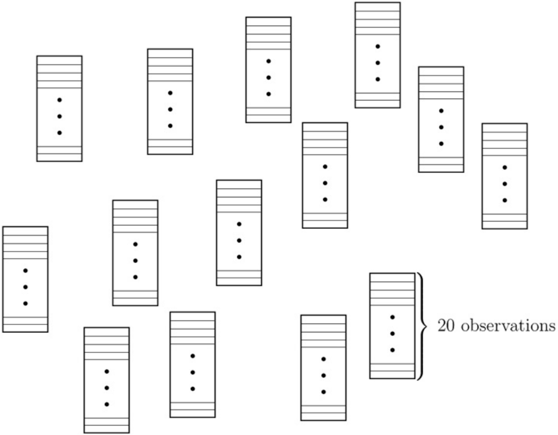
Application to psychiatric data

The problem

- Given some information about a participant, can we tell if he or she was diagnosed to have bipolar disorder, borderline personality disorder or to be healthy?
- Given a participant and information about the last few days, can we predict the mood the following day?

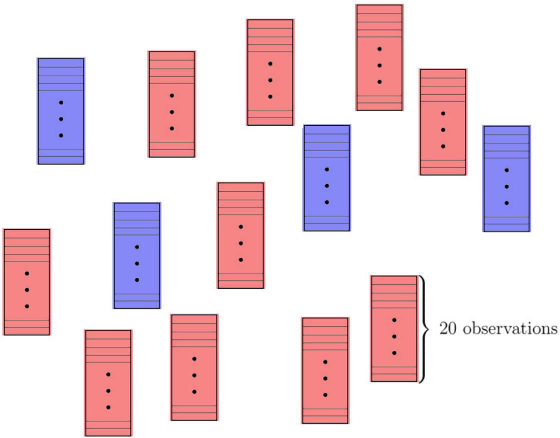
Application to psychiatric data

Methodology



Application to psychiatric data

Methodology



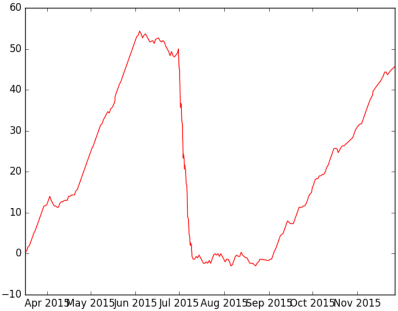
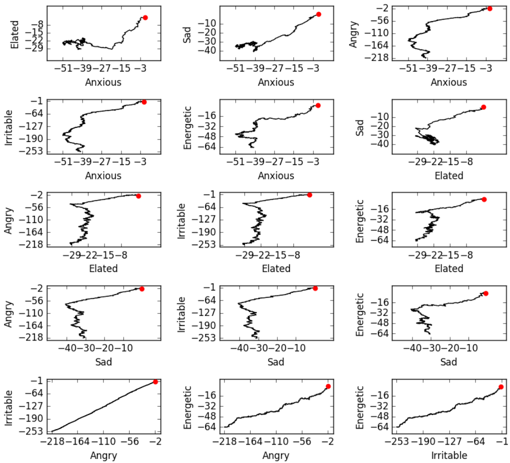


Figure: Normalised path for anxiety scores.


Elated
Anxious
Sad
Anxious
Angry
Anxious
Irritable
Anxious
Energetic
Anxious
Sad
Elated
Angry
Elated
Irritable
Elated
Energetic
Elated
Angry
Sad
Irritable
Sad
Energetic
Sad
Irritable
Angry
Energetic
Angry
Energetic
Irritable

Application to psychiatric data

Predicting if a person is healthy, has bipolar disorder or has borderline disorder

$$\{t_i, S_{t_i}\}_{i=0}^{19} \quad \rightarrow \quad \begin{cases} (-1,1), & \text{if the partcipant is healthy} \\ (-1,-1), & \text{if the participant is bipolar.} \\ (1,0), & \text{if the participant is borderline.} \end{cases}$$

Application to psychiatric data

Predicting if a person is healthy, has bipolar disorder or has borderline disorder

Order	Correct guesses
2nd	75%
3rd	70%
4th	69%

Table: Percentage of people correctly classified in the three clinical groups.

$$\{t_i, S_{t_i}\}_{i=0}^{19} \quad \rightarrow \quad S \in \{1, \ldots, 7\}^6$$

where S is the scores of the participant the following observation.

Application to psychiatric data

Predicting the future mood

Category	Healthy	Bipolar	Borderline
Anxious	98%	82%	73%
Elated	89%	86%	78%
Sad	93%	84%	70%
Angry	98%	90%	70%
Irritable	97%	84%	70%
Energetic	89%	82%	75%

Table: Percentage of correct guesses for mood predictions

Thank you!