Data Science and me

Guido Sanguinetti

ANC- School of Informatics, University of Edinburgh
Room IF1.44 gsanguin@inf

October 19, 2015
Positional statement

- I was trained as a physicist/mathematician
- Emphasis on Science in Data Science
  - I’m unconvinced by statements that large-scale data gathering will eliminate the need for theory (i.e. hypothesis driven research), except perhaps in some engineering applications.
  - However, science also produces vast amounts of data
  - Statistical models and machine learning techniques are increasingly central in turning data into knowledge.
I was trained as a physicist/mathematician

Emphasis on Science in Data Science

I’m unconvinced by statements that large-scale data gathering will eliminate the need for theory (i.e. hypothesis driven research), except perhaps in some engineering applications.

However, science also produces vast amounts of data

Statistical models and machine learning techniques are increasingly central in turning data into knowledge.
Largish group: 4 post-docs, 6 students, 8 nationalities

Funding from several sources: ERC, EPSRC, Marie Curie, School of Informatics, CDT/ DTC

Backgrounds from physics, engineering, CS and maths

Interests range from analysis of sequencing data to dynamical systems theory
1. Dynamical systems and biology

2. Two examples
   - Formal models meet machine learning
   - Epigenetics

3. Looking ahead and refs
Dynamical systems

- Abstractions of real systems focusing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$, $dx = f(x)dt + \sigma dW$, …
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for prediction and understanding, i.e. they strike a nice balance between explanatory and predictive power
Dynamical systems

- Abstractions of real systems focusing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$, $dx = f(x)dt + \sigma dW$, ...
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for prediction and understanding, i.e. they strike a nice balance between explanatory and predictive power
Dynamical systems

- Abstractions of real systems focussing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$,
  $dx = f(x)dt + \sigma dW, \ldots$
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for prediction and understanding, i.e. they strike a nice balance between explanatory and predictive power
Where does variability come into play? What can we measure? Nice example of a dynamical system with some physical knowledge and a lot of uncertainty.
Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion.

Also, more precise microscopy-based measurements give time-resolved measurements at single cells.

Each measurement is a noisy readout of one facet of a (set of) complex biological processes.

Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture.

Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings).
Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion.

Also, more precise microscopy-based measurements give time-resolved measurements at single cells.

Each measurement is a noisy readout of one facet of a (set of) complex biological processes.

Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture.

Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings).
Modelling behaviours

- In many cases, we build models to replicate qualitative behaviours, e.g. oscillations, transients, etc.
- Theoretical computer scientists have developed languages to describe and reason on behaviours, *temporal logics*, originally to reason about software failures
- A central problem is *probabilistic model checking*: given a model of a stochastic system, and a behaviour of interest, what is the probability that the behaviour will actually arise in a sampled trajectory?
  - Generally computationally intensive to answer
  - Clearly relevant beyond software: given a model of a bacterium, what is the probability that its behaviour will switch to pathogenicity? Given a model of a pacemaker and the heart, what is the probability that we will have fibrillation?
Modelling behaviours

- In many cases, we build models to replicate qualitative behaviours, e.g. oscillations, transients, etc.
- Theoretical computer scientists have developed languages to describe and reason on behaviours, *temporal logics*, originally to reason about software failures.
- A central problem is *probabilistic model checking*: given a model of a stochastic system, and a behaviour of interest, what is the probability that the behaviour will actually arise in a sampled trajectory?
- Generally computationally intensive to answer.
- Clearly relevant beyond software: given a model of a bacterium, what is the probability that its behaviour will switch to pathogenicity? Given a model of a pacemaker and the heart, what is the probability that we will have fibrillation?
In many cases, we build models to replicate qualitative behaviours, e.g. oscillations, transients, etc.

Theoretical computer scientists have developed languages to describe and reason on behaviours, *temporal logics*, originally to reason about software failures.

A central problem is *probabilistic model checking*: given a model of a stochastic system, and a behaviour of interest, what is the probability that the behaviour will actually arise in a sampled trajectory?

Generally computationally intensive to answer.

Clearly relevant beyond software: given a model of a bacterium, what is the probability that its behaviour will switch to pathogenicity? Given a model of a pacemaker and the heart, what is the probability that we will have fibrillation?
Smoothed model checking

- Model checking presumes full specification of a model
- In real applications, that is not available; in particular, parameters are always uncertain → need tools for sensitivity analysis
- We have proved (with L. Bortolussi and D. Milios) that satisfaction probabilities for a wide class of systems are smooth functions of the parameters
- We can turn the sensitivity analysis into a machine learning problem: solve at a few parameter values, then predict (emulate) everywhere else
- Technical ingredient: Gaussian process binomial regression
Model checking presumes full specification of a model

In real applications, that is not available; in particular parameters are always uncertain → need tools for sensitivity analysis

We have proved (with L. Bortolussi and D. Milios) that satisfaction probabilities for a wide class of systems are smooth functions of the parameters

We can turn the sensitivity analysis into a machine learning problem: solve at a few parameter values, then predict (emulate) everywhere else

Technical ingredient: Gaussian process binomial regression
Other work and current challenges

- As well as sensitivity analysis, we can also perform optimisation, e.g. designing a system that (robustly) satisfies a certain behaviour
- Or solve inverse problems, e.g. having observed satisfaction/not satisfaction of certain behaviours, can we determine the parameters of the system?
- Technical ingredients: Bayesian optimisation
- Challenges: most GP-based methods unfeasible beyond 5-10 dimensions (number of parameters)
- Possible solutions: sparsification, primal optimisation, dimensionality reduction (?), identifying modularities (??)
Other work and current challenges

- As well as sensitivity analysis, we can also perform optimisation, e.g. designing a system that (robustly) satisfies a certain behaviour.
- Or solve inverse problems, e.g. having observed satisfaction/not satisfaction of certain behaviours, can we determine the parameters of the system?
- Technical ingredients: Bayesian optimisation.
- Challenges: most GP-based methods unfeasible beyond 5-10 dimensions (number of parameters).
- Possible solutions: sparsification, primal optimisation, dimensionality reduction (?), identifying modularities (??)
Epigenetics

Genetics and transcription cannot be all; spatial organisation of chromosomes plays a role. This is determined by chemical modifications to DNA and histones.
Epigenetics: what the data looks like

Each row is a tiny fraction of a next-generation sequencing experiment’s data. Each row $\geq 1$GB of data. How do we determine relationships between the rows?
Current results

- Identifying statistically significant differences between the rows is already difficult: some success adapting a kernel method, *Maximum Mean Discrepancy* (Gretton et al 2008), to sequencing data (Schweikert et al, BMC Genomics 2013, Mayo et al, Bioinformatics 2015)

- Predictive models are useful: e.g., given a hypothesis that the green rows are mechanistically determined by the pink rows, we should be able to train a fairly accurate regression model

- Recent success in predicting histone modifications from binding of transcription factor proteins (Benveniste et al, PNAS 2014)

- Technical challenges: large size of the data sets, large number of covariates, inhomogeneities along chromosomes (latent variables?)
Current lines of work

- Develop predictive models to relate DBA sequence and epigenetic marks with each other, based on generalised linear models (T. Mayo)
- Model the interactions between various epigenetic factors and gene expression (consensus clustering, soon to move to more general graphical models) (A. Kapourani, CDT)
- Also important to understand processes downstream of transcription, e.g. RNA folding (A. Selega) and splicing (Y. Huang), and (remarkably) these are often also tied to epigenetics
Looking ahead

- At the moment, the two lines of work appear fairly disjointed, how do we integrate them?
  - Technical challenge 1: scaling up formal analysis methods
  - Technical challenge 2: (almost) all epigenetic data is a snapshot of a stochastic dynamical process. How do we do inference for (large scale) stochastic dynamical systems from (population/ time) average static measurements?
  - Technical challenge 3: how do we identify effective smaller (dynamical) models that match the behaviours observed in data?
At the moment, the two lines of work appear fairly disjointed, how do we integrate them?

Technical challenge 1: scaling up formal analysis methods

Technical challenge 2: (almost) all epigenetic data is a snapshot of a stochastic dynamical process. How do we do inference for (large scale) stochastic dynamical systems from (population/ time) average static measurements?

Technical challenge 3: how do we identify effective smaller (dynamical) models that match the behaviours observed in data?
References

- L. Bortolussi, D. Milios and G.S., Smoothed Model Checking for Uncertain Continuous Time Markov Chains, Information and Computation 2015
- L. Bortolussi and G. S., Learning and designing stochastic processes from logical constraints, QEST 2013 and Logical Methods in CS 2015
- G. Schweikert, B. Cseke, T. Clouaire, A. Bird and G.S., MMDiff: quantitative testing for shape changes in ChIP-Seq data sets, BMC Genomics 14:826, 2013
- D. Benveniste, H.-J. Sonntag, G.S. and D. Sproul, Transcription factor binding predicts histone modifications in human cell lines, PNAS 111(37), 13367-13372, 2014
- T. Mayo, G. Schweikert and G.S., $M^3D$: a kernel-based test for spatially correlated changes in methylation profiles, Bioinformatics 31(6), 809-816, 2015