Personalized Health with Gaussian Processes

Neil D. Lawrence

Disease Mapping Workshop, Leahurst

4th November 2013

Outline

Health

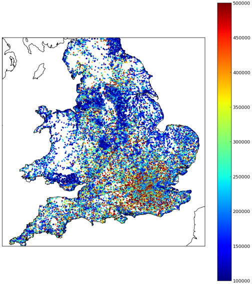
Data Heterogenity

Deep Learning

Conclusions

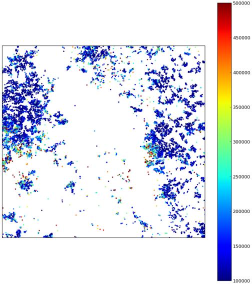
What's Changed (Changing) for Medicine?

- Modern data availability.

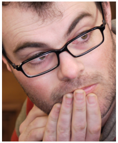


What's Changed (Changing) for Medicine?

- Modern data availability.



Hensman et al. (2013)



Gaussian Processes for Big Data

James Hensman* Dept. Computer Science The University of Sheffield Sheffield, UK

Nicolò Fusi* Dept. Computer Science The University of Sheffield Sheffield, UK

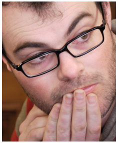
Neil D. Lawrence* Dept. Computer Science The University of Sheffield Sheffield, UK

Abstract

We introduce stochastic variational inference for Gaussian process models. This enables the application of Gaussian process (GP) models to data sets containing millions of data points. We show how GPs can be variationally decomposed to depend on a set

Even to accommodate these data sets, various approximate techniques are required. One approach is to partition the data set into separate groups [e.g. Snelson and Ghahramani, 2007, Urtasun and Darrell, 2008]. An alternative is to build a low rank approximation to the covariance matrix based around 'inducing variables' [see e.g. Csató and Opper, 2002, Seeger et al., 2003, Quiñonero Candela and Rasmussen, 2005, Tit-

Hensman et al. (2013)



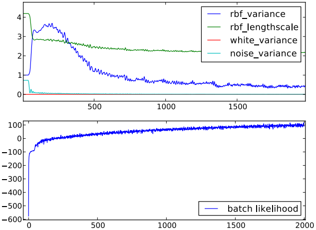


Figure 4: Convergence of the SVIGP algorithm on the two dimensional toy data

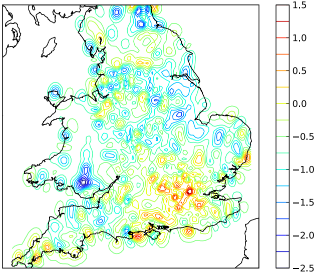


Figure 5: Variability of apartment price (logarithmically!) throughout England and Wales.

`land-registry-monthly-price-paid-data/`, which covers England and Wales, and filtered for apartments. This resulted in a data set with 75,000 entries,

ted a GP with the same covariance function as our

What's Changed (Changing) for Medicine?

- Try Googling for: "patient data "...

Image from Wikimedia Commons

Image from Wikimedia Commons

Driver and Vehicle
Licensing Agency

INF57

A brief history *of Registration*

For more information go to: **www.direct.gov.uk/motoring**

A brief history of registration

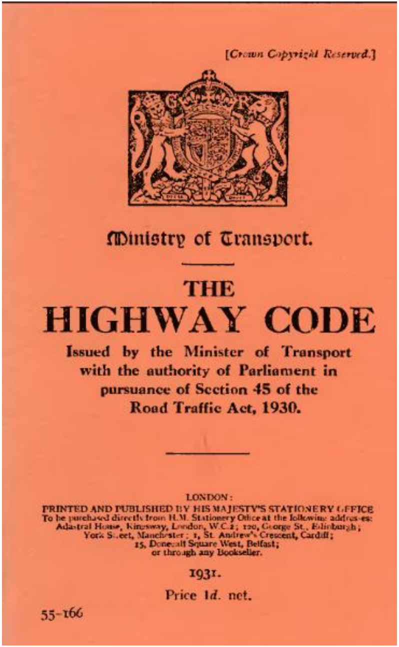
The early days

Prior to the appearance of the first railways in Britain, there was a brief development and interest in steam powered road going vehicles. In 1834, a Mr Hancock started a steam coach called the "Era", carrying up to 14 passengers from Paddington to Regents Park and the City at 6d a head. And in the following year, a Mr Church built an omnibus capable of carrying 40 passengers for the London and Birmingham Steam Carriage Company.

However, the success of the railway movement drove all such traffic off the roads. **A Parliamentary Commission of Enquiry in 1836** reported "strongly in favour of steam carriages on roads", but subsequent Acts of Parliament tended to have a discouraging and restrictive effect. **The Locomotive Act 1861** limited the weight of steam engines to 12 tons and imposed a speed limit of 10 mph.

The Locomotive Act 1865 set a speed limit of 4 mph in the country and 2 mph in towns. The 1865 Act also provided for the famous "man with a red flag". Walking 60 yards ahead of each vehicle, a man with a red flag or lantern enforced a walking pace, and warned horse riders and horse drawn traffic of the approach of a self propelled machine.

The Locomotive Amendment Act 1878 made the red flag optional under local regulations, and


[Crown Copyright Reserved.]

Ministry of Transport.

THE
HIGHWAY CODE

Issued by the Minister of Transport
with the authority of Parliament in
pursuance of Section 45 of the
Road Traffic Act, 1930.

LONDON:
PRINTED AND PUBLISHED BY HIS MAJESTY'S STATIONERY OFFICE
To be purchased directly from H.M. Stationery Office at the following addresses:
Adastral House, Kingsway, London, W.C.2; 120, George St., Edinburgh;
York Street, Manchester; 1, St. Andrew's Crescent, Cardiff;
15, Donegall Square West, Belfast;
or through any Bookseller.

1931.

Price 1*d.* net.

55-166

Image from Wikimedia Commons

Image from Wikimedia Commons

What's Changed (Changing) for Medicine?

- Genotyping.
- Epigenotyping.
- Transcriptome: detailed characterization of phenotype.
  - Stratification of data.

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M.
- Social network data, music information (Spotify), exercise

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M.

- Social network data, music information (Spotify), exercise

Open Data

- Automatic data curation: from curated data to curation of publicly available data.
- Open Data: http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M.

- Social network data, music information (Spotify), exercise

Data Sources

- UK Goverment Stipulation on Data Availability Telegraph Article
- Patient Access: http://www.patient.co.uk/patient-access.asp
- The midata project: Tescos, T-mobile ...
- A social network for personal health?? e.g. EMIS myHealth

Outline

Health

Data Heterogenity

Deep Learning

Conclusions

Missing Data

- If missing at random it can be marginalized.
- As data sets become very large (39 million in EMIS) data becomes extremely sparse.
- Imputation becomes impractical.

Imputation

- Expectation Maximization (EM) is gold standard imputation algorithm.
- Exact EM optimizes the log likelihood.
- Approximate EM optimizes a lower bound on log likelihood.
  - e.g. variational approximations (VIBES, Infer.net).
- Convergence is *guaranteed* to a local maxima in log likelihood.

Expectation Maximization

Require: An initial guess for missing data

Expectation Maximization

Require: An initial guess for missing data
repeat

Expectation Maximization

Require: An initial guess for missing data

repeat

Update model parameters (M-step)

Expectation Maximization

Require: An initial guess for missing data

repeat

Update model parameters (M-step)

Update guess of missing data (E-step)

Expectation Maximization

Require: An initial guess for missing data
  repeat
    Update model parameters (M-step)
    Update guess of missing data (E-step)
  until convergence

Imputation is Impractical

- In very sparse data imputation is impractical.
- EMIS: 39 million patients, thousands of tests.
- For most people, most tests are missing.
- M-step becomes confused by poor imputation.

Direct Marginalization is the Answer

- Perhaps we need joint distribution of two test outcomes,

$$p(y_1, y_2)$$

- Obtained through marginalizing over all missing data,

$$p(y_1, y_2) = \int p(y_1, y_2, y_3, \ldots, y_p) \mathrm{d}y_3, \ldots \mathrm{d}y_p$$

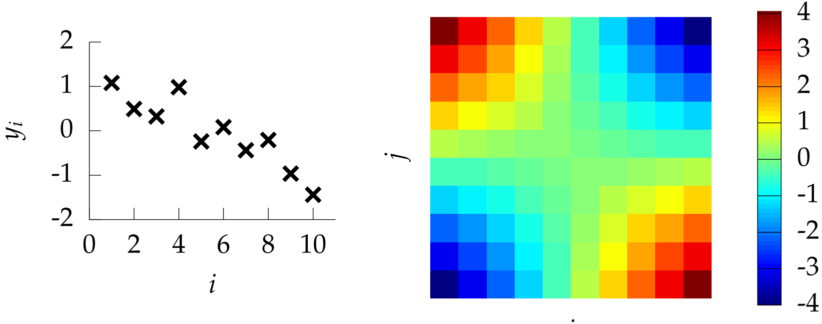
- Where $y_3, \ldots, y_p$ contains:
  1. all tests not applied to this patient
  2. all tests not yet invented!!

Magical Marginalization in Gaussians

Multi-variate Gaussians

- Given 10 dimensional multivariate Gaussian, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.
- Generate a single correlated sample $\mathbf{y} = [y_1, y_2 \dots y_{10}]$.
- How do we find the marginal distribution of y_1, y_2?

Gaussian Marginalization Property

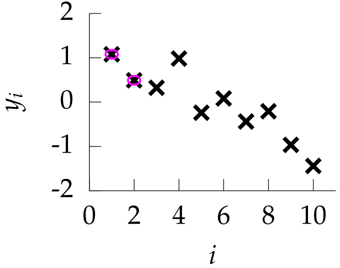


(a) A 10 dimensional sample

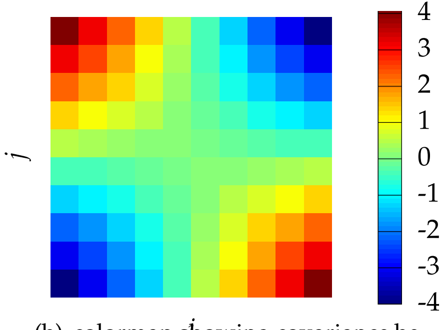
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



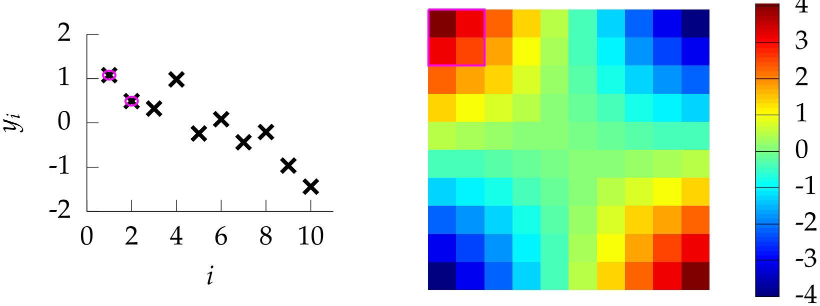
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

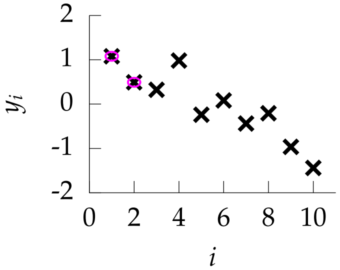


(a) A 10 dimensional sample

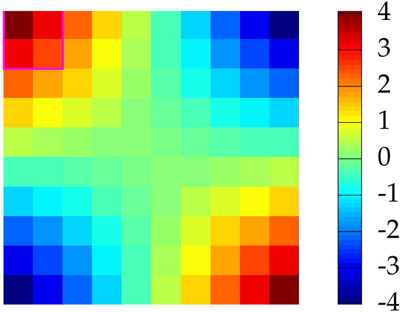
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



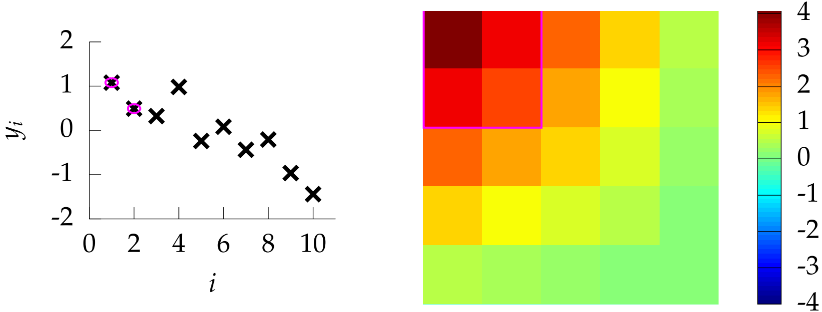
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

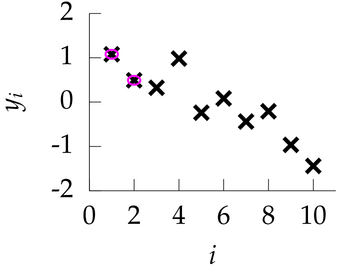


(a) A 10 dimensional sample

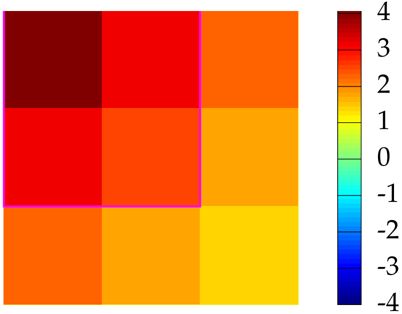
(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



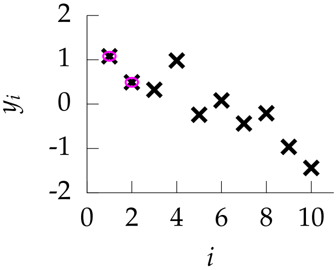
(a) A 10 dimensional sample



(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property

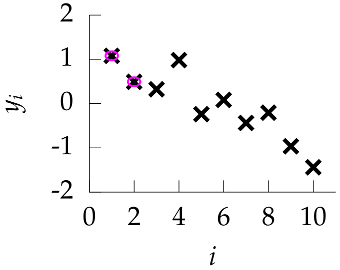


(a) A 10 dimensional sample

$$\begin{bmatrix} 1 & 0.96793 \\ 0.96793 & 1 \end{bmatrix}$$

(b) colormap showing covariance between dimensions.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.



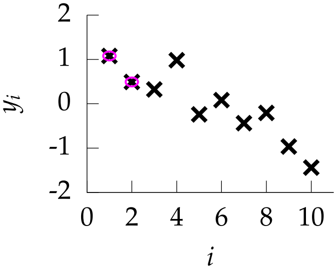
(a) A 10 dimensional sample

$$\begin{bmatrix} 4.1 & 3.1111 \\ 3.1111 & 2.5198 \end{bmatrix}$$

(b) covariance between y_1 and y_2.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Gaussian Marginalization Property



(a) A 10 dimensional sample

$$\begin{bmatrix} 1 & 0.96793 \\ 0.96793 & 1 \end{bmatrix}$$

(b) correlation between y_1 and y_2.

Figure : A sample from a 10 dimensional correlated Gaussian distribution.

Avoid Imputation: Marginalize Directly

- Our approach: Avoid Imputation, Marginalize Directly.
- Explored in context of Collaborative Filtering.
- Similar challenges:
  - many users (patients),
  - many items (tests),
  - sparse data
- Implicitly marginalizes over all future tests too.

Work with Raquel Urtasun (Lawrence and Urtasun, 2009) **and recent submission with Nicoló Fusi.**

Methods that Interrelate Covariates

- Need Class of models that interrelates data.
- Common assumption: high dimensional data lies on low dimensional manifold.
- Want to retain the marginalization property of Gaussians.

Linear Dimensionality Reduction

Linear Latent Variable Model

- Represent data, $\mathbf{Y}$, with a lower dimensional set of latent variables $\mathbf{X}$.
- Assume a linear relationship of the form

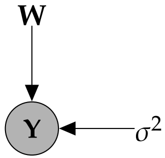
$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\epsilon}_{i,:},$$

where

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)



$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{W}\mathbf{W}^{\top} + \sigma^2\mathbf{I}\right)$$

Linear Latent Variable Model II

Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_{i,:}|\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

$$\log p(\mathbf{Y}|\mathbf{W}) = -\frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{Y}^\top\mathbf{Y}\right) + \text{const.}$$

If $\mathbf{U}_q$ are first q principal eigenvectors of $n^{-1}\mathbf{Y}^\top\mathbf{Y}$ and the corresponding eigenvalues are $\boldsymbol{\Lambda}_q$,

$$\mathbf{W} = \mathbf{U}_q\mathbf{L}\mathbf{R}^\top, \quad \mathbf{L} = \left(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I}\right)^{\frac{1}{2}}$$

where $\mathbf{R}$ is an arbitrary rotation matrix.

Dealing with Non Gaussian Data

- Marginalization property of Gaussians very attractive.
- How to incorporate non-Gaussian data?
  - Data which isn't missing at random.
  - Binary data.
  - Ordinal categorical data.
  - Poisson counts.
  - Outliers.

Project Back into Gaussian

- Combine non-Gaussian likelihood with Gaussian prior.
- Either:
  - Project back to Gaussian posterior that is nearest in KL sense.
  - Expectation propagation.

- Or:
  - Fit a locally valid Gaussian approximation.
  - Laplace Approximation.

Ongoing work with Ricardo Andrade Pacheco (EP) and Alan Saul (Laplace) also James Hensman.

Gaussian Noise

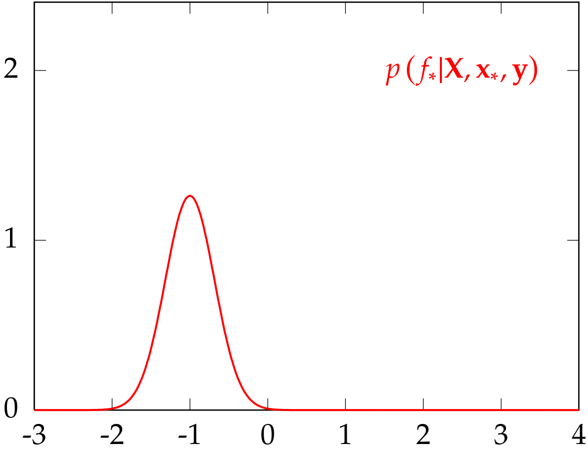


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

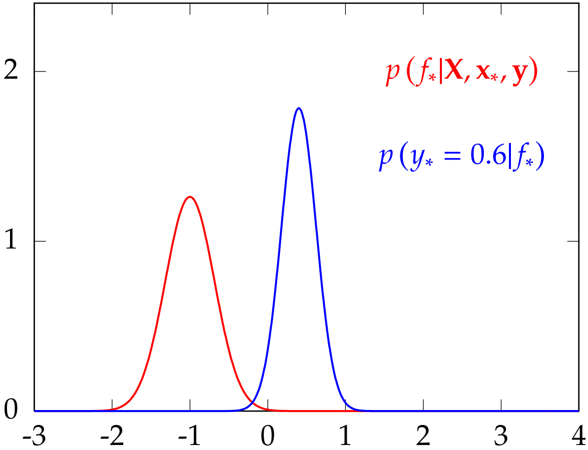


Figure : Inclusion of a data point with Gaussian noise.

Gaussian Noise

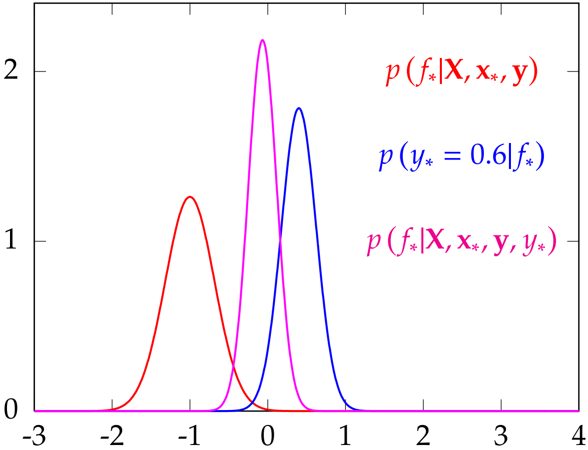


Figure : Inclusion of a data point with Gaussian noise.

Classification Noise Model

Probit Noise Model

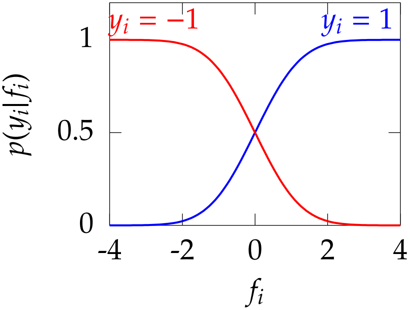


Figure : The probit model (classification). The plot shows $p\left(y_i|f_i\right)$ for different values of y_i. For $y_i = 1$ we have
$p\left(y_i|f_i\right) = \phi\left(f_i\right) = \int_{-\infty}^{f_i} \mathcal{N}\left(z|0,1\right) \mathrm{d}z.$

Classification

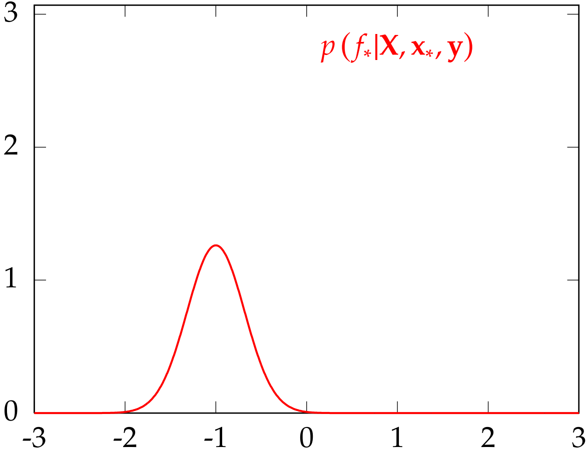


Figure : An EP style update with a classification noise model.

Classification

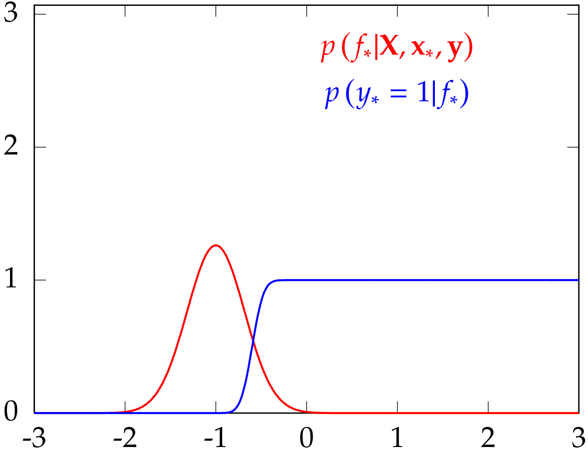


Figure : An EP style update with a classification noise model.

Classification

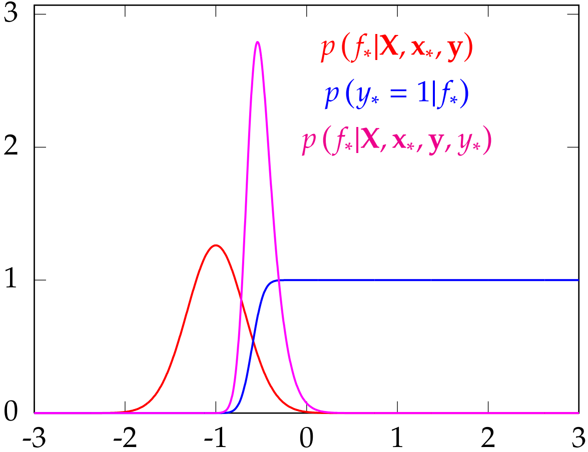


Figure : An EP style update with a classification noise model.

Classification

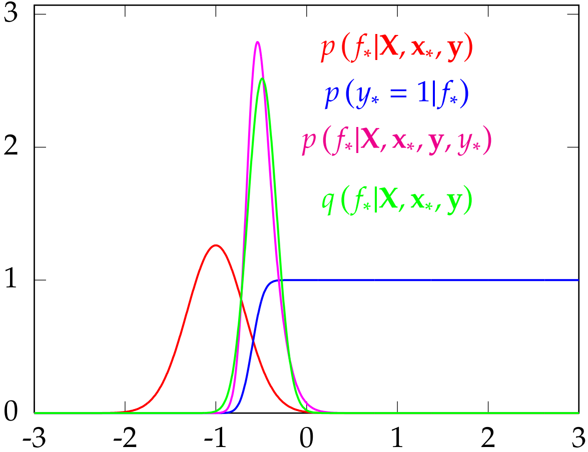


Figure : An EP style update with a classification noise model.

Ordinal Noise Model

Ordered Categories

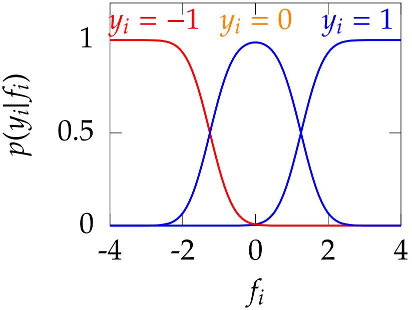


Figure : The ordered categorical noise model (ordinal regression). The plot shows $p\left(y_i|f_i\right)$ for different values of y_i. Here we have assumed three categories.

Ordinal Regression

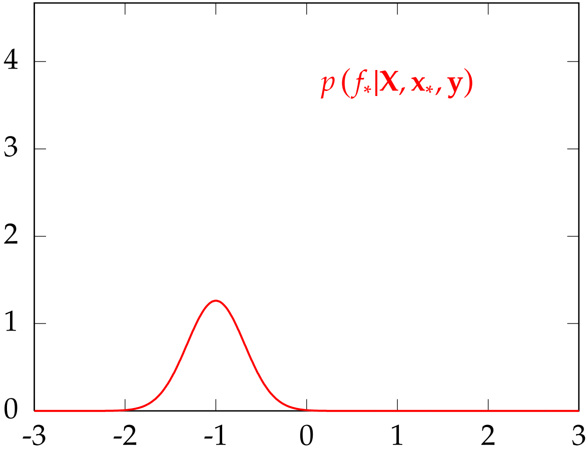


Figure : An EP style update with an ordered category noise model.

Return

Ordinal Regression

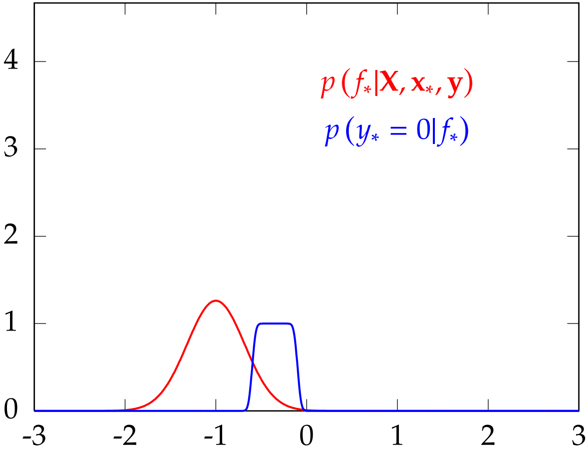


Figure : An EP style update with an ordered category noise model.

Return

Ordinal Regression

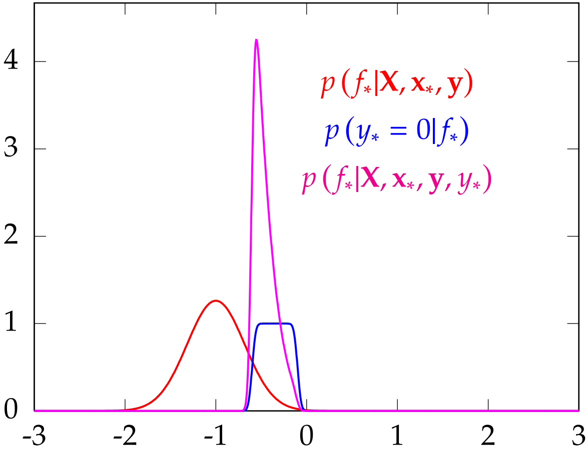


Figure : An EP style update with an ordered category noise model.

Return

Ordinal Regression

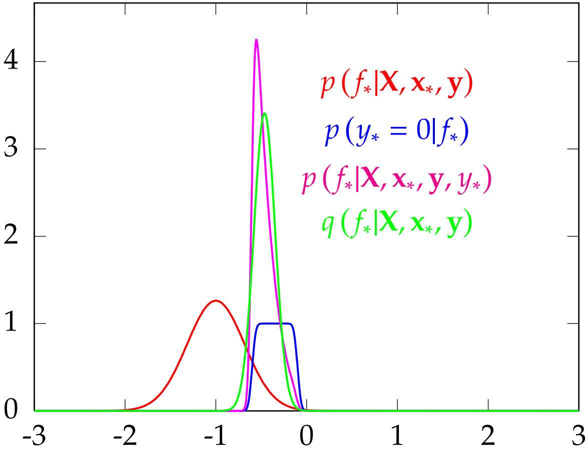


Figure : An EP style update with an ordered category noise model.

Other Challenges

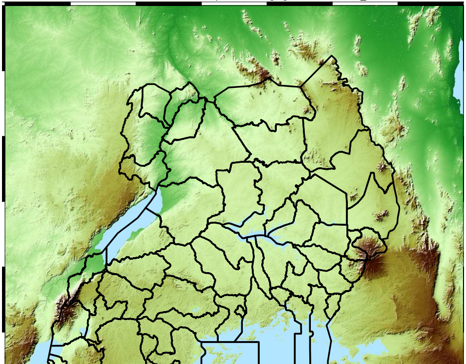
- Spatial Data (workshop in November with Peter Diggle, work with Ricardo Andrade Pacheco and John Quinn's group).
- Survival Data (work with Alan Saul and Aki Vehtari's group and HeRC).
- Image Data (work with Teo de Campos, Violet Snell and imminent arrival of Zhenwen Dai)
- Text Data (planned collaboration with Trevor Cohn)

Example: Prediction of Malaria Incidence in Uganda

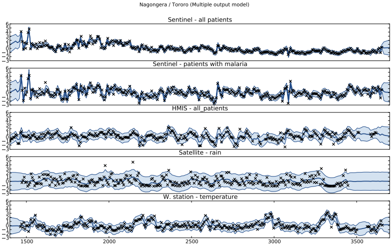
- Work with John Quinn and Martin Mubaganzi (Makerere University, Uganda)
- See http://cit.mak.ac.ug/cs/aigroup/.

Malaria Prediction in Uganda

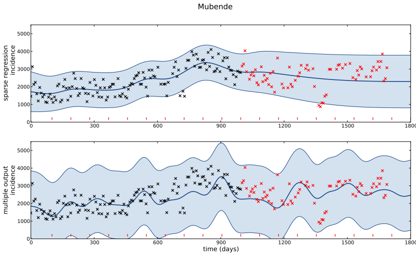
Data SRTM/NASA from http://dds.cr.usgs.gov/srtm/version2_1



Malaria Prediction in Uganda



Malaria Prediction in Uganda



Visit to Uganda

Outline

Scientists See Promise in Deep-Learning Programs

Hao Zhang/The New York Times

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese.

By JOHN MARKOFF
Published: November 23, 2012

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

点击查看本文中文版.

Connect With

The advances have led to widespread enthusiasm among researchers who design software to perform human

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?

Log In With Facebook

What's Popular Now

King Abdullah of Jordan Has Criticism for All Concerned


7 Marines Killed in Nevada Training Exercise


MOST E-MAILED | MOST VIEWED


1. WELL
Lost Sleep Can Lead to Weight Gain


2. THIS LIFE
The Stories That Bind Us


3. WELL
A New Approach to Hip Surgery


4. Unwanted Electronic Gear Rising in Toxic Piles


5. DAVID BROOKS
The Progressive Shift


6. CONTINUING EDUCATION SPECIAL SECTION
A Gray Jobs Market for All Ages


7. Vatican's Bureaucracy Tests Even the Infallible

Is "Deep Learning" a Revo × | Scientists See Advances ir × | graphics - Is it possible to ×

newyorker.com/online/blogs/newsdesk/2012/11/is-deep-learning-a-revolution-in-artificial-intelligence.html

My Boosters | LastPass - Dow... | Add to Triplt | Proverbi napol... | IEEE Xplore - On... | Google Maps | Other Bookmarks

THE NEW YORKER

SUBSCRIBE and save up to 88%

- SUBSCRIBE
- RENEW
- GIVE A GIFT
- INTERNATIONAL ORDERS
- ONLINE ARCHIVE

NEWS DESK

Reporting the latest on Washington and the world.

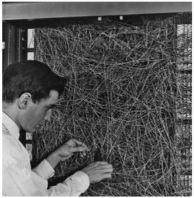
« How Susan Rice Sees the World | Main | Moral Machines »

NOVEMBER 25, 2012

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

POSTED BY GARY MARCUS

Share 677 | Tweet 361 | +1 | PRINT | MORE | 9 COMMENTS



Can a new technique known as deep learning revolutionize artificial intelligence, as yesterday's front-page article at the New York *Times* suggests? There is good reason to be excited about deep learning, a sophisticated "machine learning" algorithm that far exceeds many of its predecessors in its abilities to recognize syllables and images. But there's also good reason to be skeptical. While the *Times* reports that "advances in an artificial intelligence technology that can recognize patterns

WELCOME

SIGN IN | HELP | REGISTER

Search Web site | Find

MOST POPULAR | MOST E-MAILED | THIS ISSUE

1. Andy Borowitz: Cheney Marks Tenth Anniversary of Pretending There Was Reason to Invade Iraq
2. Amy Davidson: Life After Steubenville
3. William Finnegan: Gina Rinehart, Australia's Mining Billionaire
4. Maria Bustillos: On Video Games and Storytelling: An Interview with Tom Bissell
5. Lena Dunham: Lifelong Canine Cravings

THE NEW YORKER SUBSCRIBE TODAY! CLICK HERE

THE NEW YORKER DIGITAL

TABLET, MOBILE, AND MORE

Newsletter sign-up: Enter e-mail address | Submit

Google Hires Geoffrey Hir × | Google Hires Brains that × | Geoffrey Hinton - Google ×
www.seroundtable.com/google-hires-geoffrey-hinton-16499.html
My Boosters | LastPass - Dow... | Add to TripIt | Proverbi napol... | IEEE Xplore - On... | Google Maps | Other Bookmarks

Google To Expand Knowledge Graph Through Hire Of Geoffrey Hinton

‹ PREV STORY NEXT STORY ›

49 Tweet | 10 +1 | 16 Like | SHARE

Mar 14, 2013 • 8:23 am | (10)
by Barry Schwartz | Filed Under Google Search Engine

If I had to place one search priority above all else, I'd say right now, Google's most ambitious project is the knowledge graph. Yea, they are pushing Google+ big time, but the knowledge graph is a level above all of that technically.

Of course, Google has an outstanding team working on this project lead by one of the smartest people I've ever met Amit Singhal.

To take the knowledge graph to the next level, Google has hired/acquired Geoffrey Hinton and his team at DNNresearch. Geoffrey posted a note on his Google+ page about it:

> Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.

I know we just scratched the surface of the knowledge graph and I am excited to see where it takes us in the future.

I am just glad I don't have to figure out how to get us there. I get to just sit and enjoy the ride.

SUBSCRIBE
Enter Email Address | Subscribe Now
SUBSCRIBE OPTIONS ›
ADVERTISERS

SEARCH BUZZ VIDEO Subscribe

SUBSCRIBE › MORE VIDEOS › VIDEO DETAILS ›

ROUNDTABLE SPONSORS

BROWSE BY:
- Browse by Date
- Find by Category
- Discover by Author
- Scan Most Recent
- See Comments
- View Tag Cloud

SEM FORUM THREADS

WebmasterWorld Forums

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 6:30 AM

Follow @bobmcmillan

Like 270 | Tweet 216 | +1 114 | Share 34

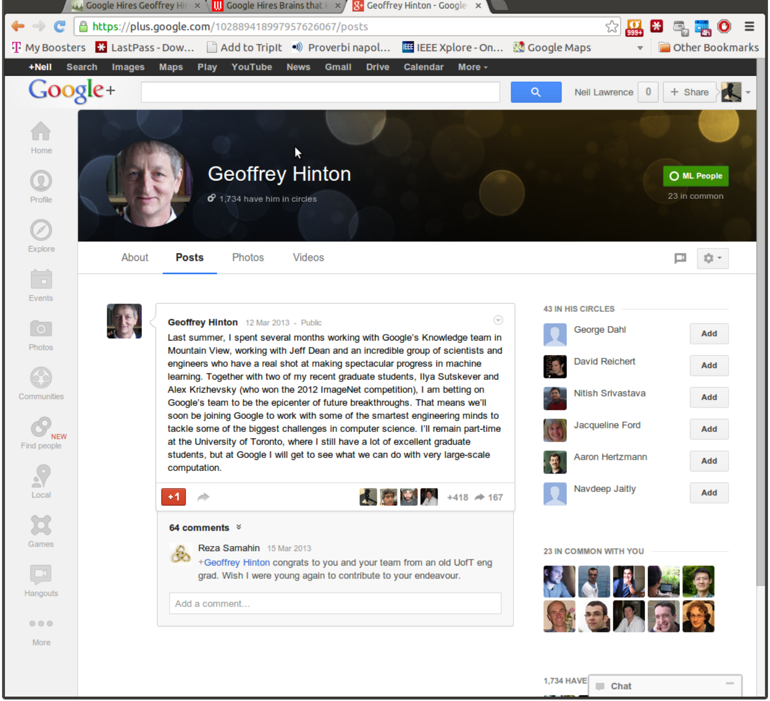
MOST RECENT WIRED POSTS


Jawbone's Up Fitness Band Is Now Android-Compatible


Review: Ecovacs Winbot, a Window-Cleaning Robot


Sherlock, Professor X and Margaery Tyrell Team for Neil Gaiman Radio Play


Video: Robo-Chopper Dives and Grabs Objects Like a Bird of Prey


Google Hires Geoffrey Hir
Google Hires Brains that
Geoffrey Hinton - Google
https://plus.google.com/102889418997957626067/posts
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+Neil Search Images Maps Play YouTube News Gmail Drive Calendar More
Google+
Neil Lawrence
0
Share
Home
Profile
Explore
Events
Photos
Communities
NEW
Find people
Local
Games
Hangouts
More
Geoffrey Hinton
1,734 have him in circles
ML People
23 in common
About
Posts
Photos
Videos
Geoffrey Hinton 12 Mar 2013 - Public
Last summer, I spent several months working with Google's Knowledge team in Mountain View, working with Jeff Dean and an incredible group of scientists and engineers who have a real shot at making spectacular progress in machine learning. Together with two of my recent graduate students, Ilya Sutskever and Alex Krizhevsky (who won the 2012 ImageNet competition), I am betting on Google's team to be the epicenter of future breakthroughs. That means we'll soon be joining Google to work with some of the smartest engineering minds to tackle some of the biggest challenges in computer science. I'll remain part-time at the University of Toronto, where I still have a lot of excellent graduate students, but at Google I will get to see what we can do with very large-scale computation.
+1
+418
167
64 comments
Reza Samahin 15 Mar 2013
+Geoffrey Hinton congrats to you and your team from an old UofT eng grad. Wish I were young again to contribute to your endeavour.
Add a comment...
43 IN HIS CIRCLES
George Dahl
Add
David Reichert
Add
Nitish Srivastava
Add
Jacqueline Ford
Add
Aaron Hertzmann
Add
Navdeep Jaitly
Add
23 IN COMMON WITH YOU
1,734 HAVE
Chat

direction for further research.

11.1. HAVE WE THROWN THE BABY OUT WITH THE BATH WATER?

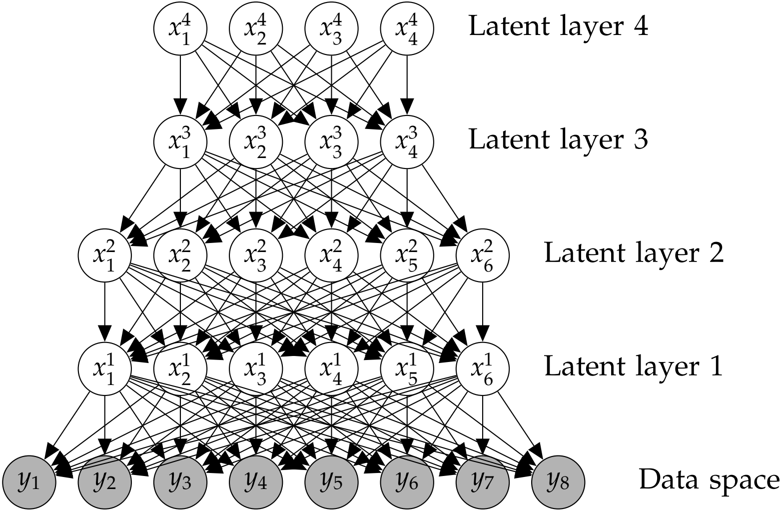
According to the hype of 1987, neural networks were meant to be intelligent models which discovered features and patterns in data. Gaussian processes in contrast are simply smoothing devices. How can Gaussian processes possibly replace neural networks? What is going on?

I think what the work of Williams and Rasmussen (1996) shows is that many real–world data modelling problems are perfectly well solved by sensible smoothing methods. The most interesting problems, the task of feature discovery for example, are not ones which Gaussian processes will solve. But maybe multilayer perceptrons can't solve them either. On the other hand, it may be that the limit of an infinite number of hidden units, to which Gaussian processes correspond, was a bad limit to take; maybe we should backtrack, or modify the prior on neural network parameters, so as to create new models more interesting than Gaussian processes. Evidence that this infinite limit has lost something compared with finite neural networks comes from the observation that in a finite neural network with more than one output, there are non–trivial correlations between the outputs (since they share inputs from common hidden units); but in the limit of an infinite number of hidden units, these correlations vanish. Radford Neal has suggested the use of non–Gaussian priors in networks with multiple hidden layers. Or perhaps a completely fresh start is needed, approaching the problem of machine learning from a paradigm different from the supervised feedforward mapping.

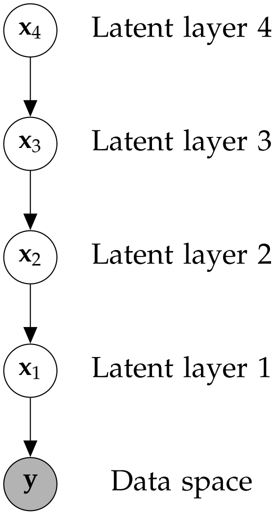
Structure of Priors

MacKay: NIPS Tutorial 1997 “Have we thrown out the baby with the bathwater?” (Published as MacKay, 1998) Also noted by (Wilson et al., 2012)

Deep Models

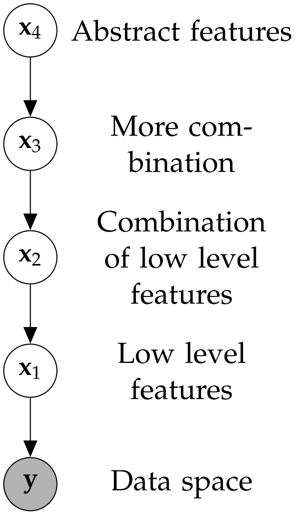

x_1^4 x_2^4 x_3^4 x_4^4
Latent layer 4
x_1^3 x_2^3 x_3^3 x_4^3
Latent layer 3
x_1^2 x_2^2 x_3^2 x_4^2 x_5^2 x_6^2
Latent layer 2
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1 x_6^1
Latent layer 1
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8
Data space

Deep Models



$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Deep Models

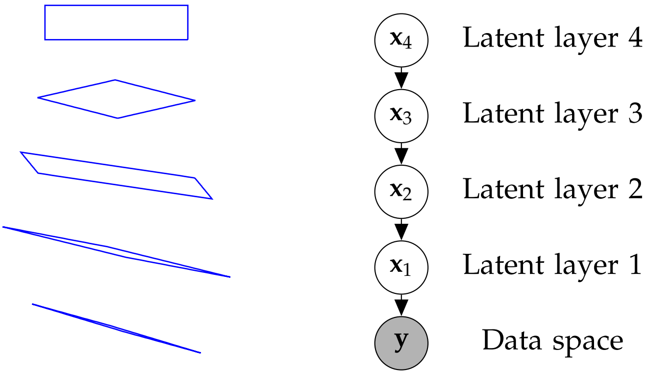

$\mathbf{x}_4$
Abstract features
$\mathbf{x}_3$
More combination
$\mathbf{x}_2$
Combination of low level features
$\mathbf{x}_1$
Low level features
$\mathbf{y}$
Data space

Deep Gaussian Processes

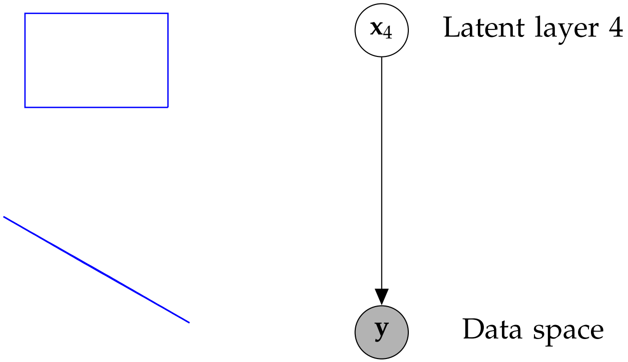
Damianou and Lawrence (2013)

- Deep architectures allow abstraction of features (Bengio, 2009; Hinton and Osindero, 2006; Salakhutdinov and Murray, 2008).
- We use variational approach to stack GP models.

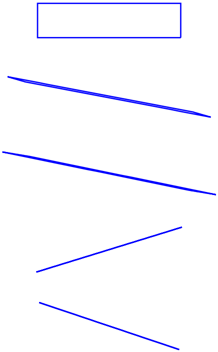
Stacked PCA

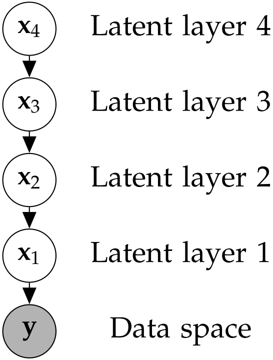

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA



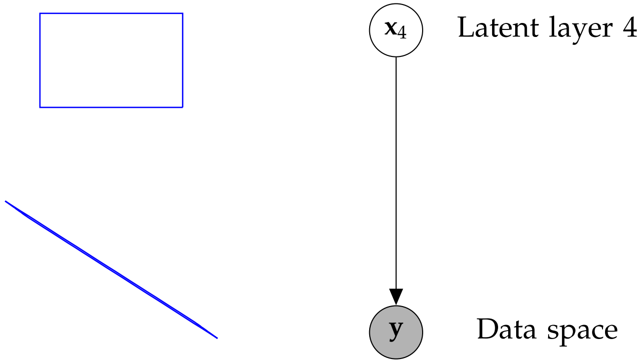
Stacked PCA



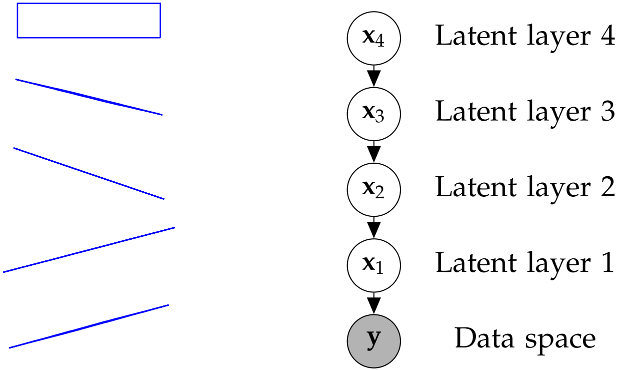


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA

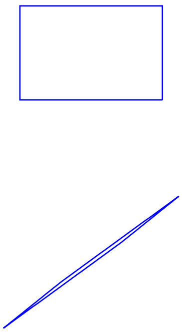


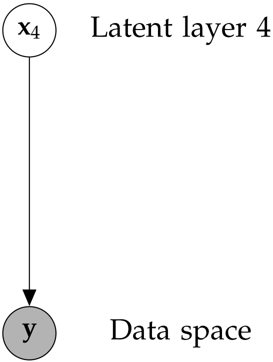
Stacked PCA



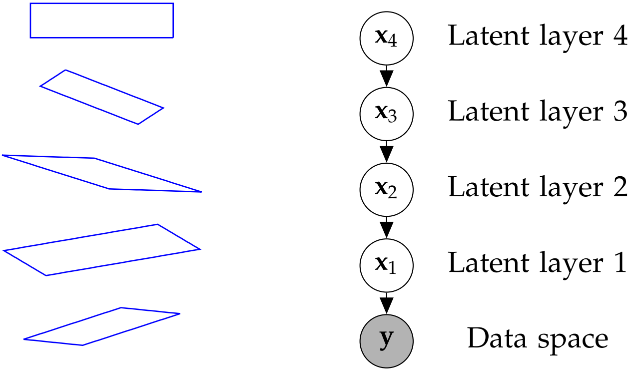
$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA

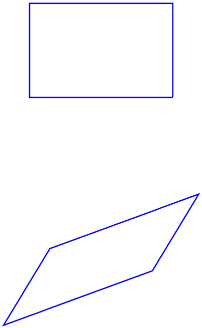


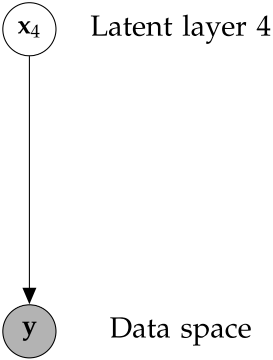


Stacked PCA

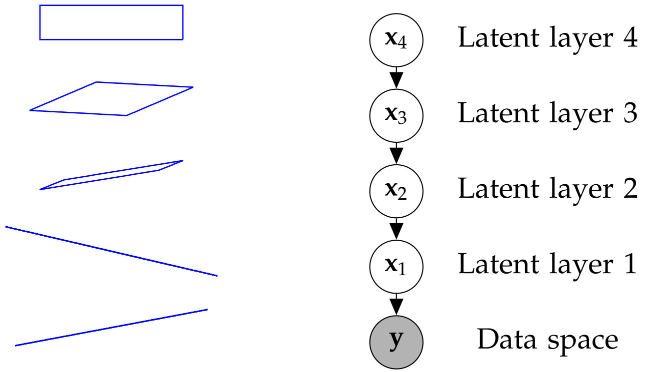

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA

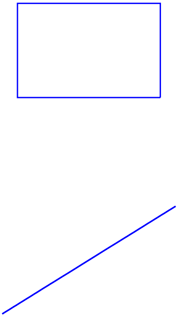


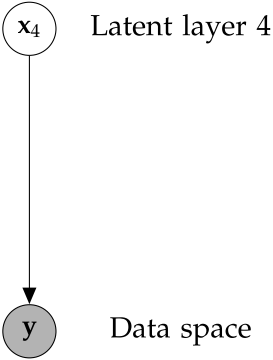


Stacked PCA

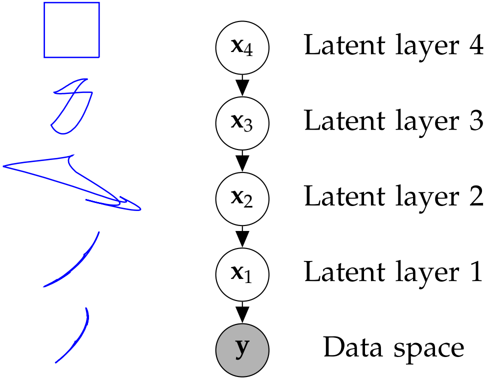

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked PCA



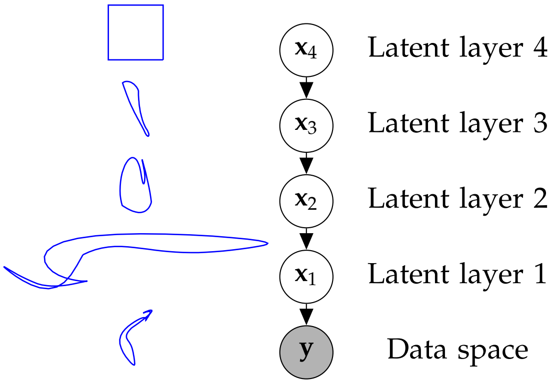


Stacked GPs



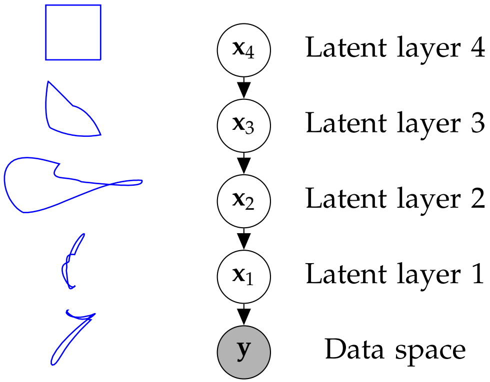
$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

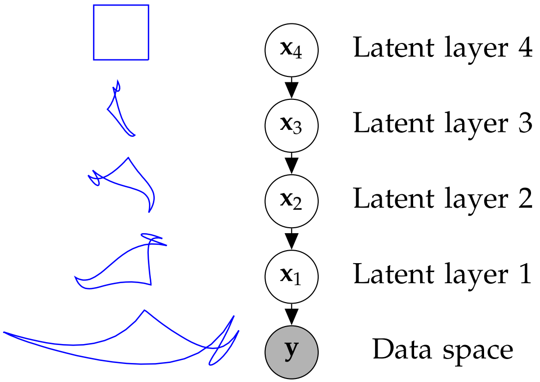


$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

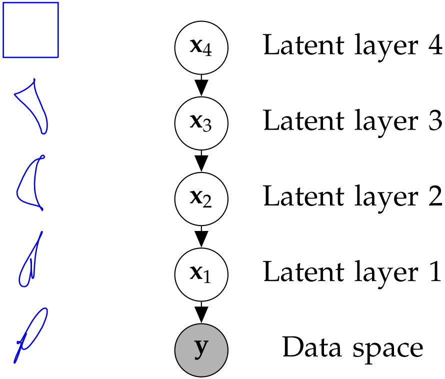

$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs



$\mathbf{x}_4$
Latent layer 4
$\mathbf{x}_3$
Latent layer 3
$\mathbf{x}_2$
Latent layer 2
$\mathbf{x}_1$
Latent layer 1
$\mathbf{y}$
Data space

Stacked GPs

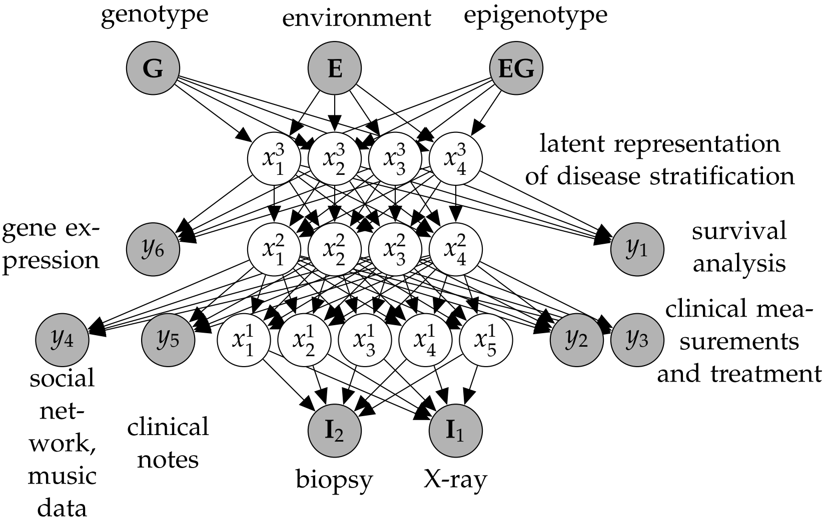


$\mathbf{x}_4$ Latent layer 4
$\mathbf{x}_3$ Latent layer 3
$\mathbf{x}_2$ Latent layer 2
$\mathbf{x}_1$ Latent layer 1
$\mathbf{y}$ Data space

What Can We Do that Google Can't?

- Google's resources give them access to volumes of data (or Facebook, or Microsoft, or Amazon).
- Is there anything for Universities to contribute?
- Assimilation of multiple views of the patient: each perhaps from a different patient.
- This may be done by small companies (with support of Universities).
- A Facebook app for your personalised health.
- These methodologies are part of that picture.

Deep Health



genotype
environment
epigenotype
G
E
EG
x_1^3 x_2^3 x_3^3 x_4^3
latent representation of disease stratification
gene expression
y_6
x_1^2 x_2^2 x_3^2 x_4^2
y_1
survival analysis
clinical measurements and treatment
y_4
y_5
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1
y_2
y_3
social network, music data
clinical notes
$\mathbf{I}_2$
$\mathbf{I}_1$
biopsy
X-ray

Deep Health: Power Ranger Model of Research

GPy: A Gaussian Process Framework in Python

- BSD Licensed software base.
- Wide availability of libraries, "modern" scripting language.
- Allows us to set projects to undergraduates in Comp Sci that use GPs.
- Available through GitHub https://github.com/SheffieldML/GPy
- Reproducible Research with IPython Notebook.

Probable Features for Next Release

- Non-Gaussian likelihoods.
- Multivariate outputs.
- Dimensionality reduction.
- Approximations for large data sets.
- Probabilistic-style programming (specify the model, not the algorithm).
- A range of covariance functions, ability to implement your own symbolically.
- Data missing at random.

Planned Features for Future Releases

- Deep models.
- Massive data: millions of points.
- Data missing not at random.
- Deep models for massive data.

GPSS: Gaussian Process Summer School

- http://ml.dcs.shef.ac.uk/gpss/
- Next one is a Winter School in Sheffield 13th-15th January 2014.
- To be followed by a workshop on Temporospatial Modelling (16th January).
- Ran a 'Gaussian Process Roadshow' at Makerere in Uganda this August.
- Will run a Roadshow in Colombia in February.
- Tentative plans for a Roadshow in Kenya in 2015.

Summary

- Gaussian models good for missing data.
- Disparate data types handled with EP and Laplace.
- Deep models allow complex abstract representation of data sets at higher levels.
- Current limitation is on data set size.
- Addressing this through work by James Hensman on Stochastic Variational Inference for GPs (recent UAI paper).
- Intention is to deploy these models for assimilating a wide range of data types in personalized health (text, survival times, images, genotype, phenotype).
- Requires population scale models with millions of features.

References I

Y. Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009. ISSN 1935-8237. [DOI].

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In A. Nicholson and P. Smyth, editors, *Uncertainty in Artificial Intelligence*, volume 29. AUAI Press, 2013.

G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.

J. Langford and J. Pineau, editors. *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman.

N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with Gaussian processes. In L. Bottou and M. Littman, editors, *Proceedings of the International Conference in Machine Learning*, volume 26, San Francisco, CA, 2009. Morgan Kauffman. [PDF].

D. J. C. MacKay. Introduction to Gaussian Processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *Series F: Computer and Systems Sciences*, pages 133–166. Springer-Verlag, Berlin, 1998.

R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In S. Roweis and A. McCallum, editors, *Proceedings of the International Conference in Machine Learning*, volume 25, pages 872–879. Omnipress, 2008.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6 (3):611–622, 1999. [PDF]. [DOI].

A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In Langford and Pineau (2012).