Data Mining in Dynamic Environments

Hyoung-joo Lee

Department of Engineering Science / Zoology, University of Oxford
hjlee@robots.ox.ac.uk

26 August, 2009

Outline

Outline

Data Mining Preliminaries

Dynamic Prediction

Dynamic Classification

A Few Others...

Conclusions

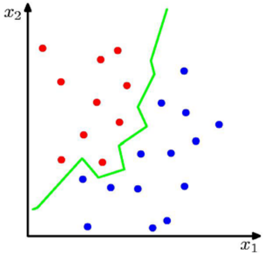
From Real World to Data Mining

Real world	Data Mining
A system	A model $\mathcal{M}$
Characteristics	Parameters $\boldsymbol{\theta}$
Observations	Data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$
Condition	Input $\mathbf{x}_i$
Behaviour	Output $\mathbf{y}_i$
Analysis	Inference
	Estimating $\boldsymbol{\theta}$ (descriptive)
	Predicting $\mathbf{y}^*$, given $\mathbf{x}^*$ (predictive)

Canonical Problems and Applications

Classification

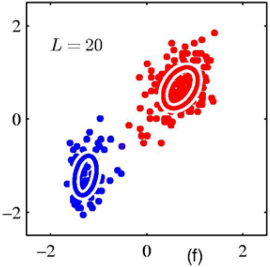
- Handwriting recognition
- Speech recognition
- Direct marketing



(b)

Clustering

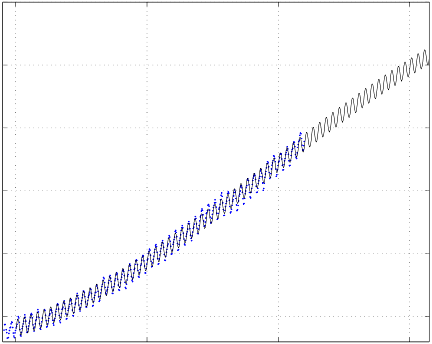
- Customer segmentation
- Webpage clustering



(f)

Regression / Prediction

- Stock price prediction
- Weather forecasting



Why Data Mining?

Because a real world system is often...

- Complex (high degrees of freedom)
- Subtle (difficult to describe expertise explicitly)
- Uncertain
- Noisy
- Dynamic

We cannot take care of all of them

- Using prior knowledge
- Relaxations on some of them
- Approximation

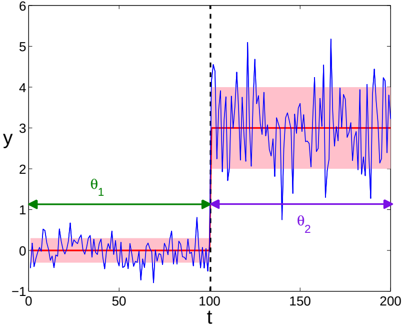
Dynamic modelling (↔ static modelling)

- System characteristics (model paramters) change over time
- To detect and adapt to changes

Some issues

- Flexibility vs. stability
- Uncertainty

  → Bayesian



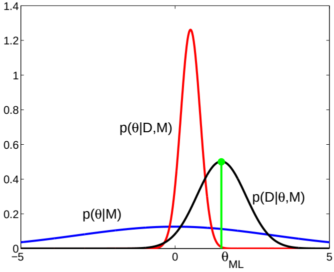
Bayesian Inference

Bayes' theorem

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})},$$

$$\text{where } p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})\mathrm{d}\boldsymbol{\theta}$$

- Updating $p(\boldsymbol{\theta}|\mathcal{M})$ to $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$
- cf) Maximum likelihood $\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$



Outline

Outline

Gaussian Distribution

$$\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}$$

Some properties

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

- A marginal of a Gaussian is Gaussian

  $$p(\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

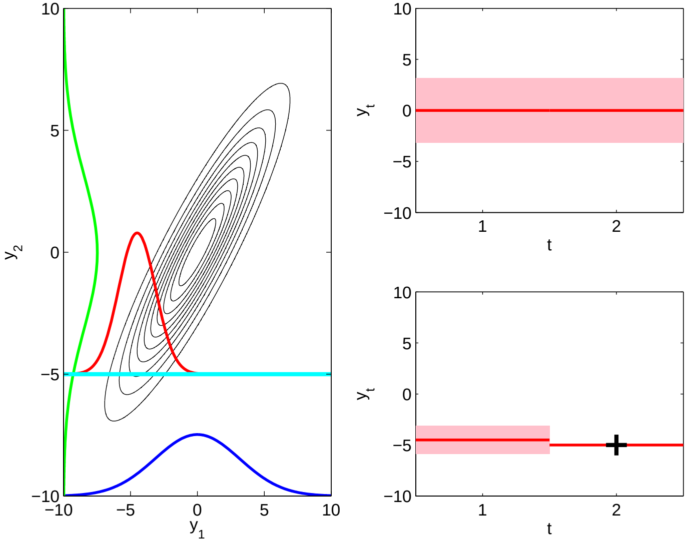
- A conditional of a Gaussian is Gaussian

  $$p(\mathbf{y}_2|\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_2; \boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}),$$

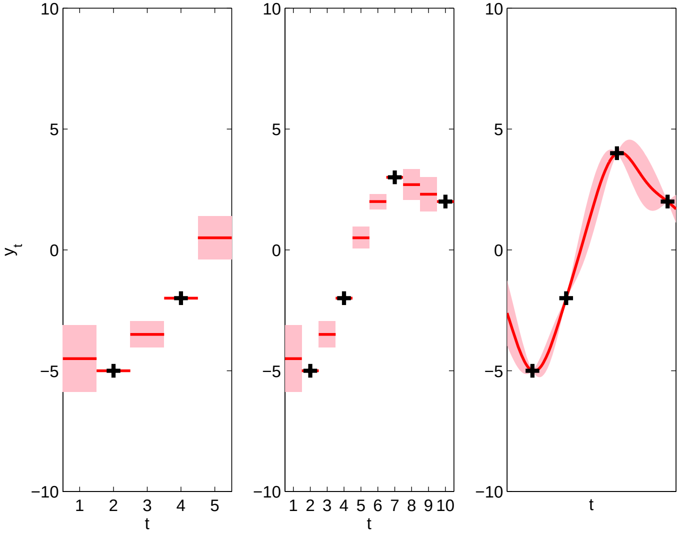
  $$\text{where } \boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),$$

  $$\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

2-D example



What If We Are Going High-dimensional?



GP Regression

Data

- Training data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Test data $\{\mathbf{x}_j^*\}_{j=1}^{n^*}$

GP prediction

- A function $\mathbf{y}_j^* = f(\mathbf{x}_j^*) + \varepsilon_j^*$
- A Gaussian distribution over the function values

$$p\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix}; \mathbf{0}, \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}\right)$$

- Prior $p(\mathbf{f}^*) = \mathcal{N}(\mathbf{f}^*; \mathbf{0}, \mathbf{K}_{22})$
- Posterior $p(\mathbf{f}^*|\mathbf{y}) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \mathbf{K}^*)$
  where $\boldsymbol{\mu}^* = \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\mathbf{y}$
  $\mathbf{K}^* = \mathbf{K}_{22} - \mathbf{K}_{21}\mathbf{K}_{11}^{-1}\mathbf{K}_{12}$

Covariance Function

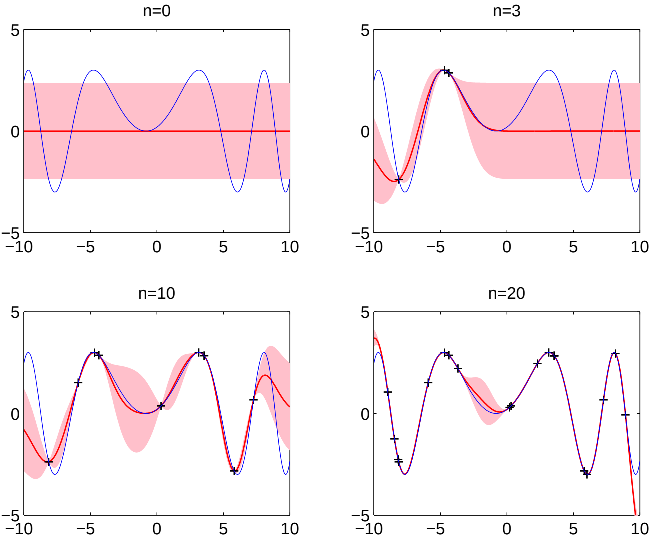
Covariance matrix

- Defined by a covariance function
- Close inputs $\Rightarrow$ similar function values
- Equivalent to kernel functions in SVMs

Some popular covariance functions

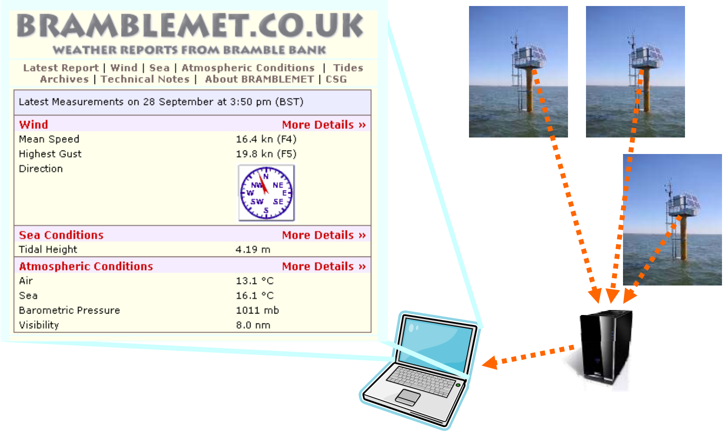
- Squared exponential $k(\mathbf{x}, \mathbf{x}') = \gamma^2 \exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right\}$
- Periodic $k(\mathbf{x}, \mathbf{x}') = \gamma^2 \exp\left\{-2\frac{\sin^2(\pi(\mathbf{x}-\mathbf{x}'))}{2\sigma^2}\right\}$
- Many others
- Hyperparameters: σ (input scale), γ (output scale)
- Combinations of different covariance functions

Examples of GP Regression



Outline

A network of wireless weather sensors on the South Coast



Weather Sensor Prediction

Prediction with Gaussian processes

- Training data $\{(t_i, y_i)\}$
- Prediction $f(t^*)$ on a sensor reading at t^*

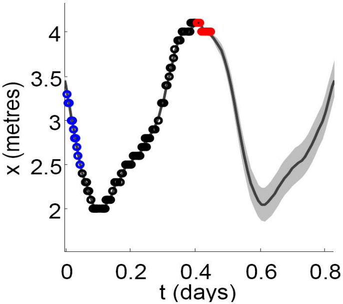
  $y^* = f(t^*) + \varepsilon^*$
- Extrapolation

Some issues

- Prediction with censored observations
- Active data sampling

Dynamic Prediction using Gaussian Processes

- Adaptively update predictions
- Moving windows
  - Adding new observations
  - Discarding uninformative, old observations
- Efficient using matrix tricks (e.g. Cholesky decomposition)



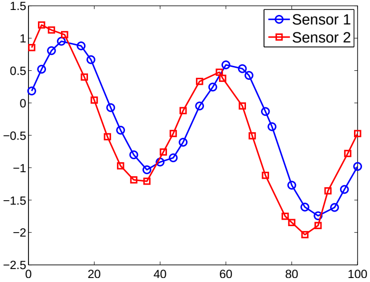
Prediction with Censored Observations

Censored observations

- Sensor faults
- Maintenance

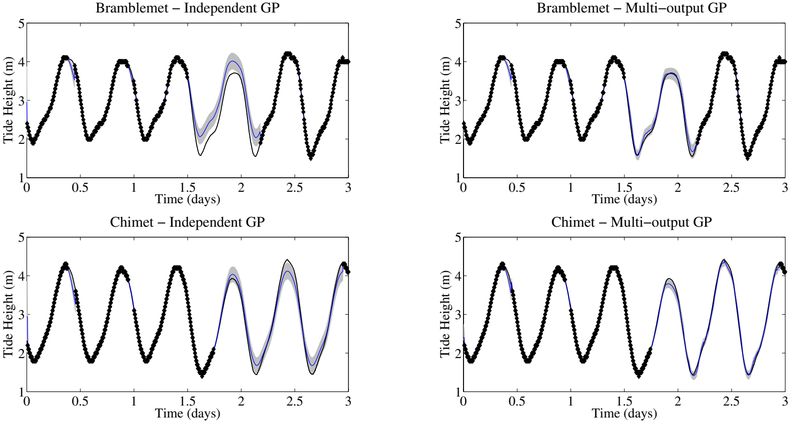
Delayed correlation between sensors

- Assuming a Gaussian distribution over sensor predictions
- Modifying one prediction considering predictions of others

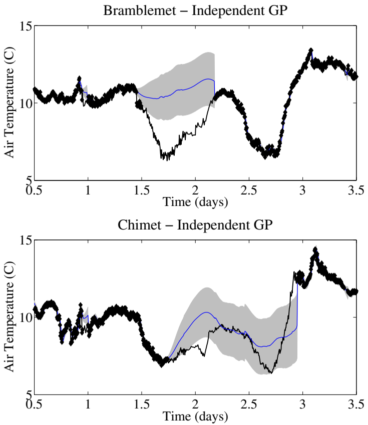


Prediction with Censored Observations

Tide heights



Air temperatures



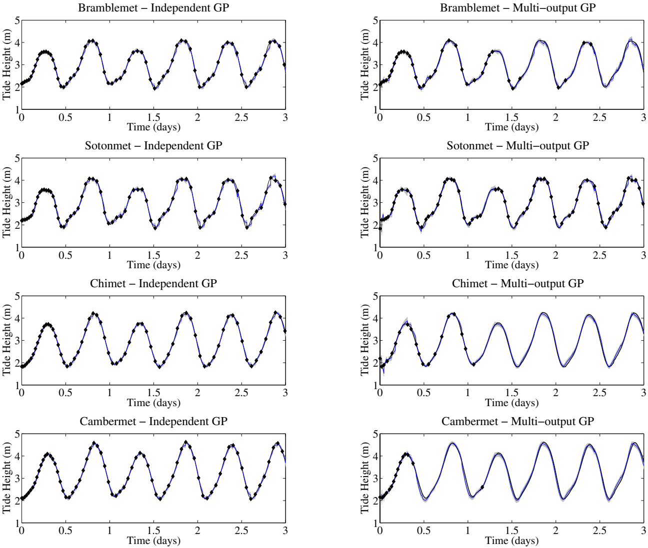
Active Sampling

Limited battery life

Selecting observations actively

- What obervations will be the most informative?
  - Which sensor to observe
  - When to observe
- Criteria
  - As few data as possible
  - Keeping accuracy intact ($\approx$ minimising uncertainty)

Active Sampling

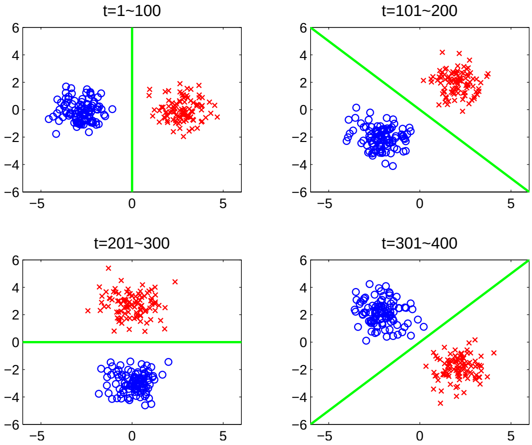


Outline

Outline

Dynamic Classification

- A decision boundary changes over time
- Adaptively update the boundary

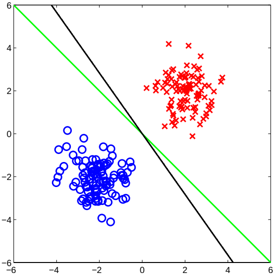


Logistic Regression

- Probability of $y_t = 1$ given an input vector $\mathbf{x}_t$
- Parameter $\mathbf{w}$ specifying the decision boundary

$$z_t = \mathbf{x}_t^\top \mathbf{w} + v_t,$$

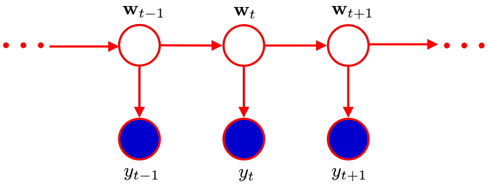
$$p(y_t = 1|\mathbf{x}_t) = \frac{1}{1 + \exp(-z_t)}$$



$$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\hat{\mathbf{w}} = \begin{bmatrix} 1.15 \\ 0.82 \end{bmatrix}$$

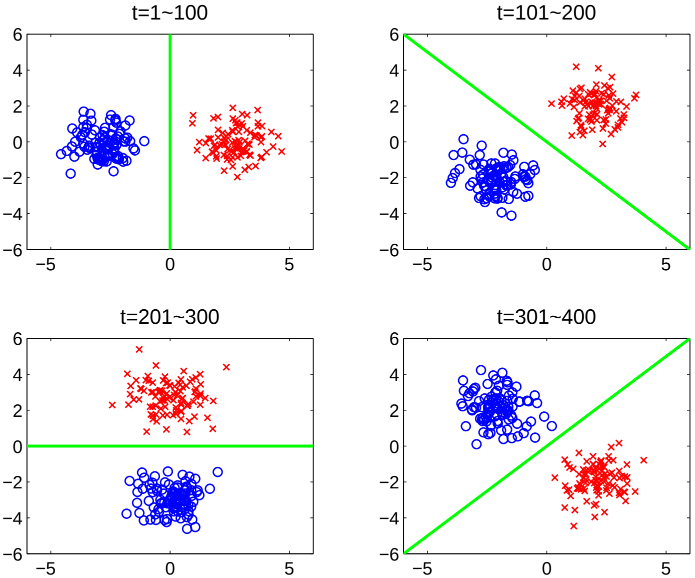
Dynamic Logistic Regression Using State Space Models



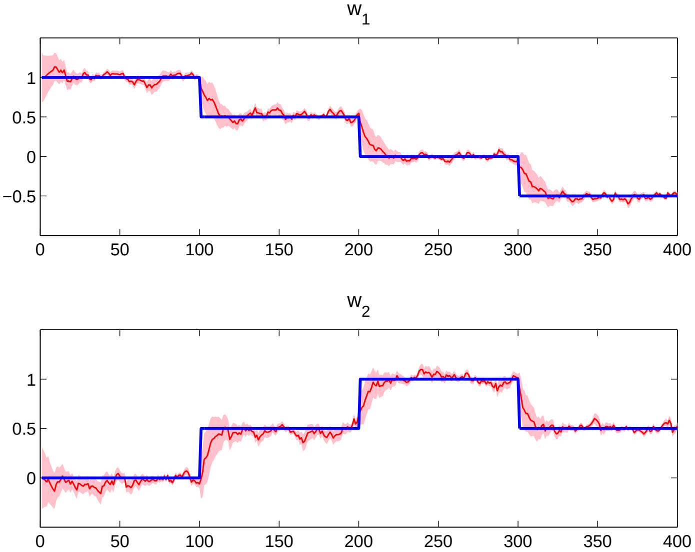
$$\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{W}_t,$$
$$z_t = \mathbf{x}_t^\top \mathbf{w}_t + v_t,$$
$$p(y_t = 1|\mathbf{x}_t) = \frac{1}{1 + e^{-z_t}}$$

- Time-varying parameter $\mathbf{w}_t$
- $\mathbf{w}_t$ treated as a hidden state variable
- Adaptively update the paramter $\mathbf{w}_t$
  - Every time a new observation y_t is given
  - Estimate the hidden state $\mathbf{w}_t$

Example



Example



w_1
1
0.5
0
−0.5
0
50
100
150
200
250
300
350
400
w_2
1
0.5
0
0
50
100
150
200
250
300
350
400

Outline

Brain-Computer Interface

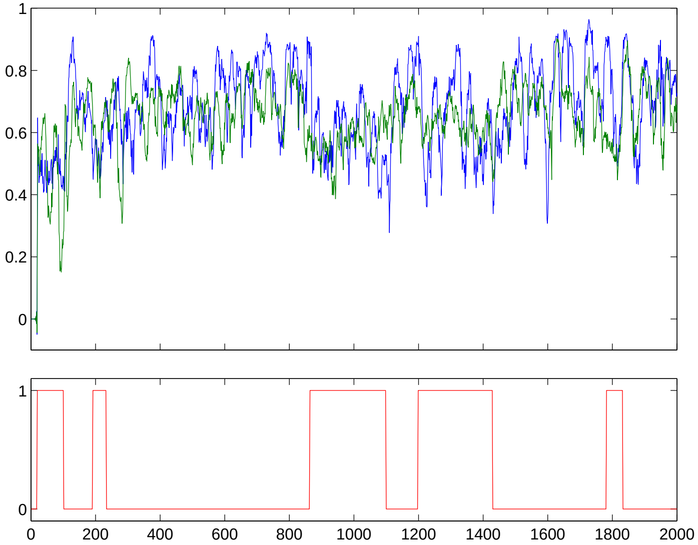
Interacting with computers using brain signals

- Useful for physically disabled people
- Manipulating computers without physical activities

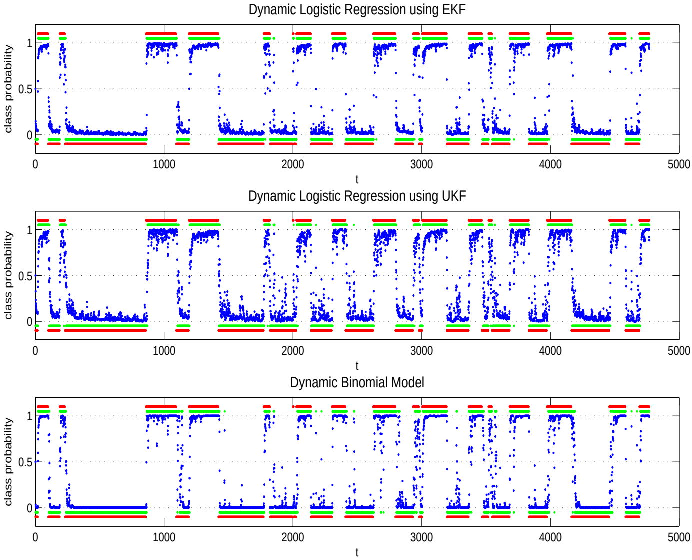
Analysing brain signals

- EEG signals
- Feature extraction
  - AR coeffients of moving windows
- Dynamic classification non-linear state space models
  - Extented/unscented Kalman filters
  - Particle filters

Imaginary Right Forearm Movement: EEG Signals



Imaginary Right Forearm Movement: Classification



Future Work

In the LONG run

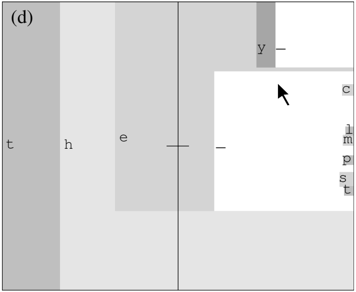
- A Dasher-like input system
- Dasher?
  - A text input system
  - By David J C MacKay (published in Nature, 2002)
  - Using 2-D eye-tracking
  - Auto-complete based on text analysis
  - 34 words/min (40-60 words/min with typical keyboards)

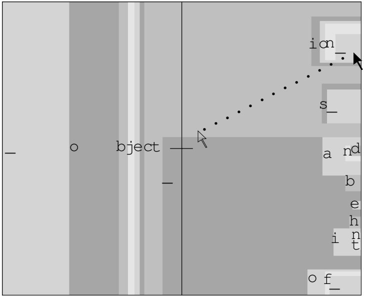
Dasher example

- Advantages of "BCI-Dasher"
  - Much faster
  - Possible without any physical movements

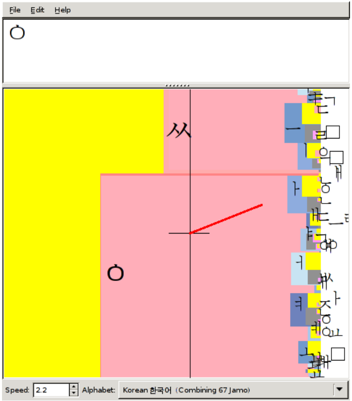
Next Skip

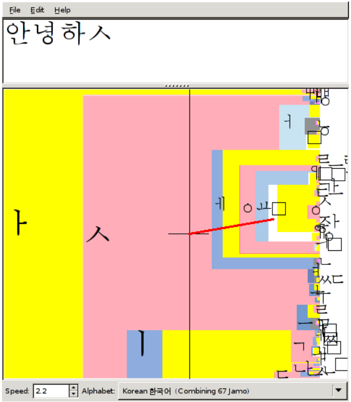
Dasher Example





Dasher Example in Korean





Back

Outline

Data Mining Preliminaries

Dynamic Prediction

Dynamic Classification

A Few Others...
- Monitoring Chickens' Behaviour
- Changepoint Detection in Pulsar Signals
- Animal Tracking

Conclusions

Outline

Welfare of Chickens

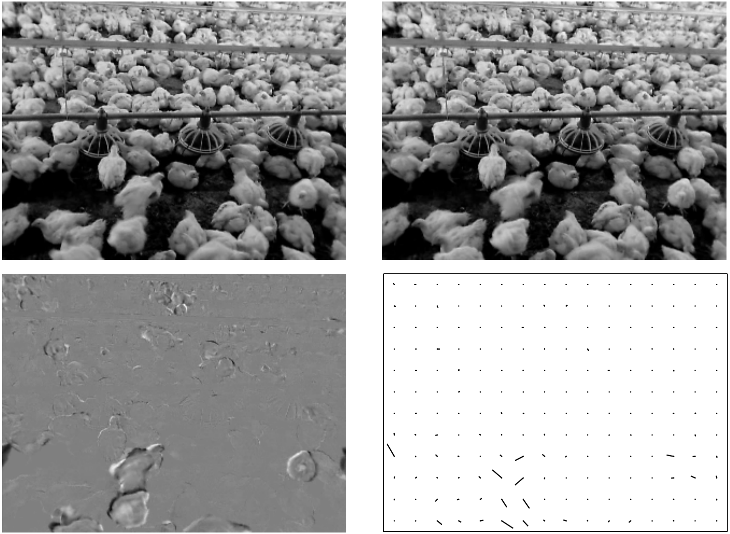
Very big business problems

- 40 billion chickens killed for meat each year
- Reach 2.5kg in less than 40 days

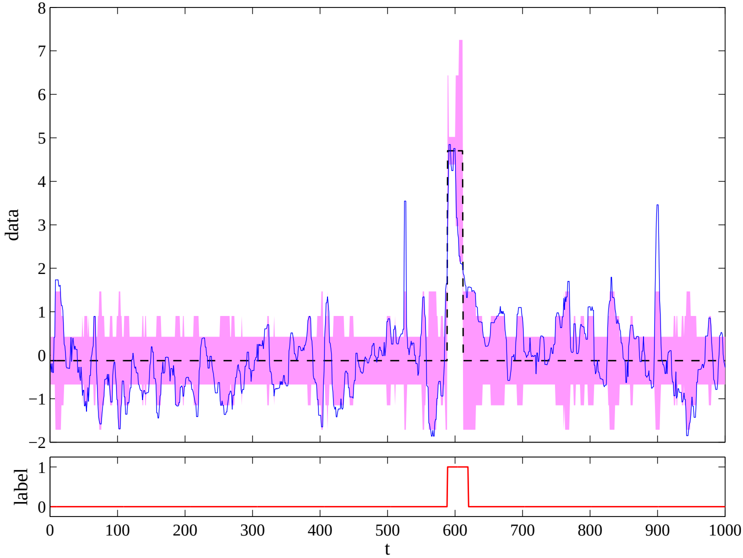
Objective

- To detect changes in chickens' behaviour
- From video footages of behaviour of chicken flocks
- Using hidden Markov models

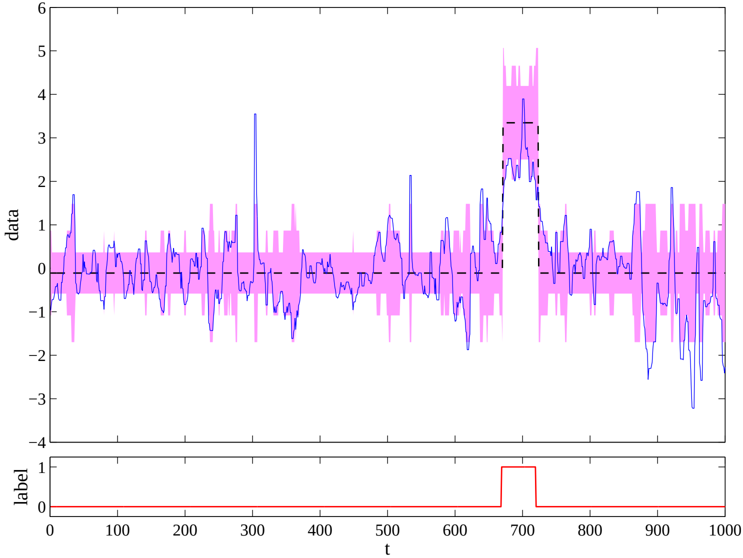
Optical Flow



State Identification in Chickens' Behaviour



State Identification in Chickens' Behaviour



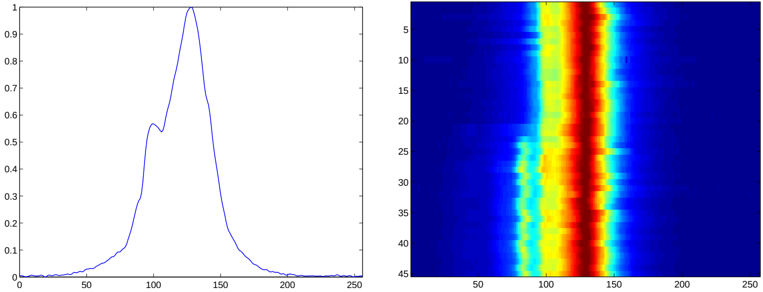
Outline

Pulsar Signals

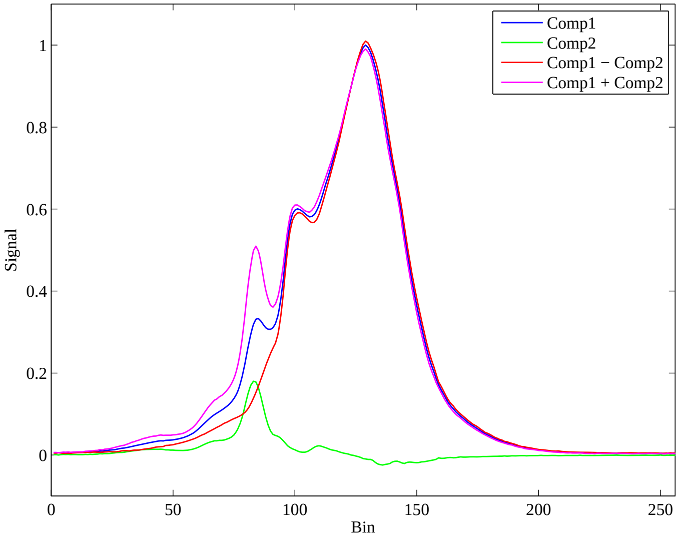
Spectra of some signals from a pulsar in the universe

- x: frequency
- y: signal intensity

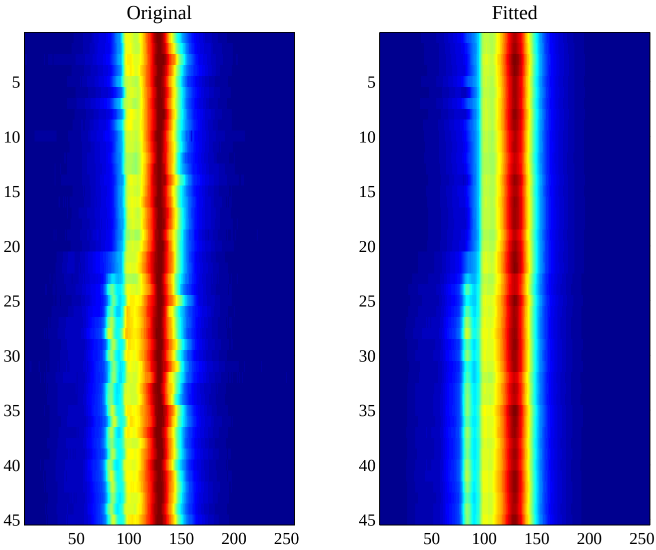
To show quantitatively that the signals are dynamic



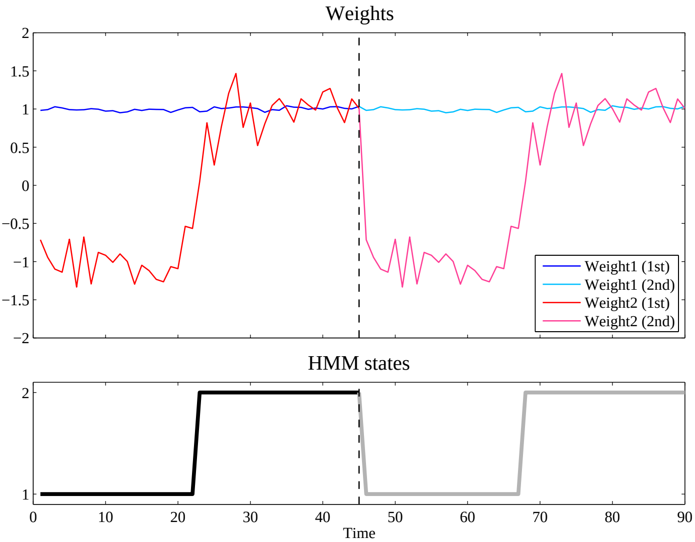
Principal Component Analysis



Fitting Results

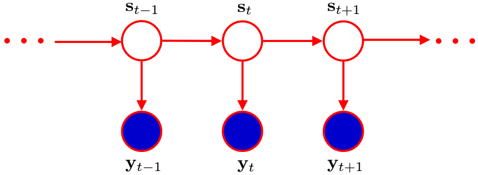


Changepoint Detection



Outline

Animal Tracking

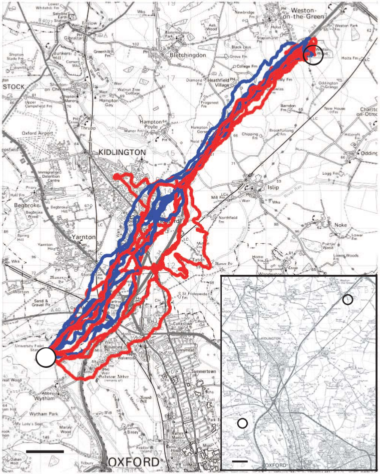


$$\mathbf{s}_t = f(\mathbf{s}_{t-1}) + \mathbf{W}_t$$
$$\mathbf{y}_t = g(\mathbf{s}_t) + \mathbf{V}_t$$

Animal tracking using state space models

- To estimate a path of an animal $\hat{\mathbf{s}}_{1:T}$
- Based on observations $\mathbf{y}_{1:T}$
  - GPS signals
  - Velocities, accelerations
  - Altitudes, temperatures, levels of sunlight

Pigeon tracking around Oxford



Outline

Data Mining Preliminaries

Dynamic Prediction

Dynamic Classification

A Few Others...

Conclusions

Conclusions

Machine learning in dynamic environments

- Many interesting (read: challenging) problems
  - Prediction
  - Classification
  - Changepoint detection
  - Tracking
- Uncertainty is important
- The Bayesian paradigm is useful

Techniques

- Gaussian processes
- State space models
  - Kalman filters (and variants thereof)
  - Hidden Markov models

A Few References

1. Pattern Analysis and Machine Learning Group Website.
   http://www.robots.ox.ac.uk/~parg/
2. David J C MacKay (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press ★FREE★
3. Christopher M Bishop (2006). Pattern Recognition and Machine Learning. Springer
4. Carl E Rasmussen and Chris K I Williams (2006). Gaussian Processes for Machine Learning. MIT Press ★FREE★
5. The Gaussian Processes Web Site.
   http://www.gaussianprocess.org/