Analysing Genomic Data

Mahesan Niranjan

Machine Learning Group
Department of Computer Science
The University of Sheffield

→ School of Electronics and Computer Science
  University of Southampton

Sheffield Winter School                                         January 2008

Service

Machine Learning

Molecular Biology

Data Driven Models

\[ \{ x_n, y_n \}_{n=1}^N \]

\[ y = f(x, \theta) + \nu \]

\[ x_{N+1} \rightarrow \hat{y}_{N+1} \]

\[ p(\theta) \]

New Algorithms

Classification, Regression
& online problems
Sampling Methods: Bayesian Inference

Parameters $\theta$

Uncertainty over parameters $p(\theta)$

Inference:

$$E[g(\theta)] = \int g(\theta) p(\theta) \, d\theta$$

$$= \frac{1}{N_s} \sum_{i=1}^{N_s} g(\theta^{(i)})$$

Importance Sampling

Density to sample from $p(\theta)$

Proposal Density $q(\theta)$

Generate $\theta^{(i)}$ from $q(\theta)$

Importance weights $w_i = \frac{p(g^{(i)})}{q(\theta^{(i)})}$

$$p(\theta) = \sum_{i=1}^{N_s} w_i \delta(\theta - \theta^{(i)})$$
Particle Filters

Bootstrap Filters (Gordon et al, Tracking)  
CONDENSATION Algorithm (Isard et al, Vision)

time: \( (n-1) \)  \[ \{ \theta^{(i)}(n-1), w_i(n-1) \}^N_{i=1} \]  
Prediction

Weights of Sample

time: \( (n) \)  \[ \{ \theta^{(i)}(n), w_i(n) \}^N_{i=1} \]

My research ambition for the next few years

From:

“I will email you some biological data  
- would you please cluster it for me?”  
(“I will make you co-author”)

To:

“According to my model, concentration  
of protein \( P \) will be \( X \) molecules per cell  
in the developing embryo at stage 5,  
– would you please measure it for me”  
(“I will make you co-author”)

Basic Biology: Cell

But structure is also important!
[Calladine et al.]
Basic Biology: *Gene*

- Purpose generally unknown
- Codes for Amino Acids
  - 3 letter NA → 1 AA

Basic Biology: *Protein*

- Primary Sequence
- Local Secondary Structure

- Primary protein structure
  - A helix
  - B helix
- Secondary protein structure
  - C helix
- Tertiary protein structure
  - D helix
- Quaternary protein structure
e.g. Sejnowski: Secondary structure prediction

Some Further Complexities

- Alternative Splicing
Some Further Complexities

• Regulatory Mechanisms
  – When to express a gene?
    transcription factors → bind to DNA → turn ON gene
  – How much protein to synthesise?
    (e.g. mRNA levels → Protein levels?)
  – Non-coding RNA; RNA Interference

Two Approaches to Pattern Recognition (Classification)

• Explicit modelling of probabilities encountered in Bayes’ formula

\[
P[\omega_i | x] = \frac{P[\omega_i] p(x | \omega_i)}{p(x)}
\]

• Parametric form for posterior probabilities / decision function and optimise it
Curse of dimensionality

Density estimation in high dimensions is difficult

Where CoD doesn’t hurt as much

Inference:

\[
E[g(\theta)] = \int g(\theta) p(\theta) \, d\theta
\]

\[
= \frac{1}{N_s} \sum_{i=1}^{N_s} g(\theta^{(i)})
\]

Convergence depends on number of samples, not dimensionality of the space
(...but practical problems of drawing samples)
Support Vector Machines

Classification, not density estimation

Support Vector Machines
Nonlinear Kernel Functions
Classifier Performance

• Error rates can be misleading
  
  – Imbalance in training/test data
    • 98% of population healthy
    • 2% population has disease
  
  – Cost of misclassification can change after design of classifier
Area under the ROC Curve: Neat Statistical Interpretation

\[ x \in A \quad y \in B \quad P \left[ g(x) > g(y) \right] \]

[David Hand: Details and Critical Review]

**Convex Hull of ROC Curves**

Provost & Fawcette
Scott, Niranjan & Prager
PARCEL: Feature subset selection

- Area under Convex Hull of multiple ROCs
- Different classifier architectures (including different features) in different operating points.

- Has been put to good use on independent implementations:
  - Oxford, UCL, Surrey
  - Sheffield Speech Group

Yeast Gene Classification:
[ Switch to MATLAB here ]

2000 yeast genes
79 experiments

Ribosome / Not
(125) (1750)

First use of SVM
Brown et al. PNAS 1999
Now for something different …

Basics: transcriptome and proteome

Gene sequences

Microarray → transcriptome

transcription

Unstable
Stochastic
Small numbers

mRNA

translation

Protein (sequence, structure, function)

Interesting action at protein levels

\[ \frac{\partial [P_i]}{\partial t} = f (P_1, P_2, \ldots, P_N) \]

2D Gel + spectrometry → proteome

mRNA often used as proxy for protein
Several authors have noted that there cannot be a one to one correspondence between mRNA and protein measurements.

Two views of the same dataset

Correlation between Protein and mRNA Abundance in Yeast
Steven P. Gygi, et. al. Molecular and Cellular Biology, March 1999

A Sampling of the Yeast Proteome
B. Futcher et. al. Molecular and Cellular Biology, November 1999

Poor Correlation < 0.7 😞

Reasonable Correlation (>0.69) 😊
Regulation

- Maintaining required protein concentration

Gene sequences → mRNA → Protein

Transcriptional

Post transcriptional (synthesis Vs decay)

Post translational (synthesis Vs decay)

Data-driven models / Machine Learning

\[
\{x_n, y_n\}_{n=1}^N
\]

\[y_n = f(x_n, \theta) + v_n\]

\[f(\quad)
\]

Radial Basis Functions
Artificial Neural Networks
Support vector machines

Good Data: \(\{x_n, y_n\}_{n=1}^{N_1}\)
where the model works

Bad Data: \(\{x_n, y_n\}_{n=N_1+1}^N\)
where the model does not work

通常是不允许发表，通常忘记。
Model “failure” can be informative I

Estimating pitch of voiced speech

- Short segment of speech
- Autoregressive model (vocal tract as an acoustic tube)
- Prediction residual by inverse filtering
- Enhanced periodicity (large errors when the model doesn’t fit)

Model “failure” can be informative II

Innovation probability in dynamical models

\[
\begin{align*}
\theta(n) & = f(\theta(n-1)) + v(n-1) \\
y(n) & = h(\theta(n)) + u(n)
\end{align*}
\]

- Prior
- Likelihood
- Posterior

\[
p(\theta|Y^n) = \frac{p(\theta|Y^{n-1}) p(y_n|\theta)}{p(y_n|Y^{n-1})}
\]

- Tracking and Switching
- Health monitoring

Innovation Probability

Novelty ➔ cue for action
Predictor for Protein Concentrations

\[ \mathbf{x} \rightarrow g(\mathbf{x}) \rightarrow y \]

\[ \{\mathbf{x}_n, y_n\}_{n=1}^{N} \]

Predictions from Gaussian process model
( also plain RBF & MLP )
- Gaussian kernel
- Hyperparameters from ML II
- 6 input features
- 76 data points

Different Sources of data, but same biological conditions
( rapidly growing yeast )

Improvement in prediction (small, but exists)
Search for “good” / “bad” data

\[ \{x_n, y_n\}_{n=1}^{N} \]

Backward deletion with

- average leave one out cross-validation error

(suboptimal) ranking of the genes/proteins

Genes/Proteins that do not fit the model are probably regulated by post translational processes

Some encouraging results?

Amongst the 76 mRNA/proteins, 4 known to be regulated post-translationally [Beilhartz, Brief. Func. Gen. Prot. 2004]:

- HAC1
- GCN4
- CPA1
- ICY2

These 4 are within the first 12 to be removed in backward deletion

Now:

- (Laborious process) of checking literature
- Larger dataset (230 mRNA protein pairs)
- Convince someone to check these experimentally
  [Repeat all measurements in same setting]
Now for something different …

Precision: 2.4601 or 2.0?

Why is precision 2.4601 unrealistic?
- mRNA from population of cells, not single cell
- mRNA inherently unstable molecule
- Other sources of stochasticity (e.g. in hybridization)

Why might this be of interest?
- Avoid using wrong tool (e.g. Gaussian noise model)
- Inference problems may be easier than thought
Classification

Performance of a support vector machine (SVM) in classifying ribosomal genes from all other yeast genes, using area under receiver operating characteristics curve as metric at various levels of quantization of the gene expression measurements.


Clustering

Will we find the same clusters at lower precision? We evaluate this by taking clusters of data from published work and comparing the average pair-wise correlations of genes within clusters with average correlations within a random set of genes at different levels of quantization.

Data: Iyer et al. (1999) Science 283: 83; cluster of 100 ('cluster A') against 417 others; human fibroblast response to serum.
Average correlations within and across cluster
Eisen et al. (1998) PNAS 95:14863; yeast; 79 hybridizations
(& ignoring genes with more than two missing values)

Area under ROC curve & Fisher ratio for one
dimensional Gaussians
Periodic expression – cell cycle

**Figure** showing raw and quantized expression levels and average pairwise correlations taken within group, random set of genes and across the two sets at various levels of quantization for 31 cell cycle generated genes.

Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis

Sean D Hooper1,2, Stephanie Boui1,2, Roland Krause2,3, Lars J Jensen4, Christopher E Mason4, Murad Ghanim5, Kevin P White6, Eileen Elliott Furlong7, and Peter Bork1,8
We may be able to work with lower precision...So what?

Distance measure suitable for binary data
Some related points

- Quantized microarray data
  - Friedman et al. *J Comput. Biol.*, 7:601, for constructing regulatory networks

- Quantize continuous data in other contexts
  - Dougherty et al. ICML 1995: Discretization improves classification performance (Naive Bayes & C4.5) → attribute the improvement to Gaussian assumption being weak.

Now for something different …