Machine Learning for Personalized Medicine

Jean-Philippe Vert

Atelier Prospectom, Grenoble, November 21, 2014
What’s in your body

1 body = $10^{14}$ human cells (and 100x more non-human cells)
1 cell = $6 \times 10^9$ ACGT coding for 20,000 genes
Many various data
A cancer cell
A cancer cell
A cancer cell
What is your risk of developing a cancer? (*prevention*)

After diagnosis and treatment, what is the risk of relapse? (*prognosis*)

What specific treatment will cure your cancer? (*personalized medicine*)
Example

Patients with same condition

DNA Profiling

Good responders

Bad side effects

No Responders
Machine learning formulation
Machine learning formulation
Machine learning formulation
Machine learning formulation
Challenges

- High dimension
- Few samples
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models
Learning with regularization

Learn

\[ f_{\beta}(x) = \beta^\top x \]

by solving

\[ \min_{\beta \in \mathbb{R}^p} R(f_{\beta}) + \lambda \Omega(\beta) \]

- \( R(f_{\beta}) \) empirical risk
- \( \Omega(\beta) \) penalty, typically:
  - \( \Omega(\beta) = \sum_{i=1}^{p} \beta_i^2 \) SVM, ridge regression, ...
  - \( \Omega(\beta) = \sum_{i=1}^{p} |\beta_i| \) Lasso, boosting, ...
1. Learning molecular classifiers with network information
2. Kernel bilinear regression for toxicogenomics
1 Learning molecular classifiers with network information

2 Kernel bilinear regression for toxicogenomics
Joint work with...

Franck Rapaport, Emmanuel Barillot, Andrei Zinovyev, Anne-Claire Haury, Laurent Jacob, Guillaume Obozinski
Breast cancer prognosis

(van 't Veer et al., 2002)
Gene selection, molecular signature

The idea

- We look for a **limited set** of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology.
Some "surprising" results

**Gene expression profiling predicts clinical outcome of breast cancer**

Laura J. van 't Veer†, Hongyue Dai‡, Marc J. van de Vijver†,
Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse†,
Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen†,
George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡,
Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

70 genes (Nature, 2002)

**Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer**

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans,
Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Bems, David Atkins, John A Foekens

76 genes (Lancet, 2005)

3 genes in common
Lack of stability of signatures

(Haury et al., 2011)
Gene networks

- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Sulfur metabolism
- Porphyrin and chlorophyll metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Purine metabolism
- Oxidative phosphorylation, TCA cycle
- Nitrogen, asparagine metabolism
- DNA and RNA polymerase subunits
Gene networks and expression data

Motivation

- Basic biological functions usually involve the **coordinated action of several proteins**:
  - Formation of **protein complexes**
  - Activation of metabolic, signalling or regulatory **pathways**
- Many pathways and protein-protein interactions are **already known**
- **Hypothesis**: the weights of the classifier should be “coherent” with respect to this **prior knowledge**
Graph based penalty

\[ f_\beta(x) = \beta^T x \]

\[ \min_{\beta} R(f_\beta) + \lambda \Omega(\beta) \]

Prior hypothesis

Genes near each other on the graph should have similar weights.

An idea (Rapaport et al., 2007)

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]

\[ \min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2. \]
Graph based penalty

\[ f_\beta(x) = \beta^T x \]
\[ \min_{\beta} R(f_\beta) + \lambda \Omega(\beta) \]

Prior hypothesis

Genes near each other on the graph should have similar weights.

An idea (Rapaport et al., 2007)

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2, \]
\[ \min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2. \]
Spectral analysis of gene expression profiles using gene networks

Fig. 5. This figure illustrates the workflow of gene expression analysis using a gene network approach. The work was supported by the grant ACI-IMPBIO-2004-47 of the French Ministry for Research and New Technologies.
Spectral penalty as a kernel

**Theorem (Rapaport et al., 2007)**

The function $f(x) = \beta^T x$ where $\beta$ is solution of

$$
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell \left( \beta^T x_i, y_i \right) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2
$$

is equal to $g(x) = \gamma^T \Phi(x)$ where $\gamma$ is solution of

$$
\min_{\gamma \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell \left( \gamma^T \Phi(x_i), y_i \right) + \lambda \gamma^T \gamma,
$$

and where

$$
\Phi(x)^T \Phi(x') = x^T K_G x'
$$

for $K_G = L^*$, the pseudo-inverse of the graph Laplacian.
The Laplacian of the graph is the matrix $L = D - A$. 

$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$
Pseufo-inverse of the Laplacian

\[
L^* = \begin{pmatrix}
0.88 & -0.12 & 0.08 & -0.32 & -0.52 \\
-0.12 & 0.88 & 0.08 & -0.32 & -0.52 \\
0.08 & 0.08 & 0.28 & -0.12 & -0.32 \\
-0.32 & -0.32 & -0.12 & 0.48 & 0.28 \\
-0.52 & -0.52 & -0.32 & 0.28 & 1.08
\end{pmatrix}
\]
Other penalties with kernels

\[ \Phi(x)^\top \Phi(x') = x^\top K_G x' \]

with:

- \( K_G = (c + L)^{-1} \) leads to

\[ \Omega(\beta) = c \sum_{i=1}^{p} \beta_i^2 + \sum_{i \sim j} (\beta_i - \beta_j)^2. \]

- The diffusion kernel:

\[ K_G = \exp_M(-2tL). \]

penalizes high frequencies of \( \beta \) in the Fourier domain.
Other penalties without kernels

- Gene selection + Piecewise constant on the graph

\[ \Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i| \]

- Gene selection + smooth on the graph

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i| \]
Example: classification of DNA copy number profiles

Aggressive (left) vs non-aggressive (right) melanoma
Fused lasso solution (Rapaport et al., 2008)

\[ \Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i| \]
Graph-based structured feature selection

\[ \Omega_1(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}, \quad \text{(Jenatton et al., 2011)} \]

\[ \Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta. \quad \text{(Jacob et al., 2009)} \]
Breast cancer prognosis
Graph Lasso signature (accuracy 0.64)

Breast cancer prognosis
Disjoint feature selection

Motivation: multiclass or multitask classification problems where we want to select features specific to each class or task.

Example: recognize identify and emotion of a person from an image (Romera-Paredes et al., 2012), or hierarchical coarse-to-fine classifier (Xiao et al., 2011; Hwang et al., 2011)
Disjoint feature selection

\[ W = (w_i)_{i \in V} \in \mathbb{R}^{p \times V} \quad \Omega(W) = \min_{-H \leq W \leq H} \sum_{i \sim j} K_{ij} \left| h_i^T h_j \right| \]

(Vervier et al., 2014)
Example: multiclass classification of MS spectra

(Vervier et al, 2014)
Outline

1. Learning molecular classifiers with network information
2. Kernel bilinear regression for toxicogenomics
Joint work with...

Elsa Bernard, Erwan Scornet, Yunlong Jiao, Véronique Stoven, Thomas Walter
Pharmacogenomics / Toxicogenomics

Patients with same condition

DNA Profiling

Good responders

No Responders

Bad side effects
Genotypes from the 1000 genome project
RNASeq from the Geuvadis project
Cell line $X$, chemical $Y$, toxicity $Z$.

**Bilinear regression model:**

$$Z = f(X, Y) + b(Y) + \epsilon,$$

**Estimation by kernel ridge regression:**

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}^p} \sum_{i=1}^{n} \sum_{j=1}^{p} (f(x_i, y_j) + b_j - z_{ij})^2 + \lambda \|f\|^2,$$
Theorem 1. Let $Z \in \mathbb{R}^{n \times p}$ be the response matrix, and $K_X \in \mathbb{R}^{n \times n}$ and $K_Y \in \mathbb{R}^{p \times p}$ be the kernel Gram matrices of the $n$ cell lines and $p$ chemicals, with respective eigenvalue decompositions $K_X = U_X D_X U_X^T$ and $K_Y = U_Y D_Y U_Y^T$. Let $\gamma = U_X^T 1_n$ and $S \in \mathbb{R}^{n \times p}$ be defined by $S_{ij} = 1/ \left( \lambda + D_X^i D_Y^j \right)$, where $D_X^i$ (resp. $D_Y^j$) denotes the $i$-th diagonal term of $D_X$ (resp. $D_Y$). Then the solution $(f^*, b^*)$ of (2) is given by

$$b^* = U_Y \text{Diag} \left( S^T \gamma \circ 2 \right)^{-1} \left( S^T \circ \left( U_Y^T Z U_X \right) \right) \gamma$$

(3)

and

$$\forall (x, y) \in X \times Y, \quad f^*(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_{i,j}^* K_X(x_i, x) K_Y(y_i, y),$$

(4)

where

$$\alpha^* = U_X \left( S \circ \left( U_X^T \left( Z - 1_n b^* U_Y^T \right) U_Y \right) \right) U_Y^T.$$
Kernel Trick

cell line descriptors

drug descriptors

mercredi 6 novembre 13
Kernel Trick

- **cell line descriptors**
- **drug descriptors**

Kernelized

- $K_{cell}$
- $K_{drug}$

jeudi 7 novembre 13
Kernel Trick

cell line descriptors

Kcell

kernelized

Kdrug

kernel bilinear regression

\hat{f}

drug descriptors
Kernel Trick

- cell line descriptors
- drug descriptors
- Kcell: kernelized
- Kdrug: kernel bilinear regression
- Kernel choice?
- descriptors
- data integration
- missing data
- jeudi 7 novembre 13
1. $K_{\text{cell}}$:
   - 29 cell line kernels tested
   - 1 kernel that *integrate all information*
   - deal with missing data

2. $K_{\text{drug}}$:
   - 48 drug kernels tested
   - *multi-task* kernels
Kernel choice

1. $K_{\text{cell}}$:
   - $\Rightarrow$ 29 cell line kernels tested
   - $\Rightarrow$ 1 kernel that *integrate all information*
   - $\Rightarrow$ deal with missing data

2. $K_{\text{drug}}$:
   - $\Rightarrow$ 48 drug kernels tested
   - $\Rightarrow$ multi-task kernels
Cell line data integration

**Covariates**
- linear kernel

**SNPs**
- 10 gaussian kernels

**RNA-seq**
- 10 gaussian kernels
Cell line data integration

Covariates
- linear kernel

SNPs
- 10 gaussian kernels

RNA-seq
- 10 gaussian kernels

Integrated kernel
Dirac

Multi-Task

Feature-based

Empirical

Integrated

independent regression for each drug
Multi-task drug kernels

1. Dirac
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated

sharing information across drugs
Multi-task drug kernels

Dirac

Multi-Task

Feature-based

Empirical

Integrated

Linear kernel and 10 gaussian kernels based on features:

- CDK (160 descriptors) and SIRMS (9272 descriptors)
- Graph kernel for molecules (2D walk kernel)
- Fingerprint of 2D substructures (881 descriptors)
- Ability to bind human proteins (1554 descriptors)
Multi-task drug kernels

1. Dirac
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated

Empirical correlation

Color Key and Histogram
Multi-task drug kernels

- Dirac
- Multi-Task
- Feature-based
- Empirical
- Integrated

Integrated kernel:

\[ K_{int} = \sum_i K_i \]

- Combine all information on drugs
29x48 kernel combinations: CV results

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Value

0 150

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53

Color Key and Histogram

Value

Count

0.5 0.51 0.52 0.53
29x48 kernel combinations: CV results

Color Key and Histogram

Value

Count

0 150

CI

integrated
and
covariates
kernels
29 x 48 kernel combinations: CV results

Color Key and Histogram

Count 0 - 150
Value 0.5, 0.51, 0.52, 0.53

CI

sightly multi-task
on drugs

covariates kernel
on cell lines

jeudi 7 novembre 13
Kernel on cell lines: CV results

Mean CI for cell line kernels

Integrated kernel

Batch effect

jeudi 7 novembre 13
Kernel on drugs: CV results

Mean CI for chemicals kernels

Many new problems and lots of data in computational genomics

Computational constraints $\Rightarrow$ fast sparse models (FlipFlop)

Small $n$ large $p$ $\Rightarrow$ regularized models with prior knowledge

Heterogeneous data integration $\Rightarrow$ kernel methods

Personalized medicine promising but difficult!
Thanks


