Machine Learning for Scientific Discovery

Cheng Soon Ong

Machine Learning Research Group
NICTA Canberra

4 December 2014
2014 Australian Frontiers of Science

www.ong-home.my/download/frontiers2014.pdf

Machine Learning and Physics

Higgs challenge
the HiggsML challenge
May to September 2014
When High Energy Physics meets Machine Learning

What is machine learning?

Machine learning is about prediction

Examples/features	$x_1, \ldots, x_n \sim \mathcal{X}$
Labels/annotations	$y_1, \ldots, y_n \sim \mathcal{Y}$
Predictor	$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Estimate best predictor = training

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

- No mechanistic model of the phenomenon
- There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
- The outcomes (labels, y usually binary) are well defined

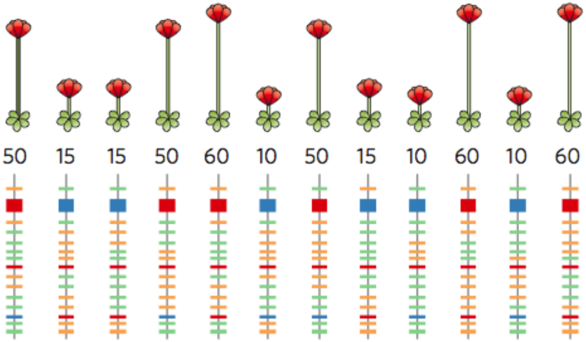
Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

What are good features?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

What are good biomarkers?



Genome Wide Association Studies

- Which mutations are associated with tall poppies?
- Identify biomarkers with hypothesis tests

Finding stable biomarkers

- Split cohort into two (cross validation)
- Use p-value as a score
- Investigate rank correlation between scores

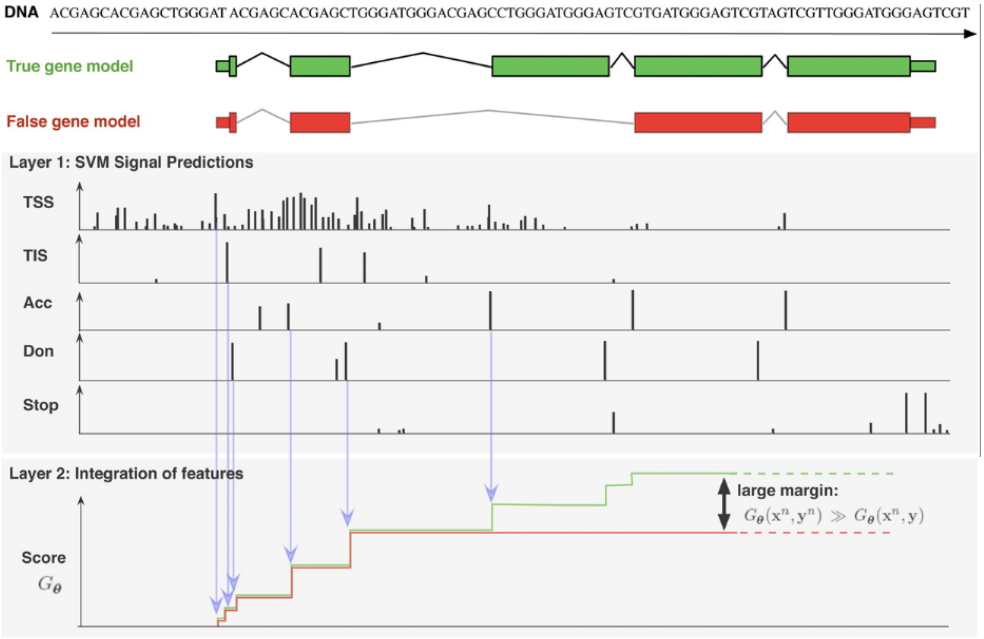
bioinformatics.research.nicta.com.au/software/gwis/

Not standard binary classifcation

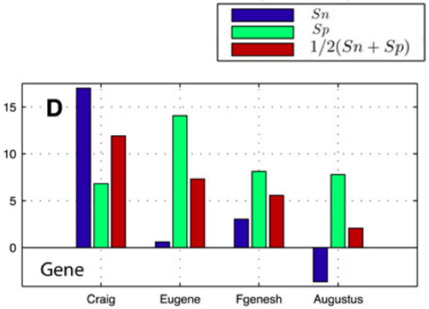
$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Gene finding

Predict a sequence of binary decisions


DNA ACGAGCACGAGCTGGGAT ACGAGCACGAGCTGGGATGGGACGAGCCTGGGATGGGAGTCGTGATGGGAGTCGTAGTCGTTGGGATGGGAGTCGT
True gene model
False gene model
Layer 1: SVM Signal Predictions
TSS
TIS
Acc
Don
Stop
Layer 2: Integration of features
Score
G_θ
large margin:
$G_\theta(x^n, y^n) \gg G_\theta(x^n, y)$

www.mgene.org/web

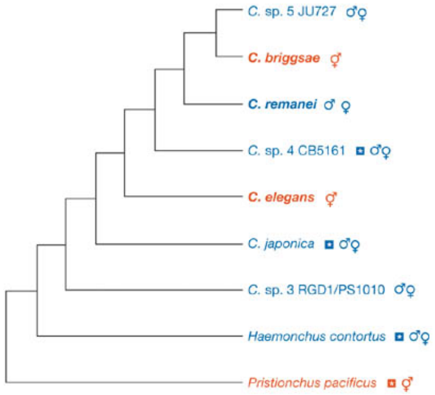


Improving annotation

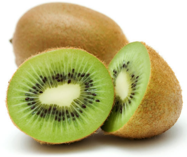
Improving well studied genomes

	Total	Tested	Confirmed	Fraction
New genes	2197	57	24	42%
Missed unconf. genes	205	24	2	8%

Annotating new genomes



Unknown objects

During training				
After deployment				

Identifying wheel defects in trains

- Wheel defects destroy infrastructure
- Classify type of defect from time series

Collaboration with Swiss National Railway

Classifying celestial objects

- Skymapper southern sky survey
- Rare objects not available at training

Discussion with Christian Wolf, RSAA, ANU

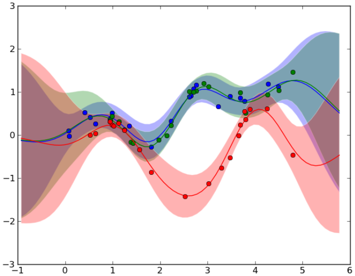
What to measure?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Active Learning / Expt. Design

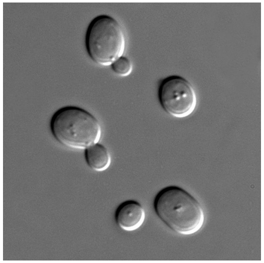
Use predictor to identify good candidates

- Annotate top-k items
- Confidence interval improves performance
- Explore - exploit tradeoff



Glucose metabolism in Yeast

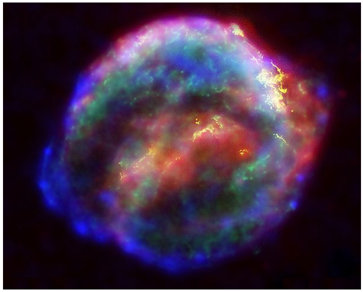
- Multiple possible models
- Design biological experiments that maximise information gain



Collaboration with SystemsX Switzerland

Finding supernovae

- Machine learning to classify images
- Show 10 candidates to expert daily

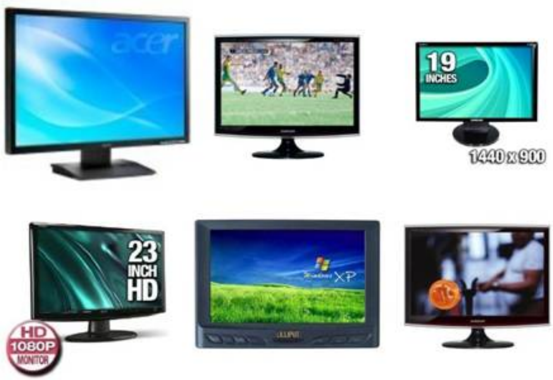


Discussion with Richard Scalzo at RSAA, ANU

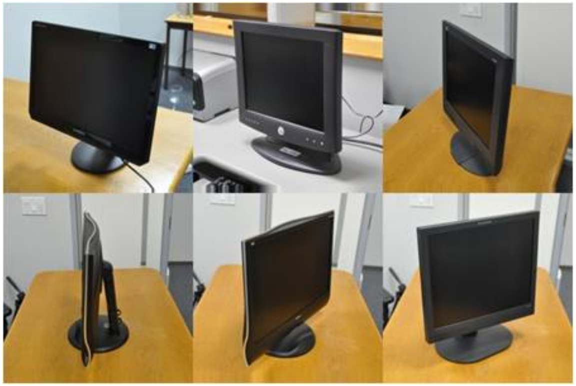
Challenges to ML4Science

What is the keyword? (1)

Training



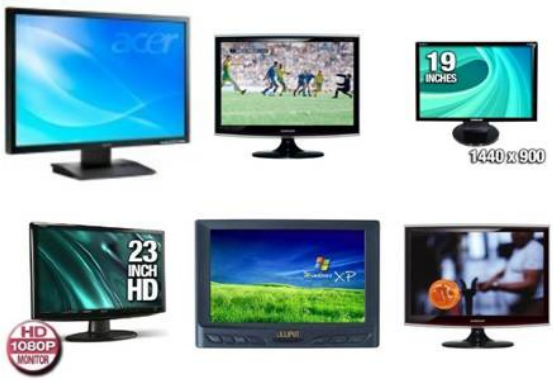
Deployed



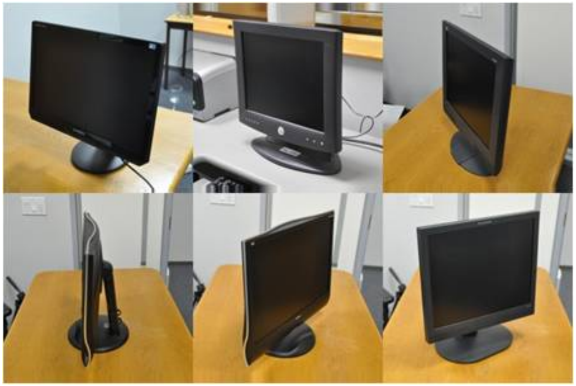
www.cs.uml.edu/∼saenko/projects.html#data

What is the keyword? (1)

Training



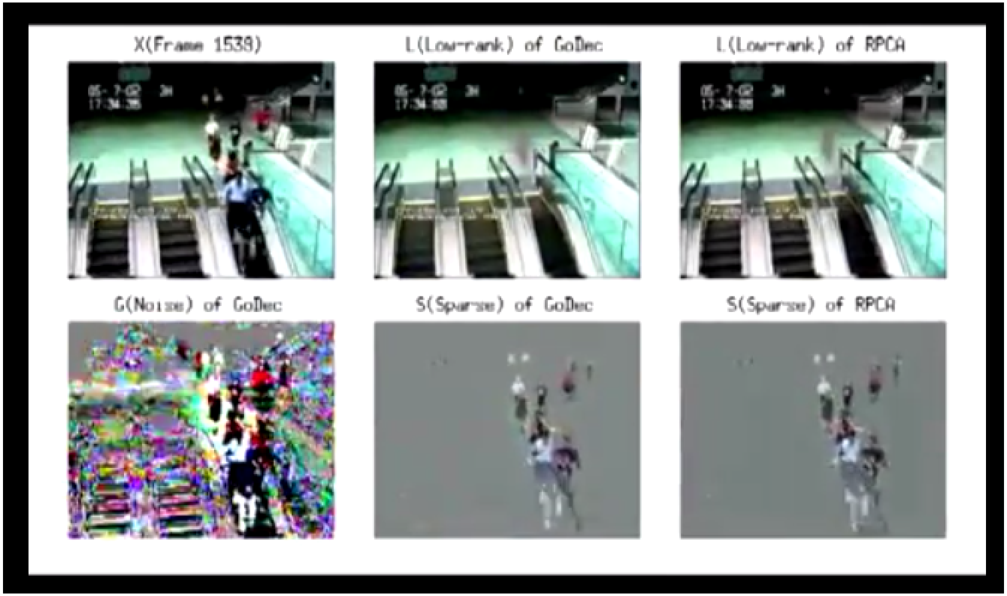
Deployed



www.cs.uml.edu/∼saenko/projects.html#data

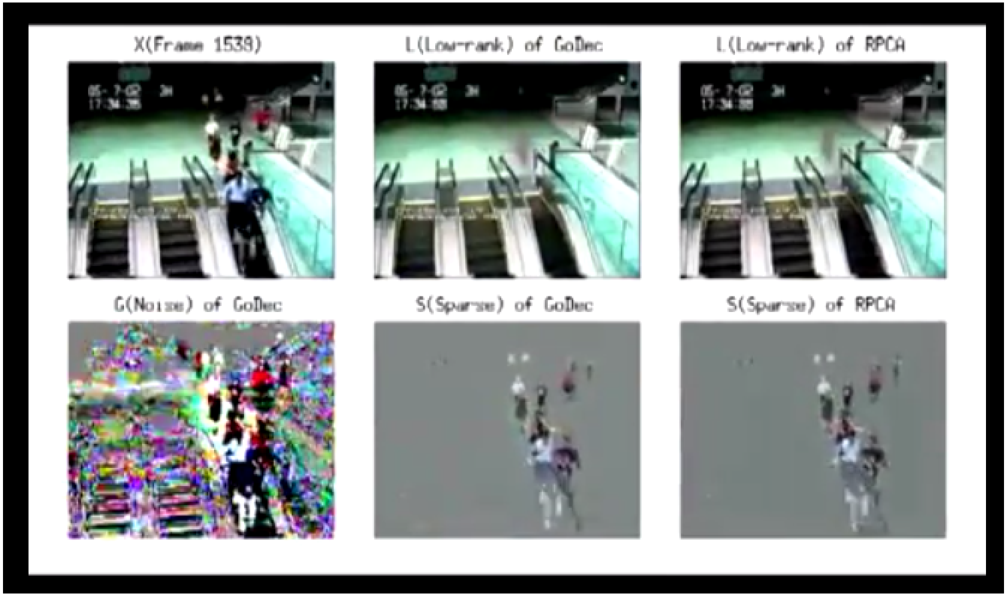
Domain adaptation

What is the keyword? (2)



https://www.youtube.com/watch?v=YpdCvbJI2eg

What is the keyword? (2)



sites.google.com/site/godecomposition/home

Robust principal component analysis

ML Open Source Software

Wider adoption of methods

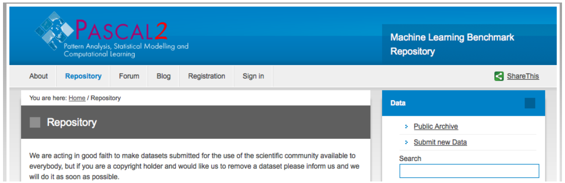
- Domain experts can use machine learning core
- Available for teaching

Scientific reproducibility

- Fair comparison of methods
- Access to scientific tools

Community growth

- “Given enough eyeballs, all bugs are shallow”
- Combination of advances


PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning
Machine Learning Benchmark Repository
About
Repository
Forum
Blog
Registration
Sign in
ShareThis
Data
Public Archive
Submit new Data
Search

You are here: Home / Repository

Repository

We are acting in good faith to make datasets submitted for the use of the scientific community available to everybody, but if you are a copyright holder and would like us to remove a dataset please inform us and we will do it as soon as possible.

LIBSVM 2.9

by cjlin - February 27, 2010, 01:09:23 CET [ ] 4935 views, 1079 downloads,

About: LIBSVM is an integrated software for support vector classification, (C-SVC, nu-estimation (one-class SVM). It supports multi-class [...]

mloss.org   mldata.org

Plug and Pray

Machine Learning Open Source Software

Do We Need Hundreds of Classifiers
to Solve Real World Classification Problems?

jmlr.org/papers/v15/delgado14a.html

Spoiler: No

Usability and Reproducibility

- (too much) focus on new algorithms
- Documentation, modularity issues
- Literate programming

  ipython.org/notebook.html yihui.name/knitr jupyter.org
- Scientific computing workflows

  galaxyproject.org

Dream: App Bazaar for data science

Summary

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Three extensions

- What are good features? $f_{\mathbf{w}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$
- Not standard binary classification $f_{\mathbf{w}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$
- What to measure? $f_{\mathbf{w}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$

Plug and pray

- Finding the right keyword
- Software, software, software

Thank You

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Three extensions

- What are good features? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- Not standard binary classification $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$
- What to measure? $f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

Plug and pray

- Finding the right keyword
- Software, software, software

Please make your research open

www.nicta.com.au/research/machine_learning

www.ong-home.my

References

Open Science

- Sören Sonnenburg, Mikio L. Braun, Cheng Soon Ong, et. al. The need for open source software in machine learning. Journal of Machine Learning Research, 8:2443âĂŞ2466, 2007.
- Joaquin Vanschoren, Mikio Braun, Cheng Soon Ong, Open Science in Machine Learning, Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society (CLADAG) 2013, Modena, Italy
- Mikio L. Braun, Cheng Soon Ong, Open Science in Machine Learning. Book chapter in Implementing Reproducible Research, 2014

Stability of feature selection

Justin Bedő, David Rawlinson, Benjamin Goudey, Cheng Soon Ong, Stability of bivariate GWAS biomarker detection PLoS ONE, 9(4), e93319

Gene finding

- Gabriele Schweikert, Jonas Behr, Alexander Zien, Georg Zeller, Cheng Soon Ong, SÃűren Sonnenburg and Gunnar RÃď'tsch. mGene.web: a web service for accurate computational gene finding. Nucleic Acids Research, Volume 37, Web Server Issue, 2009.
- Gabriele Schweikert, et. al. mGene: Accurate SVM-based gene finding with an application to nematode genomes. Genome Research, 19:2133–2143, 2009.

Confidence sets

Fan Shi, Cheng Soon Ong, Christopher Leckie. Applications of Class-Conditional Conformal Predictor in Multi-Class Classification International Conference on Machine Learning and Applications, 2013

Active Learning

- Alberto Giovanni Busetto, Cheng Soon Ong and Joachim M. Buhmann. Optimized Expected Information Gain for Nonlinear Dynamical Systems. In Proceedings of the International Conference on Machine Learning, pages 97–104, 2009.
- Alberto Giovanni Busetto, et. al. Near-optimal experimental design for model selection in systems biology Bioinformatics, 29 (20): 2625-2632. doi:10.1093/bioinformatics/btt436
- Andreas Krause, Cheng Soon Ong. Contextual Gaussian Process Bandit Optimization. Advances in Neural Information Processing, 2011.