Machine learning for biomedical problems, including radiomics and (radio)genomics

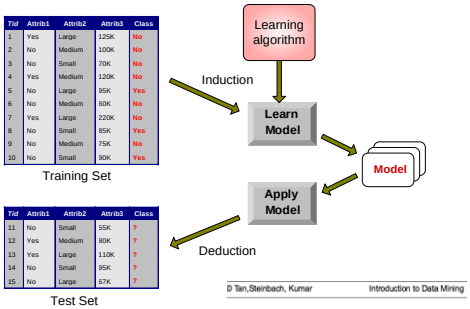
Gaurav Pandey
Department of Genetics and Genomic Sciences
Icahn Institute for Genomics and Multiscale Biology
Icahn School of Medicine at Mount Sinai, New Yo
http://research.mssm.edu/gpandey/
gaurav.pandey@mssm.edu

Mount Sinai

What is Machine Learning?

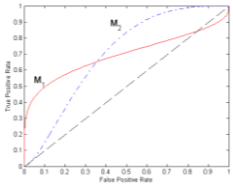
- ML is the science (art?) of discovering **actionable** models/patterns/knowledge directly from data.
- ML methods try to:
  - Make as few assumptions and be as computationally efficient as possible (vis-à-vis traditional statistical methods)
  - Be as unbiased w.r.t. current knowledge as possible (vis-à-vis traditional bioinformatics and computational biology methods)
- Several types of (machine) learning:
  - Supervised: Classification, regression.
  - Unsupervised: Clustering, anomaly detection.
  - Others: Semi-supervised, ensemble, deep, feature selection, spatio-temporal.

Supervised learning



Evaluation of supervised ML models

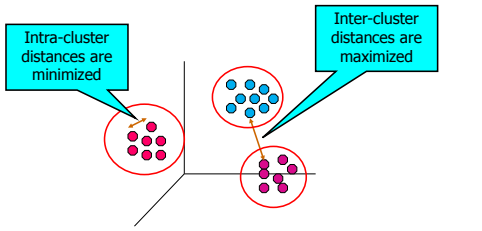
- Evaluation setups:
  - Training-test sets
  - Cross-validation
- Evaluation metrics:
  - Accuracy
  - ROC Curve
    - Shows relationship between True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity) across a variety of thresholds applied to classifier output scores.
    - Area Under the Curve (AUC)
      - Ideal: AUC = 1
      - Random: AUC = 0.5
    - Model with higher AUC generally considered better



- More specialized metrics needed for unbalanced data sets
  - Typical biomedical problems (e.g. healthy-vs-diseased) are unbalanced
  - Right metrics to use: Precision-Recall-Fmeasure
- Excellent review: Lever et al, Nature Methods, 2016.

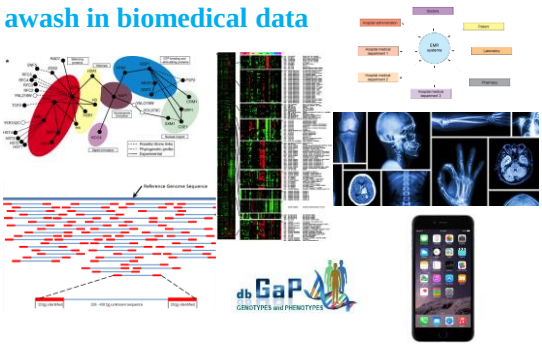
Unsupervised learning: Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



© Tan,Steinbach, Kumar Introduction to Data Mining

Why should we care? Because we are awash in biomedical data



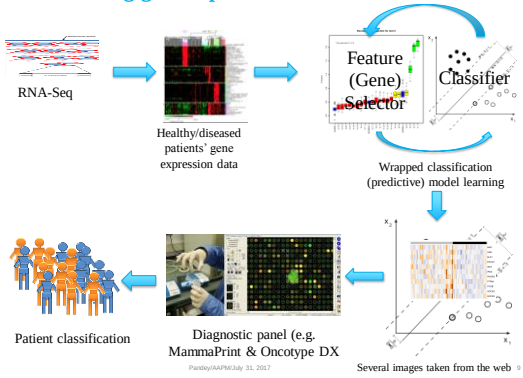
Several images taken from the web

Abundant data and ML provide opportunities to address problems related to personalized/precision medicine

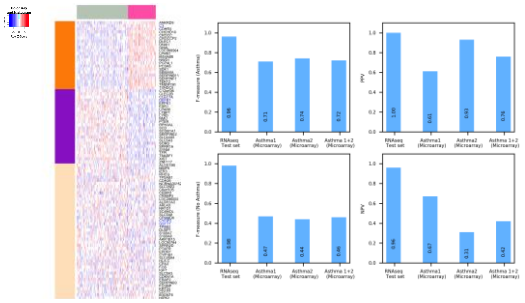
- Supervised learning
  - Discovery of factors affecting/related to health/disease (biomarkers)
    - Genetic/genomic factors
    - Environmental factors (exposome)
    - Gene X Environment interactions
  - Prediction of disease phenotypes, progression, survival rates etc.
    - Imaging data (radiomics, deep learning etc.)
    - Genetic, EMR and other data types
  - Prediction of drug sensitivity/effectiveness and adverse/side effects
- Unsupervised learning
  - Disease subtype discovery
  - Deconvolution of cell types in a mixture
  - Drug repositioning and discovery of effective drug combinations
  - Clustering of (disease-related) gene/proteins into functions/pathways etc.

Some applications of ML in biomedical problems

Discovering gene expression biomarkers of diseases

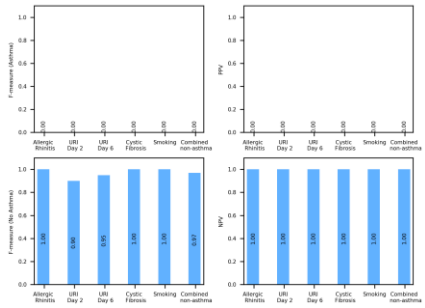


Accurate diagnostic panel for (mild/moderate) asthma



Pandey et al, bioRxiv 145771, 2017

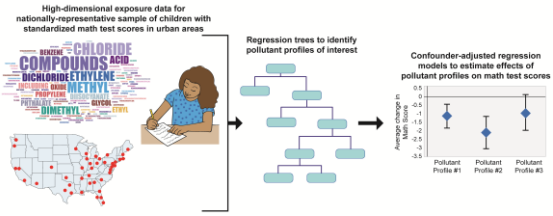
Specificity of diagnostic panel to (mild/moderate) asthma vis-à-vis respiratory diseases with similar symptoms



Pandey et al, bioRxiv 145771, 2017

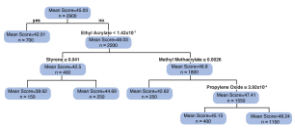
Discovering air pollutant combinations affecting children's health

"You Can't Change Your Genes, but You Can Change the Environment: How the Environment Affects Your Health": Dr. Linda S. Birnbaum, Director, National Institute of Environmental Health Sciences and National Toxicology Program



Stingone et al, Environmental Pollution, 2017

Pollutant combinations can help define at-risk population profiles



Selected demographic characteristics[a] of children within the Early Childhood Longitudinal Study, Birth Cohort living in highly-populated urban ZIP codes, by ambient bisphenol A level.

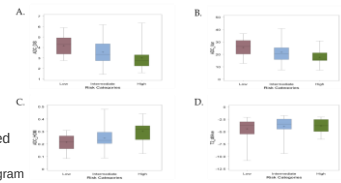
	ZIP Code's estimated ambient concentration of bisphenol A	
Geographic Region	≥0.47 ng/m^3	<0.47 ng/m^3
Northeast	36.3	20.0
Southern/Southeastern	28.1	33.5
Central	23.2	27.2
Western	12.3	19.3
Community Demographics Average Proportion of Residents with Less than a HS Education	22.0	19.7
Average Proportion of Residents who are unemployed	3.5	4.4
Average Proportion of Residents who are living in poverty	8.8	11.8
Average Proportion of Residents who are Black, non-Hispanic	19.2	7.8
Average Proportion of Residents who Rent their Homes	42.7	25.9
Average Proportion of Male Residents who have Professional Occupations	16.8	14.1
Average Proportion of Female Residents who have Professional Occupations	23.5	23.0
Average Proportion of Residents who lived in the same community for the previous five years	54.6	58.4

a Demographic data from the 2000 US Census.

Stingone et al, Environmental Pollution, 2017

Radiomics and ML for tumor classification

- Data set: 68 prostate tumor captured using mpMRI (ADC and T2)
  - 54 low and intermediate
  - 14 highly aggressive tumors
- 116 radiomics features derived from images:
  - Mostly texture-based (Histogram analysis, GLCM, GLDM and Fourier analysis)



Box plots of GLCM Dissimilarity (A), Variance (B) and Homogeneity (C) on ADC and Difference Average (D) on T2 for each risk category

Frank Chen, Bino Varghese, Darryl Hwang, Steve Cen and Vinay Duddalwar, USC

- Goal: Can supervised ML methods be applied to this data set to improve tumor classification by identifying a combination of radiomics features?

ML methodology and current results

Bino Varghese and Vinay Duddalwar (USC)

Challenges with biomedical ML

- Type and amount of data being analyzed should be relevant and representative for the target problem.
- Interpretability of ML models: "Block Box" characterization
  - Much of this comes from incomplete understanding of how ML methods work
- Data issues:
  - Noise
  - Missing data
  - Incompatibility of data from different sources
    - Same data type: Different scales/distributions (batch effects, normalization etc.)
    - Different data types: Different representations, not always clear how to integrate
  - Integral to any data analysis, not just ML
    - Best practices should be followed, unless better solutions available

Summary

- ML methods of several types hold great potential in the data-rich era of biomedical sciences to address challenging problems and derive actionable knowledge directly from data.
- Several useful applications:
  - Development of diagnostic gene expression panels for diseases (e.g. asthma)
  - Identification of air pollutant combinations that affect children's health
  - Tumor classification based on radiomics data derived from mpMRI images
- Substantial challenges remain and efforts are being made!

Acknowledgements

- To you for your attention and meeting and session organizers for the invitation ☺
- ML method development: Sean Whalen, Ana Stanescu, Om P. Pandey and other lab members at Mount Sinai
- Asthma: Supinda Bunyavanich (Mount Sinai)
- Air pollution and environmental health: Jeanette Stingone and Luz Claudio (Mount Sinai)
- Radiomics: Bino Varghese and Vinay Duddalwar (USC)
- Financial and technical support: NIH R01GM114434, P30ES023515, IBM Faculty Award and Mount Sinai Institute for Genomics & Multiscale Biology and Minerva supercomputing team.