Model-free machine learning methods for personalized breast cancer risk prediction -SWISS PROMPT

Chang Ming, 22.11.2017
University of Basel

Swiss Public Health Conference 2017

Breast Cancer & personalized cancer prevention

- In Switzerland each year, about 5’250 women develop breast cancer and 1’350 die from it.[1]
- Prediction model in personalized cancer prevention:
  - ability to forecast breast cancer risk or presence before clinical symptoms appear
  - opportunity to act on the breast cancer through early intervention.
  - guide surveillance and preventive treatment (such as increased frequency of mammography, prophylactic surgery, chemoprevention and medication)

What is a good prediction model?

- **Calibration**
  - Does the model correctly predict the number of people will develop breast cancer?
- **Discriminatory accuracy**
  - Does the model correctly predict exactly who will develop breast cancer?

The Area Under an ROC Curve

- The area measures **discrimination**, that is, the ability of the test to correctly classify those with and without the disease

.90-1 = excellent (A) .80-.90 = good (B) .70-.80 = fair (C) .60-.70 = poor (D) .50-.60 = fail (F)

Population level or Personalized level?[3]

Table 1

Examples of risk prediction models for asymptomatic individuals that have been validated in different populations

	Summary of performance in validation studies	
Risk model	**Discrimination (AUROC, 95% CI)**	**Calibration (O/E, 95% CI)**
Breast (Meads *et al*, 2012)		
Colditz	0.63 (0.63–0.64)	1.01 (0.94–1.09)
Gail 2	0.63 (0.59–0.67)	0.95 (0.88–1.01)
Rosner and Colditz	0.57 (0.55–0.59)	0.96 (0.92–1.02)
Tyrer and Cusick	0.762 (0.70–0.82)[a]	1.09 (0.85–1.41)[a]

Gail model

- **Based** on case-control data from 284,780 women.
- **Risk factors** included :
    - Age
    - Reproductive history
    - Family history
    - Personal history (Biopsies)

- **Validated** using data from NCI's Surveillance, Epidemiology, and End Results (SEER).
    - Caucasian women and African American
    - Asian and Pacific Islander

- **Guideline** based on Gail model

  From The American Society of Clinical Oncology (ASCO)

  five-year risk ≥1.67%: Clinical breast examination at least once per year, annual mammogram, consider high-risk counseling or risk reducing medication.(e.g. tamoxifen)

BOADICEA model

- Based on 2785 UK families
- Included Family pedigree and cancer history, Mutation BRCA 1&2, Ethnicity, and several Biomarkers
- Validated in a large series of families from UK genetics clinics
- In UK and several European countries , it is recommended as a risk assessment tool in clinical guideline
- In Newest UK guideline, Lifetime risk > 30%
    - Screening starting at 30-35 yrs
    - Consider annual MRI starting at 30 yrs
    - Clinical breast exam (annual)
    - Preventive treatment: Consider chemoprevention and preventive mastectomy

Methodology – Machine learning -learn from experience

Three characters:

Learning

- Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model.

Learning more

- The algorithms adaptively improve their performance as the number of samples available for learning increases.

Generate insight for prediction

Leaning techniques/algorithms

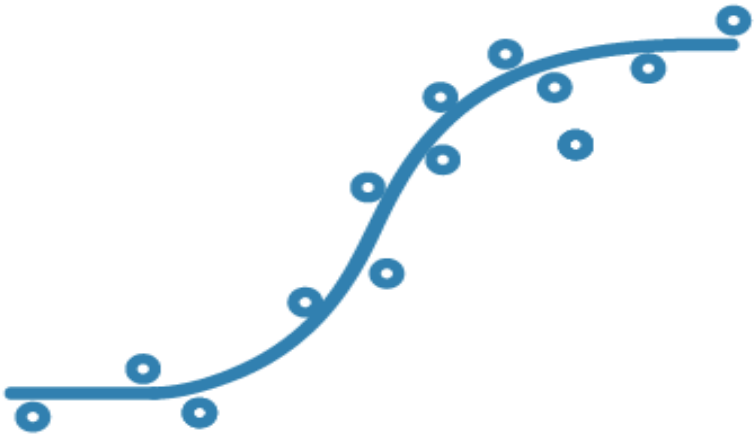
--How should the machine search for “pa
--Depends on whether known responses
learning >>>**Supervised**

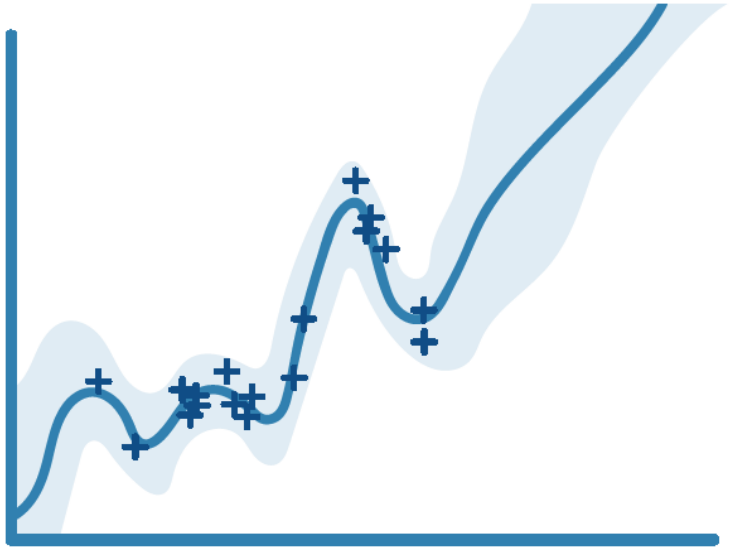
Classification techniques
predict discrete responses

- Binary vs. Multiclass Classification
- Logistic Regression
- k Nearest Neighbor (kNN)
- Neural Network
- Bagged and Boosted Decision Trees

Regression techniques
predict continuous responses

- Generalized Linear Model
- Gaussian Process Regression Model





Study Setting and Materials

- ML v.s. Gail
  - a random population-based
  - US breast cancer patients and their cancer-free female relatives (N=1232)
  - CDC
- ML v.s. BOADICEA
  - a clinic-based sample
  - Swiss breast cancer patients and cancer-free women seeking genetic evaluation and/or testing
  - Geneva University Hospitals (N=1967 Families and 112,482 individual) collected since 1998

Results1

- Always same input data for Gail v.s. ML
  1. simulated, with no signal; N=800

	Gail	ML-ada
acc	0.345	0.384

  1. simulated, with artificial signal; N=800

	Gail	ML-ada
acc	0.711	0.958

Adapt boosting (ada)
Linear discriminant (lda)
Random forest (rf)
Linear Model (lm)
Logistic Regression (Logistic)
k-Nearest Neighbors algorithm (k-NN)
Quadratic Discriminant (qda)

  2. Real data N=1232

	Gail	ML-ada	ML-lda	ML-rf	ML-logistic	ML-knn	ML-qda	ML-lm
acc	0.658	0.897	0.828	0.734	0.855	0.783	0.782	0.334

Results2

- Always same input data for BOADICEA v.s. ML
  1. simulated, with no signal; N=800

	BOADICEA	ML-Logistic	ML-ada
acc	0.279	0.301	0.234

Adapt boosting (ada)
Linear discriminant (lda)
Random forest (rf)
Linear Model (lm)
Logistic Regression (Logistic)
k-Nearest Neighbors algorithm (k-NN)
Quadratic Discriminant (qda)

  1. simulated, with artificial signal; N=800

	BOADICEA	ML-Logistic	ML-ada
acc	0.699	0.953	0.939

  2. Real data N=112,482

	BOADICEA	ML-rf	ML-Logistic	ML-lda	ML-ada	ML-knn	ML-qda	ML-lm
acc	0.671	0.924	0.894	0.881	0.625	0.858	0.812	0.534

Conclusion and Next steps

Advantages of ML:

- **Big improvement in predictive discriminatory accuracy**
- **Not limited by various epidemiology assumptions**
- **Model-free: Free to add any risk factors, e.g Mammographic density**
- **The "bigger" the data, the better the prediction**
- **Easy adaption in application**

Limitations: If not having enough data

SWISS PROMPT:

- first project internationally to apply machine-learning methods in individual breast cancer risk prediction and compares its predictive accuracy with existing models;
- first risk prediction model which is developed using primarily data from Swiss populations;
- will incorporate additional risk factors than existing models. E.g. Modifiable and non-modifiable risk factors

Reference

1. BOUCHARDY MAGNIN, Christine, LOREZ, Matthias, ARNDT, Volker. Effects of age and stage on breast cancer survival in Switzerland. In: Bulletin suisse du cancer
2. Gagnon, J. et al. “Recommendations on Breast Cancer Screening and Prevention in the Context of Implementing Risk Stratification: Impending Changes to Current Policies.” Current Oncology 23.6 (2016): e615–e625. PMC. Web. 20 Nov. 2017.
3. Usher-Smith, Juliet et al. “Risk Prediction Tools for Cancer in Primary Care.” British Journal of Cancer 113.12 (2015): 1645–1650. PMC. Web. 20 Nov. 2017.
4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Shairer C, Mulvihill JJ: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81(24):1879-86, 1989.
5. W. L. Chao, J. J. Ding, “Integrated Machine Learning Algorithms for Human Age Estimation”, NTU, 2011.
6. Easton et al., Nature 2007; 447: 1087-1095
7. Cox et al., Nature Genetics 2007; 39: 352-358
8. Stacey et al., Nature Genetics 2007; 39: 865-869
9. Friedman, Hastie, and Tibshirani, “Additive logistic regression: a statistical view of boosting”, Annals of Statistics, 2000
10. Christopher Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995

Thank you
for your attention.