Generalized additive model for disease risk prediction

Guodong Chen

Chu Kochen Honors College, Zhejiang University
Channing Division of Network Medicine, BWH & HMS

Advised by: Prof. Yang-Yu Liu

12/07/2016
Is it safe to use machine learning in medical data analysis?

• Data for 1M patients
• 1000’s good features
• Train state-of-the-art machine learning model on data
• Model accuracy looks great.
Is it safe to use machine learning in medical data analysis?

- Data for 1M patients
- 1000’s good features
- Train state-of-the-art machine learning model on data
- Model accuracy looks great.

- Is it safe to deploy this model for use on real patient?
- Is accurate prediction on test data enough to deploy?
- What could go wrong?
Is it safe to use machine learning in medical data analysis?

Trained an accurate black box, but don’t understand what’s in the box
What could go wrong…

AUC=0.95

*http://medcitynews.com/wp-content/uploads/Black-Box-Art.png*
AUC = 0.95

Studies compared various learning methods:

- Logistic regression
- Rule-based learning (RBL)
- K-nearest neighbor
- Neural nets
- \ ...
- SVM
- Random Forest

<table>
<thead>
<tr>
<th>ML methods</th>
<th>Interpretability</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>Very good</td>
<td>Poor</td>
</tr>
<tr>
<td>Rule-based learning (RBL)</td>
<td>Very good</td>
<td>Poor</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>Poor</td>
<td>Very good</td>
</tr>
<tr>
<td>Neural nets</td>
<td>Poor</td>
<td>Very good</td>
</tr>
<tr>
<td>SVM</td>
<td>Poor</td>
<td>Very good</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Poor</td>
<td>Very good</td>
</tr>
</tbody>
</table>

Most accurate ML method: ensemble of multitask neural nets

Clinical trail with most accurate model: neural nets?

No! went to clinical trial with logistic regression

Why?
Case study: Predicting Pneumonia Risk (mid 90’s)

- Goal of clinical trial: more informed admission decision
  1. **Low risk:** antibiotics, call us if not feeling better
  2. **High risk:** admit to hospital (5-10% of pneumonia patients die)
- Why go to clinical trials with logistic regression, one of the least accurate methods?
- Rule-based learning (RBL) yields a strange: $\text{HasAsthma}(x) \Rightarrow \text{LessRisk}(x)$

- **True** pattern in the data:
  1. Doctors consider asthmatics presenting with pneumonia very high risk
  2. admitted immediately to hospital, often of ICU (critical care)
  3. asthmatics receive very **aggressive treatment**
  4. Good news: treatment so effective it lowers risk of dying compared to general population
  5. Bad news: if we learned model for in-patient vs. out-patient decision, **could hurt asthmatics**

Question: Probably could prevent neural net from learning this...
• ...not ethical to collect data without treating asthmatics, so we’d have to use ML “tricks” to eliminate the asthma effect
  – remove asthma feature?
  – Train separate models for asthmatics and non-asthmatics?
• Are there other “bad patterns” we don’t know about, e.g., pregnancy?

• Key to discovering HasAsthma(x)… was intelligibility of rules
  1. Neural net is more accurate, so it probably learns things RBL missed
  2. But neural nets are not intelligible, don’t know what they learned
  3. Can’t depend on RBL to discover all of the problems when using more accurate models
• Unfortunately, this is a serious problem…

• Maybe shouldn’t always use most accurate models
Is this fundamental like: can’t have both high accuracy and high intelligibility at same time?

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Intelligibility</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+++</td>
<td>+</td>
</tr>
<tr>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>+++</td>
<td>+</td>
<td>+++</td>
</tr>
</tbody>
</table>
Simple Model

- Linear regression, logistics regression
- Regression: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)
- Classification: \( P[y = 1|x] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \)

Simple, intelligible but less accurate
### Complex Model

- Random forest, SVMs with RBF kernel, etc.
- \( y = f_1, \ldots, n(x_1, \ldots, x_n) \)

Complex and opaque, but often more accurate
Is this fundamental like: can’t have both high accuracy and high intelligibility at same time?
Each feature is “Shaped” by shape function $f_i$

1. Each feature can have a complex non-linear shape, thus the accuracy of addictive models can be significantly higher than the accuracy of simple linear models.

2. Full complexity models such as ensembles of trees are more accurate because they model both non-linearity and interaction, but are so complex that it is nearly impossible to interpret them.

*Walker, SH; Duncan, DB (1967). *Biometrika*.  
*Hastie, T. J.; Tibshirani, R. J. (1990).*
GA$^2$M Algorithm Sketch

1. Build the best additive model $F$ using only one-dimensional Components
   - Additive effects are now modeled
   - If Stage 1 done perfectly, only have interaction (and noise) in residual
2. Fix the one-dimensional functions
   - Detect pairwise interaction on residual (new FAST algorithm)
   - Build shape models for important pairwise interactions on residuals
3. Post-process shape plots
   - Center average prediction of each plot to improve modularity
   - Sort terms by importance to model or patient to aid intelligibility
   - Bag(repeat) process 10-100 times to create pseudo-confidence intervals and further reduce overfitting.

*Yin and Rich, KDD 2013*
Pneumonia in 1995
- 14,199 pneumonia patients
- 70:30 train:test split (train=9847; test=4352)
- 46 features
- **Goal:** Predict POD (probability of death)
- 10.86% of patients (1542) died

30-day Hospital Readmission
- Larger, modern dataset
- Records from large urban hospital 2011-2014
- Train=195,901 (2011-12); test=100,823 (2013)
- 3,956 features for each patient
- **Goal:** Predict probability patient will be readmitted to hospital within 30 days of release
- 8.91% of patients are readmitted within 30 days
Data: Pneumonia 46 Features

<table>
<thead>
<tr>
<th>Patient-history findings</th>
<th>age</th>
<th>gender</th>
<th>diabetes mellitus</th>
<th>asthma</th>
<th>cancer</th>
<th>number of diseases</th>
<th>history of seizures</th>
<th>renal failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>chronic lung disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>re-admission to hospital</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>admitted through ER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>admitted from nursing home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>congestive heart failure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ischemic heart disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cerebrovascular disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chronic liver disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>history of chest pain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical examination findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diastolic blood pressure</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gastrointestinal bleeding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>respiration rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>altered mental status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heart rate</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laboratory findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>liver function tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glucose level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>potassium level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hematocrit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percentage bands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pO2</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sodium level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chest X-ray findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive chest x-ray</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pleural effusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cavitation/empyema</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lobe or lung collapse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| BUN level                        |     |        |                   |        |        |                    |                     |              |
| creatinine level                 | C   |        |                   |        |        |                    |                     |              |
| albumin level                    | C   |        |                   |        |        |                    |                     |              |
| WBC count                        | C   |        |                   |        |        |                    |                     |              |
| pH                               | C   |        |                   |        |        |                    |                     |              |
| pCO2                             | C   |        |                   |        |        |                    |                     |              |
| lung infiltrate                  |     |        |                   |        |        |                    |                     |              |
| pneumothorax                     |     |        |                   |        |        |                    |                     |              |
| chest mass                       |     |        |                   |        |        |                    |                     |              |

Table 1: Pneumonia attributes, grouped by type. Continuous features that will be shaped by GAM/GA²M models are marked with a “C”. 

*Rich and Yin, KDD 2015
**GA^2Ms** Yield State-of-Art Accuracy on these Problems

<table>
<thead>
<tr>
<th>Model</th>
<th>Pneumonia</th>
<th>Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.8432</td>
<td>0.7523</td>
</tr>
<tr>
<td>GAM</td>
<td>0.8542</td>
<td>0.7795</td>
</tr>
<tr>
<td>GA^2M</td>
<td>0.8576</td>
<td>0.7833</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.8460</td>
<td>0.7671</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>0.8493</td>
<td>0.7835</td>
</tr>
</tbody>
</table>

Table 2: AUC for different learning methods on the pneumonia and 30-day readmission tasks.

- GAMs are 26 years old, why are they working so well now?
  - With GA^2Ms we are doing each step very carefully
  - Pairwise interactions are sometimes very important

*Rich and Yin, KDD 2015*
Suppose age is the \( i \)th feature, Risk Score of \( i \) for the patient with age:\( x_i \) is: \( f_i(x_i) - \text{Mean of } f_i \)

\[
\begin{align*}
R_{\text{tot}} &= R_{\text{base}} + \sum_i R_i + \sum_{i,j} R_{i,j} \\
POD &= 1/(1 + \exp(-R_{\text{tot}}))
\end{align*}
\]

\( R_{\text{base}} \) is the average risk score of all patients in the dataset

*Rich and Yin, KDD 2015*
Age Shape Plot for 30-Day Hospital Readmission

- Less affective
- **Very** different datasets with similar graph
- Kind of assurance
- Have the data of the newborn

*Rich and Yin, KDD 2015*
Age Shape Plot for 30-Day Hospital Readmission: Children

- Risk score of readmission for the newborn
- -0.05 is kind of effective
- Most babies are healthy
- Model learns it…

*Rich and Yin, KDD 2015*
Age vs. Cancer Pairwise Interaction

- Risk increases as your age increases without cancer
- **Childhood Cancer**
- This is the most important interaction term

*Rich and Yin, KDD 2015*
Summary of Experiments

• High accuracy on both problems
• Model is pretty understandable
• Terms (i.e. Shape functions) are modular
  • Removing term does not introduce bias, just reduces discrimination
  • Makes it easy to edit/delete terms

• Pneumonia-95
  • Interesting detail on Age, pH, … shape plots
  • Could have gone to clinical trial with this model

• 30-day Hospital Readmission
  • Easy to explain predictions for each patients

*Rich and Yin, KDD 2015
Discussion

• How to sort terms to present to user
  • Affects intelligibility—still not sure how to do this best
  • Few terms: present terms in logical groups
  • Many terms: sort terms by importance to model

• What do we mean by “Intelligible”? 
  • Model is an accurate, intelligible description of itself 
  • Exactly explains the predictions it makes 
  • Not necessarily an accurate, intelligible model of the world. 
  • Ultimately need to do user studies…
  • Makes it easy to edit/delete terms

*Rich and Yin, KDD 2015
My Case: GAM in breast cancer dataset for disease diagnosis

Breast cancer attribute:
1. Clump Thickness 0 - 9
2. Uniformity of Cell Size 0 - 9
3. Uniformity of Cell Shape 0 - 9
4. Marginal Adhesion 0 - 9
5. Single Epithelial Cell Size 0 - 9
6. Bare Nuclei 0 - 9
7. Bland Chromatin 0 - 9
8. Normal Nucleoli 0-9
9. Mitoses 0 - 9

AUC:
- Logistic boosting: 0.9886554621848747
- GAM: 1.0
- GA2M: 1.0
- Random forest: 1.0

*Dataset: http://mlr.cs.umass.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
Case 2: GAM in breast cancer dataset for disease diagnosis

- Risk Score Plot for the Shape functions in breast cancer dataset
Thank you

Guodong Chen

12/07
For More Information

Papers, software and codes:
KDD’12, Yin Lou et al August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08
[2] KDD’13, Yin Lou et al
August 11–14, 2013, Chicago, Illinois, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2174-7/13/08
August 10-13, 2015, Sydney, NSW, Australia. ACM 978-1-4503-3664-2/15/08