Interactive Machine Learning via Transparent Modeling: Putting Human Experts in the Driver's Seat

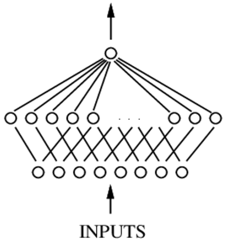
Rich Caruana
Microsoft Research

Joint Work with
Sarah Tan & Yin Lou
Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad

Thanks to
Greg Cooper MD PhD, Mike Fine MD MPH, Eric Horvitz MD PhD
Nick Craswell, Tom Mitchell, Jacob Bien, Giles Hooker, Noah Snavely

When is it Safe to Use Machine Learning in Healthcare?

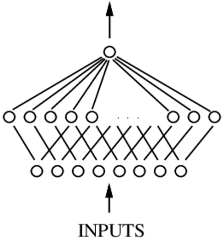
- data for 1M patients
- 1000's great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC $= 0.95$



- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?

When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000's great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC $= 0.95$



- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?

When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000's great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC $= 0.95$

- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?

When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000's great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC $= 0.95$

- is it safe to deploy this model and use on real patients?
- NO! — human expert MUST be able to understand and edit model before use!

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital ($\approx$10% of pneumonia patients die)
- One goal was to compare various ML methods:
  - logistic regression
  - rule-based learning
  - k-nearest neighbor
  - neural nets
  - Bayesian methods
  - hierarchical mixtures of experts
  - ...
- Most accurate ML method: **multitask neural nets**
- Safe to use neural nets on patients?
- **No — we used logistic regression instead...**
- **Why???**

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital ($\approx$10% of pneumonia patients die)

- One goal was to compare various ML methods:
  - logistic regression
  - rule-based learning
  - k-nearest neighbor
  - neural nets
  - Bayesian methods
  - hierarchical mixtures of experts
  - ...

- Most accurate ML method: **multitask neural nets**
- Safe to use neural nets on patients?

- **No — we used logistic regression instead...**
- **Why???**

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital ($\approx$10% of pneumonia patients die)
- One goal was to compare various ML methods:
  - logistic regression
  - rule-based learning
  - k-nearest neighbor
  - neural nets
  - Bayesian methods
  - hierarchical mixtures of experts
  - ...
- Most accurate ML method: **multitask neural nets**
- Safe to use neural nets on patients?
- **No — we used logistic regression instead...**
- **Why???**

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- RBL learned rule: **HasAsthma(x)** => **LessRisk(x)**

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive agressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering **HasAsthma(x)...** was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don't we know about that RBL missed?

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- RBL learned rule: **HasAsthma(x)** => **LessRisk(x)**

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive agressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering **HasAsthma(x)...** was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don't we know about that RBL missed?

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- RBL learned rule: **HasAsthma(x)** => **LessRisk(x)**

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive agressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering **HasAsthma(x)...** was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don't we know about that RBL missed?

Motivation: Predicting Pneumonia Risk Study (mid-90's)

- RBL learned rule: **HasAsthma(x)** => **LessRisk(x)**
- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive agressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population
- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics
- Key to discovering **HasAsthma(x)...** was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don't we know about that RBL missed?

Lessons Learned

- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don't understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop

Lessons Learned

- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don't understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop

Lessons Learned

- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don't understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop

Lessons Learned

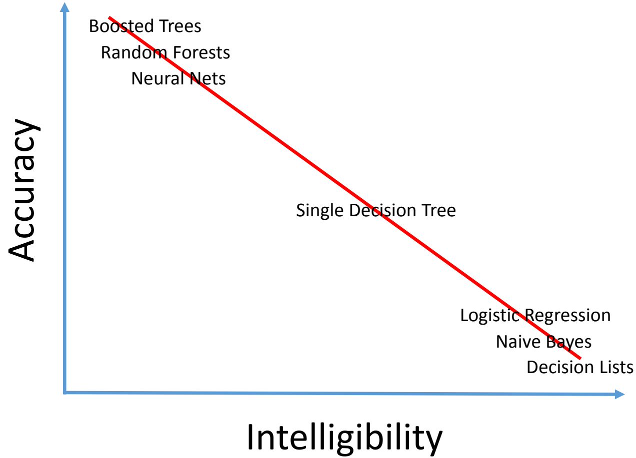
- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don't understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop

Lessons Learned

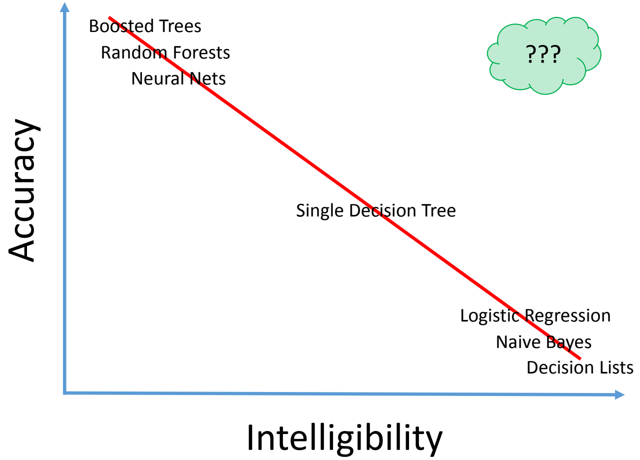
- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don't understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop

To put humans in the driver's seat
all we need is an accurate, intelligible model

Problem: The Accuracy vs. Intelligibility Tradeoff



Problem: The Accuracy vs. Intelligibility Tradeoff



Model Space from Simple to Complex

- Linear Model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
- Additive Model: $y = f_1(x_1) + ... + f_n(x_n)$
- Additive Model with Interactions: $y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ...$
- Full Complexity Model: $y = f(x_1, ..., x_n)$

Model Space from Simple to Complex

- Linear Model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
- Additive Model: $y = f_1(x_1) + ... + f_n(x_n)$
- Additive Model with Interactions: $y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ...$
- Full Complexity Model: $y = f(x_1, ..., x_n)$

Model Space from Simple to Complex

- Linear Model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
- Additive Model: $y = f_1(x_1) + ... + f_n(x_n)$
- Additive Model with Interactions: $y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ...$
- Full Complexity Model: $y = f(x_1, ..., x_n)$

Model Space from Simple to Complex

- Linear Model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
- Additive Model: $y = f_1(x_1) + ... + f_n(x_n)$
- Additive Model with Interactions: $y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ...$
- Full Complexity Model: $y = f(x_1, ..., x_n)$

Add ML-Steroids to old Stats Method: GAMs $\rightarrow$ GA2Ms

- Generalized Additive Models (GAMs)
  - Developed at Stanford by Hastie and Tibshirani in late 80's
  - Regression: $y = f_1(x_1) + ... + f_n(x_n)$
  - Classification: $logit(y) = f_1(x_1) + ... + f_n(x_n)$
  - Each feature is "shaped" by shape function f_i

T. Hastie and R. Tibshirani.
Generalized additive models.
Chapman & Hall/CRC, 1990.

Skip technical details of algorithm and jump to results

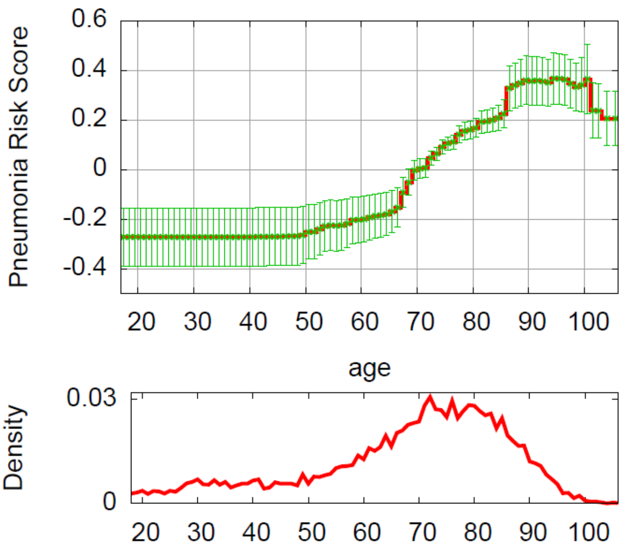
Motivation: Predicting Pneumonia Risk Study (mid-90's)

- Pneumonia Data (dataset from early 1990's)
  - 14,199 pneumonia patients
  - 70:30 train:test split (train=9847; test=4352)
  - 46 features
  - predict POD (probability of death)
  - 10.86% of patients (1542) died

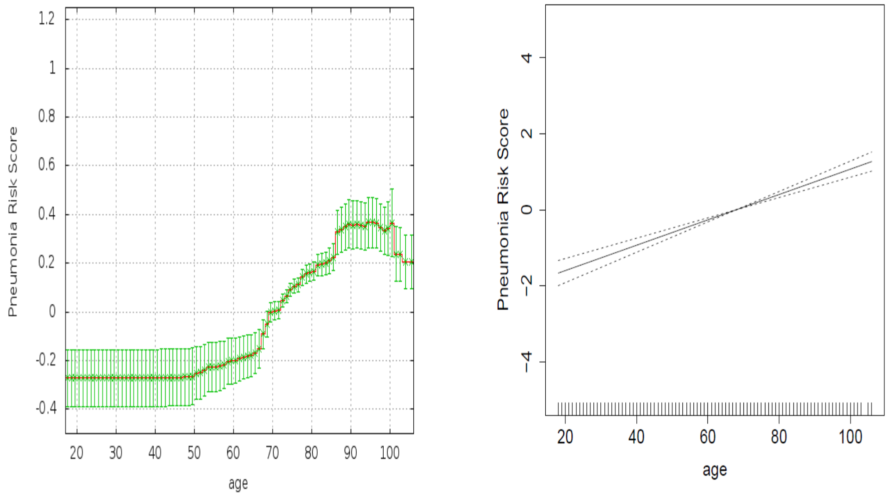
Pneumonia Dataset (mid-90's): 46 Features

Physical examination findings	
Respiration rate (resps/min)	≤ 29*, ≥ 30
Heart rate (beats/min)	≤ 124*, 125–150, ≥ 151
Systolic blood pressure (mmHg)	≤ 60, 61–70, 71–80, 81–90, ≥ 91*
Temperature (°C)	≤ 34.4, 34.5–34.9, 35–35.5, 35.6–38.3*, 38.4–39.9, ≥ 40
Altered mental status (disorientation, lethargy, or coma)	no*, yes
Wheezing	no*, yes
Stridor	no*, yes
Heart murmur	no*, yes
Gastrointestinal bleeding	no*, yes
Laboratory findings	
Sodium level (mEq/l)	≤ 124, 125–130, 131–149*, ≥ 150
Potassium level (mEq/l)	≤ 5.2*, ≥ 5.3
Creatinine level (mg/dl)	≤ 1.6*, 1.7–3.0, 3.1–9.9, ≥ 10.0
Glucose level (mg/dl)	≤ 249*, 250–299, 300–399, ≥ 400
BUN level (mg/dl)	≤ 29*, 30 to 49, ≥ 50
Liver function tests (coded only as normal* or abnormal)	SGOT ≤ 63 and alkaline phosphatase ≤ 499*, SGOT > 63 or alkaline phosphatase > 499
Albumin level (gm/dl)	≤ 2.5, 2.6–3, ≥ 3.1*
Hematocrit	6–20, 20.1–24.9, 25–29, ≥ 30*
White blood cell count (1000 cells/μl)	0.1–3, 3.1–19.9*, ≥ 20
Percentage bands	≤ 10*, 11–20, 21–30, 31–50, ≥ 51
Blood pH	≤ 7.20, 7.21–7.35, 7.36–7.45*, ≥ 7.46
Blood pO_2 (mmHg)	≤ 59, 60–70, 71–75, ≥ 76*
Blood pCO_2 (mmHg)	≤ 44*, 45–55, 56–64, ≥ 65

What GA2Ms on Steroids Learn About Risk vs. Age



Age Shape Plots: GA^{2}M vs. Splines



Splines tend to be too smooth

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses
- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions
- **Important:** Must keep potentially offending features in model!

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses
- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions
- **Important:** Must keep potentially offending features in model!

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk

- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses

- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions

- **Important:** Must keep potentially offending features in model!

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk

- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses

- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions

- **Important:** Must keep potentially offending features in model!

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses
- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions
- **Important:** Must keep potentially offending features in model!

- Some of the things the intelligible model learned:
  - Age 105 is safer than Age 95
  - We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk

- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses

- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions

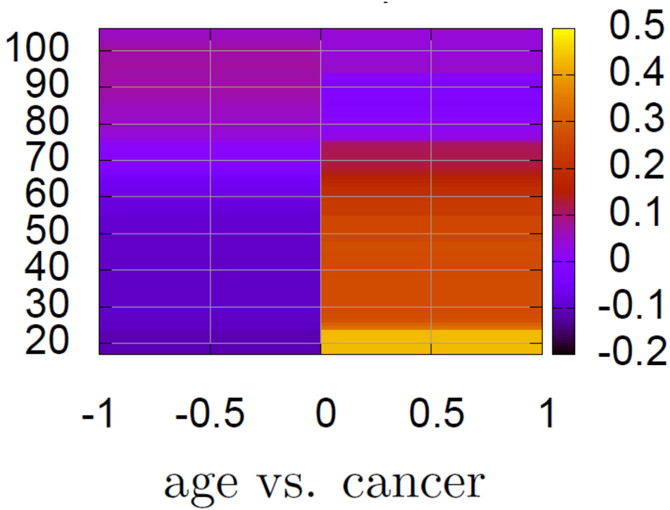
- **Important:** Must keep potentially offending features in model!

Pairwise Interactions?

	0	**1**
0	0	1
1	1	0

- Parity is the classic (extreme) interaction
  - For N-bit parity, need all N bits at same time to calculate parity
  - No correlation between any of the bits and parity signal
  - No information in any subset of the bits
- Interactions can't be modeled as sum of independent effects
- Interactions important on some problems, less on others

Age vs. Cancer Pairwise Interaction (Pneumonia-95)



Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
- Smoothness
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
- Smoothness
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

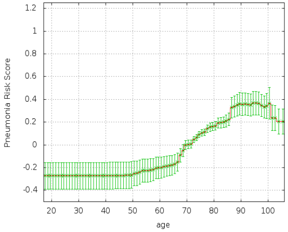
Work in Progress: Can We Make GA2Ms More Intelligible?

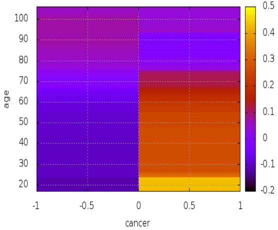
- Over-Paramterization
- Smoothness
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

Over-Parameterization

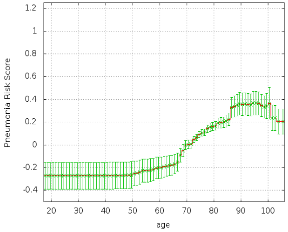
- GA2Ms with pairwise interactions are over-parameterized
- What is over-parameterization?
  - suppose $y = a * x_1 + b * x_1$
  - many ways to set a and b to yield same model because $y = (a + b) * x_1$
  - suppose we want $y = 10 * x_1$
  - then $a = 10$ and $b = 0$, or $a = 5$ and $b = 5$, or even $a = 100$ and $b = -90$ all work
- There's a similar over-parameterization between mains and interactions of those mains

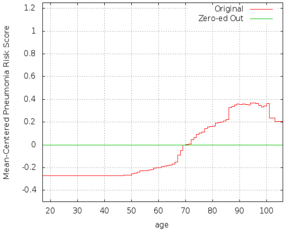
Over-Parameterization: Before Moving Mass from Main to Interaction

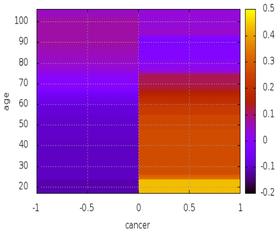




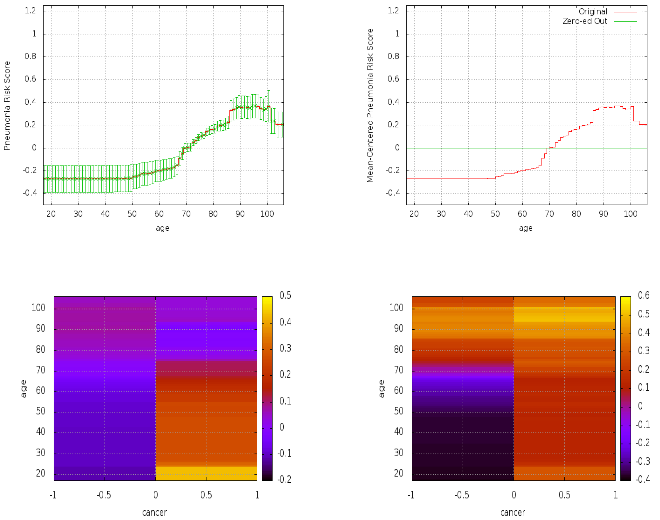
Over-Parameterization: After Moving All Mass From Main to Interaction







Over-Parameterization: After Moving All Mass From Main to Interaction



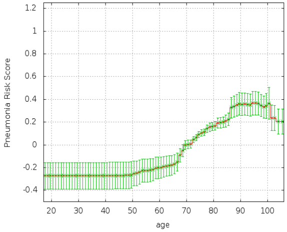
Work in Progress: Can We Make GA2Ms More Intelligible?

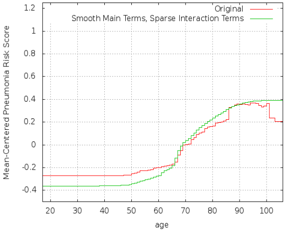
- Over-Paramterization
  - Pushing all mass into interactions can reduce number of terms because some mains go away
  - But can make model harder to interpret because interactions can become more complex
  - If main is involved in more than one interaction, many ways to distribute mass
  - GA2M algorithm currently tries to push mass into mains so pairs are just residuals
  - But over-parameterization and mass-moving provide interesting interactive opportunities
- Smoothness
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

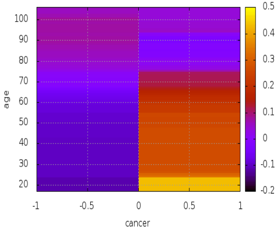
Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
  - Pushing all mass into interactions can reduce number of terms because some mains go away
  - But can make model harder to interpret because interactions can become more complex
  - If main is involved in more than one interaction, many ways to distribute mass
  - GA2M algorithm currently tries to push mass into mains so pairs are just residuals
  - But over-parameterization and mass-moving provide interesting interactive opportunities
- Smoothness
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

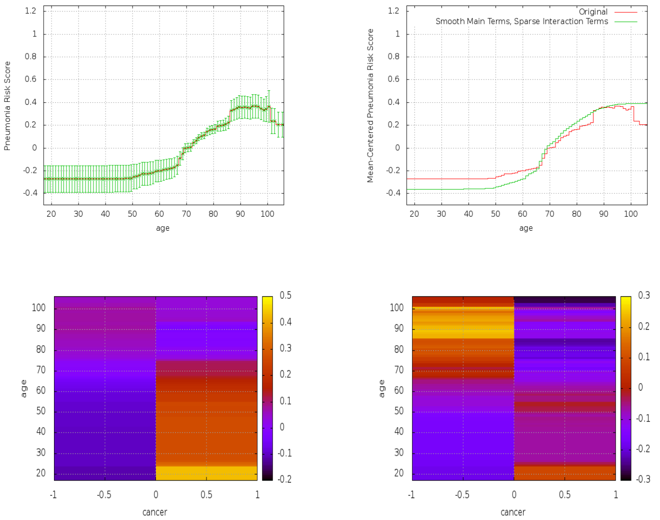
Smoothness: Before and After Optimizing Smoothness of Main Effect







Smoothness: Before and After Optimizing Smoothness of Main Effect



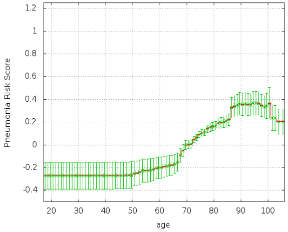
Work in Progress: Can We Make GA2Ms More Intelligible?

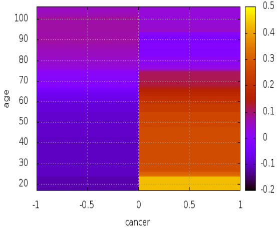
- Over-Paramterization
- Smoothness
  - to add constraint like smoothness to main must add counter-balancing contraint to interactions otherwise optimization will happily move all mass from main to interaction!
  - can achieve simpler, cleaner main but at expense of pushing detail into interactions
  - in general, we don't find extreme smoothness of mains is to be prefered
  - smoothness created monotonicity (by accident), making it look like $age > 100$ is solved
  - but adding explicit contraint for monotonicity is a better way to achieve monotonicity
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

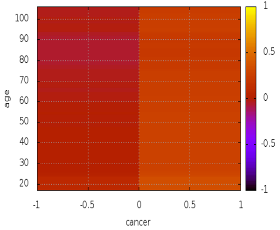
Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
- Smoothness
  - to add constraint like smoothness to main must add counter-balancing contraint to interactions otherwise optimization will happily move all mass from main to interaction!
  - can achieve simpler, cleaner main but at expense of pushing detail into interactions
  - in general, we don't find extreme smoothness of mains is to be prefered
  - smoothness created monotonicity (by accident), making it look like $age > 100$ is solved
  - but adding explicit contraint for monotonicity is a better way to achieve monotonicity
- Sparsity
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

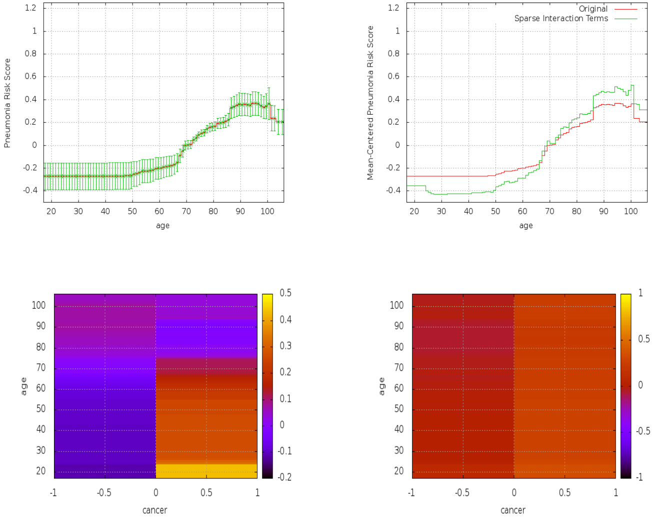
Sparsity: Before and After Optimizing Sparsity of Interactions







Sparsity: Before and After Optimizing Sparsity of Interactions



Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
- Smoothness
- Sparsity
  - Don't have to add counter-balance because can't move all of an interaction to the mains
  - Adding sparsity (or smoothness) to interactions can make them easier to interpret
  - Sometimes seems to hurt mains a little, sometimes doesn't
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

Work in Progress: Can We Make GA2Ms More Intelligible?

- Over-Paramterization
- Smoothness
- Sparsity
    - Don't have to add counter-balance because can't move all of an interaction to the mains
    - Adding sparsity (or smoothness) to interactions can make them easier to interpret
    - Sometimes seems to hurt mains a little, sometimes doesn't
- Monotonicity
- Lasso L1 Regularization (feature selection)
- Tradeoff between simplicity/intelligibility and prediction accuracy
- More causal?
- ...

Smoothness + Sparsity + Monotonicity + Simplicity + L1 + ...

- Most machine learning is about optimizing well-defined criteria such as accuracy
- For each term in a GA2M model (can be 100's or 1000's of terms)
- For each main M and pairwise interaction PI in a GA2M model
- Have the opportunity to optimize smoothness, sparsity, monotonicity, simplicity, L1, ...
- To optimize things like intelligibility, editability, trust, ...
- Don't have objective measures for these so we can't do optimization automatically
- Currently need interactive exploration by human to examine the possibilities

Smoothness + Sparsity + Monotonicity + Simplicity + L1 + ...

- Most machine learning is about optimizing well-defined criteria such as accuracy

- For each term in a GA2M model (can be 100's or 1000's of terms)
- For each main M and pairwise interaction PI in a GA2M model
- Have the opportunity to optimize smoothness, sparsity, monotonicity, simplicity, L1, ...
- To optimize things like intelligibility, editability, trust, ...

- Don't have objective measures for these so we can't do optimization automatically

- Currently need interactive exploration by human to examine the possibilities

Smoothness + Sparsity + Monotonicity + Simplicity + L1 + ...

- Most machine learning is about optimizing well-defined criteria such as accuracy

- For each term in a GA2M model (can be 100's or 1000's of terms)
- For each main M and pairwise interaction PI in a GA2M model
- Have the opportunity to optimize smoothness, sparsity, monotonicity, simplicity, L1, ...
- To optimize things like intelligibility, editability, trust, ...

- Don't have objective measures for these so we can't do optimization automatically

- Currently need interactive exploration by human to examine the possibilities

Work in Progress: Can We Make GA2Ms Easier to Edit?

- Centering to increase modularity
- HCI tools to help experts edit model and understand the impact of those edits
- Statistical tools to help experts understand the impact of their edits on accuracy

Work in Progress: Can We Make GA2Ms Easier to Edit?

- Centering to increase modularity
- HCI tools to help experts edit model and understand the impact of those edits
- Statistical tools to help experts understand the impact of their edits on accuracy

Work in Progress: Can We Make GA2Ms Easier to Edit: Centering

- Modularity of terms makes GA2Ms easier to edit — can we improve modularity?
- Yes, one easy fix: add intercept term to make each term easier to remove

- Suppose $y = mx + b$
- Can't change m or b without changing model

- Now suppose $y = m * graph(x) + b$
- Can shift graph up or down, and just compensate by adjusting b
- $y = m * (graph(x) + c) + b'$ where $b' = b - m * c$

- This is useful for GA2Ms because removing a term (graph) introduces bias
- By centering each graph so mean prediction is zero, we make graphs removable

Work in Progress: Can We Make GA2Ms Easier to Edit: Centering

- Modularity of terms makes GA2Ms easier to edit — can we improve modularity?
- Yes, one easy fix: add intercept term to make each term easier to remove

- Suppose $y = mx + b$
- Can't change m or b without changing model

- Now suppose $y = m * graph(x) + b$
- Can shift graph up or down, and just compensate by adjusting b
- $y = m * (graph(x) + c) + b'$ where $b' = b - m * c$

- This is useful for GA2Ms because removing a term (graph) introduces bias
- By centering each graph so mean prediction is zero, we make graphs removable

Work in Progress: Can We Make GA2Ms Easier to Edit: Centering

- Modularity of terms makes GA2Ms easier to edit — can we improve modularity?
- Yes, one easy fix: add intercept term to make each term easier to remove

- Suppose $y = mx + b$
- Can't change m or b without changing model

- Now suppose $y = m * graph(x) + b$
- Can shift graph up or down, and just compensate by adjusting b
- $y = m * (graph(x) + c) + b'$ where $b' = b - m * c$

- This is useful for GA2Ms because removing a term (graph) introduces bias
- By centering each graph so mean prediction is zero, we make graphs removable

Work in Progress: Can We Make GA2Ms Easier to Edit: Centering

- Modularity of terms makes GA2Ms easier to edit — can we improve modularity?
- Yes, one easy fix: add intercept term to make each term easier to remove

- Suppose $y = mx + b$
- Can't change m or b without changing model

- Now suppose $y = m * graph(x) + b$
- Can shift graph up or down, and just compensate by adjusting b
- $y = m * (graph(x) + c) + b'$ where $b' = b - m * c$

- This is useful for GA2Ms because removing a term (graph) introduces bias
- By centering each graph so mean prediction is zero, we make graphs removable

Work in Progress: Can We Make GA2Ms Easier to Edit: Analysis Tools

- What is the impact of editing model on overall accuracy?
- What is the impact to different kinds of patients?
- Could edit(s) be accomplished just by pushing mass around?
- NO! — this is cheating, using mass moving to hide/shuffle mass!

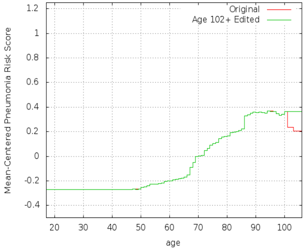
Work in Progress: Can We Make GA2Ms Easier to Edit: Analysis Tools

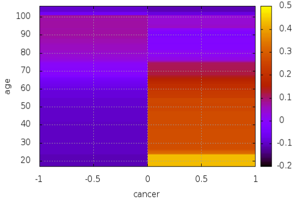
- What is the impact of editing model on overall accuracy?
- What is the impact to different kinds of patients?
- Could edit(s) be accomplished just by pushing mass around?
- NO! — this is cheating, using mass moving to hide/shuffle mass!

Work in Progress: Can We Make GA2Ms Easier to Edit: Analysis Tools

- What is the impact of editing model on overall accuracy?
- What is the impact to different kinds of patients?
- Could edit(s) be accomplished just by pushing mass around?
- NO! — this is cheating, using mass moving to hide/shuffle mass!

- What is the impact of editing model on overall accuracy?
- What is the impact to different kinds of patients?
- Could edit(s) be accomplished just by pushing mass around?
- NO! — this is cheating, using mass moving to hide/shuffle mass!





Advantages of Transparent Modeling for Interactive Data Analysis

- It is much easier to understand a model than to understand the data
- The model will tell you things about the data you never expected
- Don't have to know what to look for in advance
- Don't have to design statistical tests for biases in advance
- Just train model, and look at what it learned — the model will surprise you

- Modularity of GAMs makes many problems easier to recognize
- Modularity of GAMs makes many problems easier to correct

- High accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data

Advantages of Transparent Modeling for Interactive Data Analysis

- It is much easier to understand a model than to understand the data
- The model will tell you things about the data you never expected
- Don't have to know what to look for in advance
- Don't have to design statistical tests for biases in advance
- Just train model, and look at what it learned — the model will surprise you

- Modularity of GAMs makes many problems easier to recognize
- Modularity of GAMs makes many problems easier to correct

- High accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data

Advantages of Transparent Modeling for Interactive Data Analysis

- It is much easier to understand a model than to understand the data
- The model will tell you things about the data you never expected
- Don't have to know what to look for in advance
- Don't have to design statistical tests for biases in advance
- Just train model, and look at what it learned — the model will surprise you

- Modularity of GAMs makes many problems easier to recognize
- Modularity of GAMs makes many problems easier to correct

- High accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data

Comments

- GA2Ms are not causal models
  - because they're simple and transparent, often find causal effects
  - but it's up to the user to figure out what's really going on
- GA2Ms do not cure the curse of dimensionality and correlation
- GA2Ms are intelligible only if features are intelligible
- GA2Ms are not a replacement for deep learning on raw signals
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans
- GA2Ms are not perfect yet...
  - we're still doing research to make the GA2Ms better
  - but they're now good enough to be used instead of logistic regression

Comments

- GA2Ms are not causal models
  - because they're simple and transparent, often find causal effects
  - but it's up to the user to figure out what's really going on
- GA2Ms do not cure the curse of dimensionality and correlation
- GA2Ms are intelligible only if features are intelligible
- GA2Ms are not a replacement for deep learning on raw signals
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans
- GA2Ms are not perfect yet...
  - we're still doing research to make the GA2Ms better
  - but they're now good enough to be used instead of logistic regression

Comments

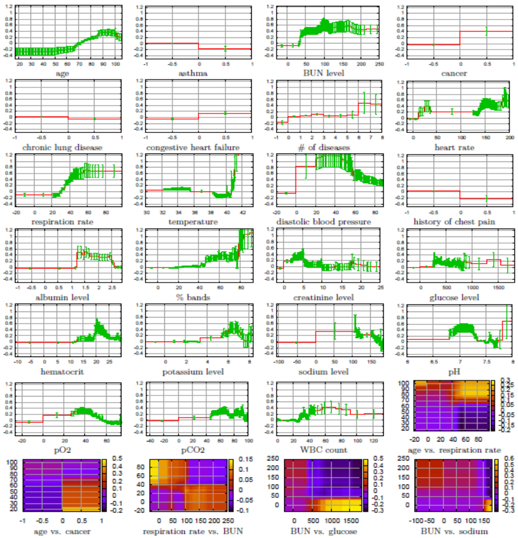
- GA2Ms are not causal models
  - because they're simple and transparent, often find causal effects
  - but it's up to the user to figure out what's really going on
- GA2Ms do not cure the curse of dimensionality and correlation
- GA2Ms are intelligible only if features are intelligible
- GA2Ms are not a replacement for deep learning on raw signals
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans
- GA2Ms are not perfect yet...
  - we're still doing research to make the GA2Ms better
  - but they're now good enough to be used instead of logistic regression

Comments

- GA2Ms are not causal models
  - because they're simple and transparent, often find causal effects
  - but it's up to the user to figure out what's really going on
- GA2Ms do not cure the curse of dimensionality and correlation
- GA2Ms are intelligible only if features are intelligible
- GA2Ms are not a replacement for deep learning on raw signals
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans
- GA2Ms are not perfect yet...
  - we're still doing research to make the GA2Ms better
  - but they're now good enough to be used instead of logistic regression

Summary

- High accuracy on test set is not enough
- There are land mines hidden in the data
- You need magic glasses to see the landmines
- It's critical to understand model before deploying it
- Model correctness depends on how model will be used
- New GA2Ms give us accuracy and intelligibility at same time
- Important to keep potentially offending variables in model so bias can be detected and then removed after training
- If you eliminate offending variables before training you:
  - can't tell you have a problem
  - make it harder to correct the problem
- Transparency allows you to detect problems you didn't anticipate in advance
- Working to develop tools to put expert in the driver's seat — need your help!


age
asthma
BUN level
cancer
chronic lung disease
congestive heart failure
# of diseases
heart rate
respiration rate
temperature
diastolic blood pressure
history of chest pain
albumin level
% bands
creatinine level
glucose level
hematocrit
potassium level
sodium level
pH
pO2
pCO2
WBC count
age vs. respiration rate
age vs. cancer
respiration rate vs. BUN
BUN vs. glucose
BUN vs. sodium


?

30-day Hospital Readmission (joint work with NYP)

- 30-day Hospital Readmission Data
  - larger, modern dataset
  - records from NYP 2011-2014
  - train=195,901 (2011-12); test=100,823 (2013)
  - 3,956 features for each patient
  - goal: predict probability patient will be readmitted within 30 days
  - 8.91% of patients readmitted within 30 days

Quick look at two 30-day Readmission Patients

References

Y. Lou, R. Caruana, and J. Gehrke.
Intelligible Models for Classification and Regression.
In *KDD*, 2012.

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker.
Accurate Intelligible Models With Pairwise Interactions.
In *KDD*, 2013.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, Noemie Elhadad.
Intelligible Models for Healthcare.
In *KDD*, 2015.

T. Hastie and R. Tibshirani.
Generalized additive models.
Chapman & Hall/CRC, 1990.

Thank You!

GA^{2}M Algorithm Sketch

- Stage 1: build best additive model using only 1-dim components
  - Additive effects are now modeled
  - If Stage 1 done perfectly, only have interaction (and noise) in residuals
- Stage 2: fix the one-dimensional functions
  - Detect pairwise interactions on residuals (new FAST algorithm)
- Stage 3: build shape models for most important pairwise interactions on residuals
- Stage 4: post-process shape plots
  - center average prediction of each plot to improve modularity
  - sort terms by importance to aid intelligibility
- Bag (repeat) process 10-100 times to create pseudo-confidence intervals and further reduce overfitting