Taking Advantage of Text for Healthcare, Financial and Other Modeling

Steve Gallant

VP for Research
Textician

MinneBos Conference
Questrom School of Business
Boston University
August 23, 2018

Topics

- Where can text help?
- Why is text hard to use for modeling?
- What are the main technical approaches for modeling (and text representation)?
- Case study: predicting severe sepsis
- Discussion

Where Text Can Help

- **Healthcare** – Under-utilized text in the Electronic Health Record
  - Medical coding (ICD-10, CPT, E&M, …)
  - Predict Severe Sepsis and other important conditions
  - Estimate risk of quick re-admission
- **Financial**
  - Classify news stories
  - Route email to folders
  - Detect spam and unallowable posts
- **Marketing**
  - Select customers for promotions
  - Decide which web pages and content to display

Text Modeling is Hard

- Traditionally, modeling involves things that look like a row in a database.

Age	Income	Have Product A	…	Buy Product B?
36	50,000	1		0
55	75,000	0		1

Doesn't Look Like Text!

- Language is subtle

  "Time flies like an arrow"

  "Fruit flies like a banana"

  -- Groucho Marx

- Medical text presents special difficulties
  - Abbreviations and jargon
  - Negation is critical and difficult, for example with medical coding

    "Not B, C"

    "A, Not B, C, Not D, E"

Overview of Main Approaches for Modeling with Text

1. Text pattern match using human-generated rules

   IF TEXT CONTAINS “Transient”, “myocardial”, “ischemia” AND “newborn”
   THEN CODE= “P29.4”

   - Advantage: Can be very precise, and accommodate special case actions
   - Disadvantages: Brittle, can give poorer results, requires lots of effort and hard to maintain (can be >500,000 rules!)

2. Deep Learning / top-down approach
   - Advantage: Can achieve best predictive performance
   - Disadvantages: Requires lots of data, lots of computation, and lots of hand-holding when building models

Overview of Main Approaches for Modeling with Text (continued)

3. Flat Models / bottom-up approach
   - Start with a vector for each term (eg., word2vec, GloVe), and add/average vectors for terms in a document

   - Takes advantage of pre-training, so similar terms have similar vectors
   - Produces a single vector for a document (eg., distributed vector with 300 dimensions)
   - Advantages:
     - Good performance
     - Faster and easier to build models, and to automate model building
     - Easy to combine with numerical data
     - Can introduce structured elements into the (single) document vector, for example negated terms
   - Disadvantages: Deep learning can give better performance with sufficient data, computation and hand-holding

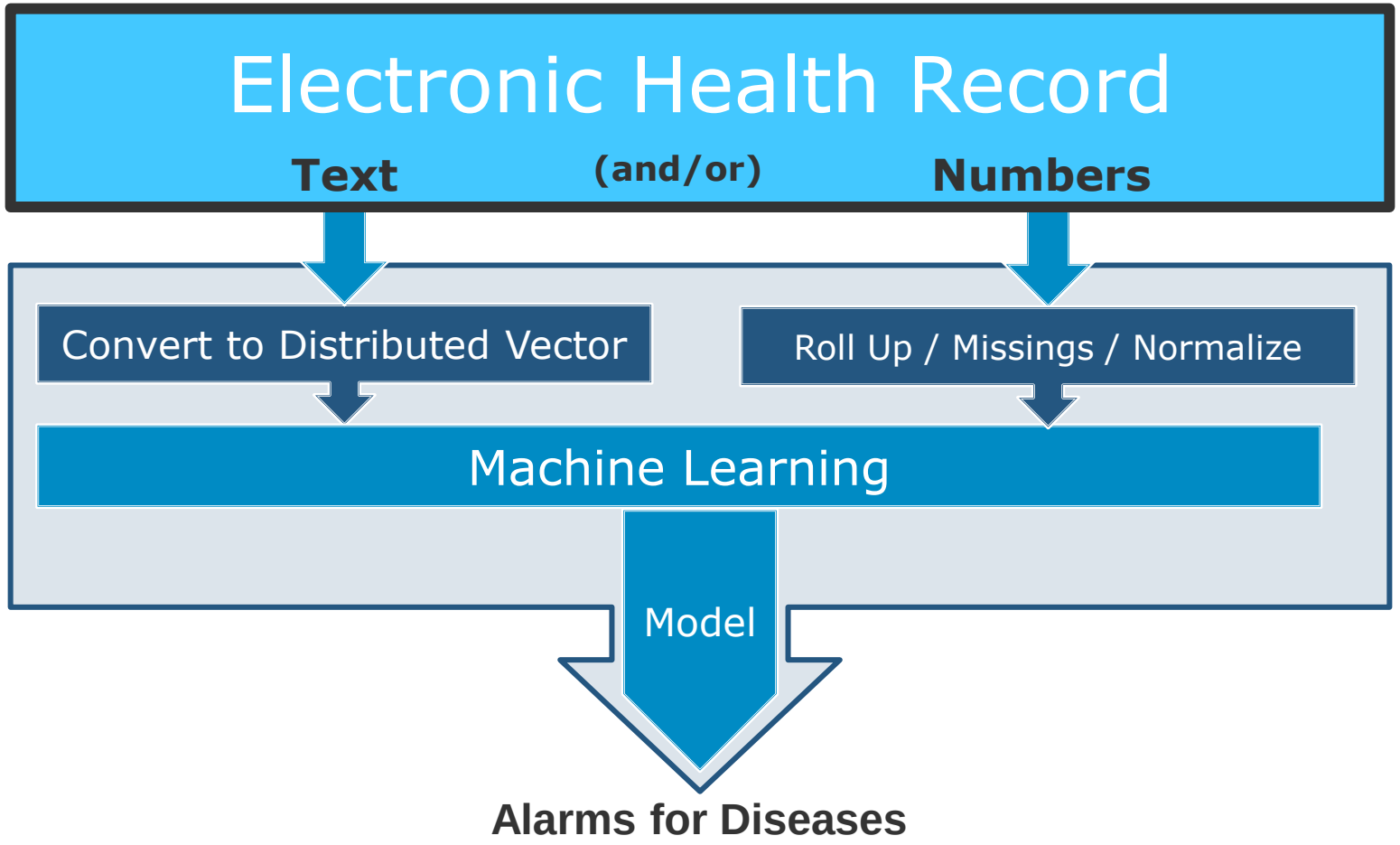
Case Study: Predicting Severe Sepsis

- Sepsis is a life-threatening organ dysfunction caused by infection
- More people die from sepsis than prostate cancer, breast cancer and AIDS combined.
- Sepsis is the No.1 financial burden for hospitals: $24 billion per year

➢ Worked with Baystate Health to predict severe sepsis directly from the Electronic Health Record

- Used unstructured text, structured (numeric) data, and both
- Predicted 4, 8 and 24 hours ahead

Modeling Overview



Findings

- Can we predict Severe Sepsis by using **machine learning** on EHR records from previous patients? **Yes**.
- Can we predict Severe Sepsis using only **textual notes** in the EHR? **Yes**.
- Compare machine learning capability using:
  - Unstructured text only (progress notes, …)
  - Structured Data only (blood pressure, …)
  - Both Unstructured and Structured data

For Sepsis, unstructured models slightly better than structured models. Unstructured + Structured models boost performance an additional 5%.

- Can we use the same techniques to generate other Alarms (Over Sedation, Re-admission risk, …)? **Yes**.

-- S. Gallant, P. Culliton, M. Levinson, A. Ehresman, J. Wherry, J. Steingrub (2018). Predicting Severe Sepsis from the Electronic Health Record Using Machine Learning. American Thoracic Society (ATS) 2018 Conference (abstract), May 18-23, 2018, San Diego.

-- Phil Culliton, Michael Levinson, Alice Ehresman, RN, Joshua Wherry, Jay S. Steingrub, MD, Stephen I. Gallant (2017). Predicting Severe Sepsis Using Text from the Electronic Health Record. Neural Information Processing Systems (NIPS) Workshop on Machine Learning For Health. http://arxiv.org/abs/1711.11536

Discussion

Surprisingly, **text can be easier to use than numerical data**! For example, to build models from information in the Electronic Health Record:

Structured Numeric Data

- Decide which data to include *for this model*
- Determine which are corresponding fields in the Electronic Health Record
- Get this data extract
- Roll up repeated measurements
- Deal with missing data (etc)
- Build models

Unstructured Text

- Extract text from Electronic Health Record (once!)
- Select which types of text (eg, omit billing data)
- Build Models

Doctors often include structured data in their text notes, but *just the important data*

Conclusion

Text is helpful for modeling in Healthcare, Financial, Marketing and other areas, but is under-utilized!

If you have relevant text for modeling, use it!

Thank You!

Steve Gallant

VP for Research
Textician

sgallant@textician.com

Appendix: Predicting Severe Sepsis From EHR Text

Table 1: Predicting Severe Sepsis Using Only Text from the EHR

Predict Ahead		Encounters with Usable Data in Modeling Window	Top 1% Predicted	Top 5% Predicted	Top 10% Predicted
4 hours					
	Sample Size	129,421	1,294	6,471	12,942
	Targets found	2,527	521	801	952
	% of Sample		**40%**	**12%**	**7%**
	% of All Targets		**21%**	**32%**	**38%**
	AUC	**0.636**			
8 hours					
	Sample Size	117,768	1,178	5,888	11,777
	Targets found	2,158	503	769	916
	% of Sample		**43%**	**13%**	**8%**
	% of All Targets		**23%**	**36%**	**42%**
	AUC	**0.660**			
24 hours					
	Sample Size	68,482	685	3,424	6,848
	Targets found	1,427	412	707	829
	% of Sample		**60%**	**21%**	**12%**
	% of All Targets		**29%**	**50%**	**58%**
	AUC	**0.727**			

Appendix: Predicting Severe Sepsis Using Text only, Structured Numeric only, or a Combination

Table 2: Predicting Severe Sepsis 24 Hours in the Future Using Unstructured, Structured, or Combination EHR Data

Type of EHR Data	AUC	Predicted Top 1%		Predicted Top 5%		Predicted Top 10%	
		Number In Set	Severe Sepsis	Number In Set	Severe Sepsis	Number In Set	Severe Sepsis
Unstructured Text only	0.81	136	115	680	217	1,360	247
Structured Data only	0.80	136	112	680	206	1,360	248
Both Unstructured Text And Structured Data	0.85	136	125	680	239	1,360	272