Predicting Hospital Readmission using Machine Learning Models

August 7, 2020

Overview

- Objective:
  - To improve the accuracy of prediction for 30-day unplanned readmission or death using machine learning methods
  - To provide a prediction model that can be used in practice to inform targeting resource-intensive interventions
- Study population
  - All patients discharged from medical inpatient units in any of the acute facilities in the province of Alberta between January 1, 2015 and December 31, 2016
  - We excluded any obstetric or gynecologic hospitalizations

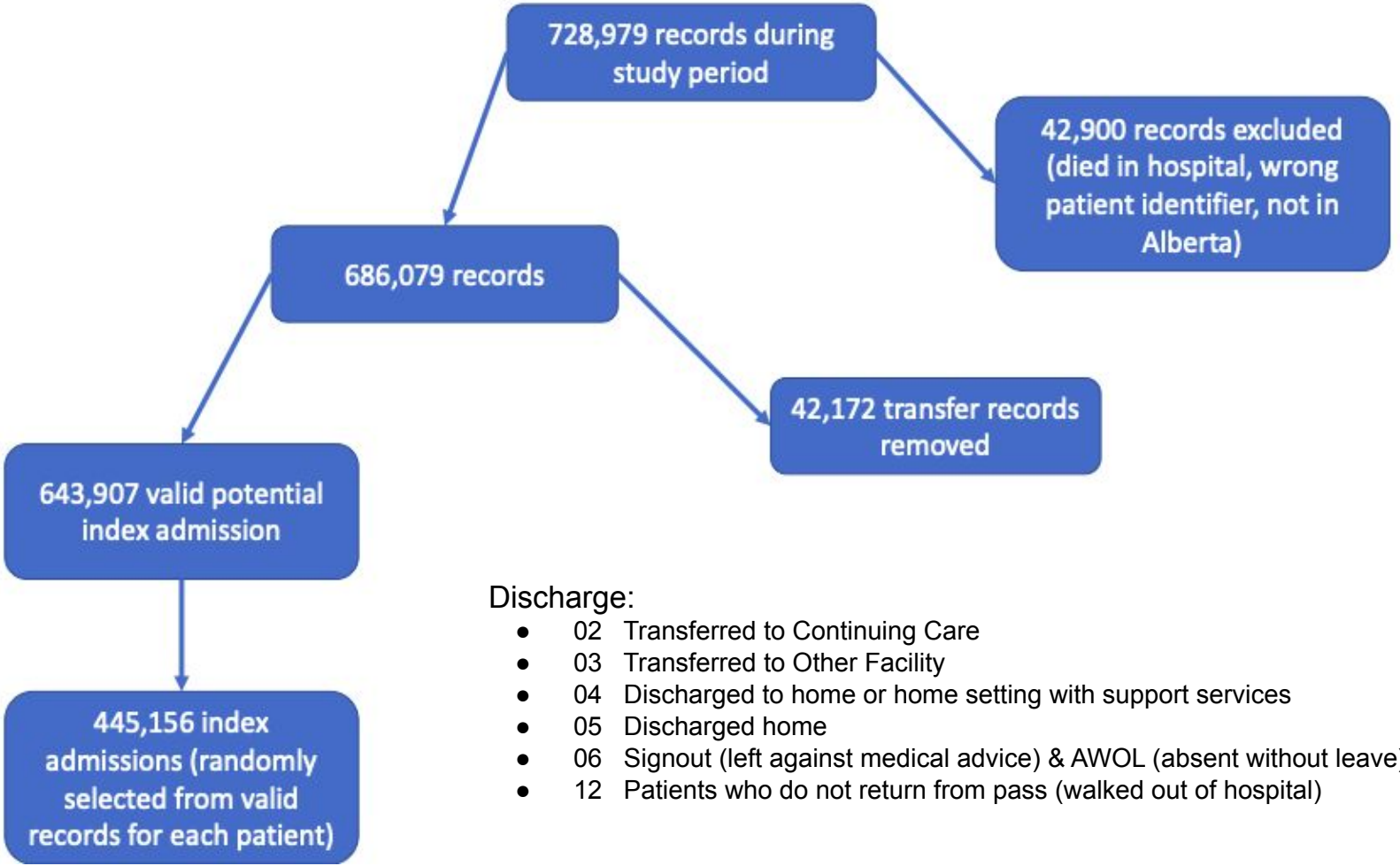
Data

- We have the following for the target population over 2011-2017
- DAD
  - Sex, age, zip codes
  - Institution number, admit date, discharge date, discharge disposition
  - Information about the healthcare providers involved in the patient's care
    - Their role and level of training
  - ICD-10-CA codes and their types
  - CCI codes and their start dates
- NACRS
  - Sex, age, zip codes
  - Institution, mode of visit, date, disposition
  - ICD-10-CA codes
  - CCI codes

Data (continued)

- Lab
  - Date, test code, test name
  - Test reference range, test result, unit of measure
- Claims
  - Start and end dates of the service event
  - Provider’s skill code
  - ICD-9 codes
  - CCPX codes
  - Paid dollar amount
- PIN
  - Date
  - Drug identification number, ATC code
  - Dispensed quantity and daily dosage

Index Admission Selection



Discharge:

- 02 Transferred to Continuing Care
- 03 Transferred to Other Facility
- 04 Discharged to home or home setting with support services
- 05 Discharged home
- 06 Signout (left against medical advice) & AWOL (absent without leave)
- 12 Patients who do not return from pass (walked out of hospital)

Index Episode of Care

Join other DAD records to the index based on the following rules:

- If a transfer DAD record connects to or overlaps with the index's admission date, then we join them
- If another DAD record overlaps with the index's discharge date, we remove the patient (40 were removed)
- If another DAD record is connected to the index's discharge date, then we see it as a same-day readmission

(If a discharge DAD record overlaps with the index, we remove the patient as well, and 818 patients were removed)

42,172 transfers records (disp == "01") in dad data.

- 37,728 (89.46%) is connected with another record
  - 11,472 connections when previous record is real discharge
- 680 (1.61%) overlaps with other records
  - 528 overlaps when the previous record is real discharge
- 1,465 (3.47%) records don't have the following record

Clinical Features

Features that are derived from the data but not through deep learning

DAD

- Directly used
  - **Sex**
  - **Age**
  - **Discharge position**
- Calculated based on data
  - **Length of Stay (admission and discharge date)**
  - **Charlson Score (ICD codes)**
  - **Number of discharged admissions during the past 6 months (previous records)**
  - **Number of discharged admissions during the past year (previous records)**
  - Number of unique procedures during the index episode (CCI codes)
  - Number of total procedures during the index episode (CCI codes)
  - Number of unique diagnosis codes during the index episode (ICD codes)
  - Total length of stay past 6 months and 2 years

NACRS

- Calculated based on data
  - **Number of ED visits records during the past 6 months (previous records)**
  - **Number of ED visits records during the past year (previous records)**
  - **Number of non-ED visits records during the past 6 months (previous records)**
  - **Number of non-ED visits records during the past year (previous records)**

Claims

- Calculated based on data
  - **General Practitioner, General Surgery and Internal Medicine visits in the past 6m and 1 year (Service provider skill code)**
  - **Access to healthcare** - whether a patient has a claim record during the past 2, 3 and 4 years (previous record)
  - Total claimed amount over the past 2, 3 and 4 years (claim amount)

PIN

- Calculated based on data
  - Number of prescriptions in the past 2 years
  - Number of prescriptions during the index hospitalization episode
  - Total prescription days in the past 2 years
  - Total prescription days during index admission
  - Number of drugs patient prescribed over the past two years
  - Number of drugs patient prescribed during index admission

Census

- Calculated based on data
  - **Median after-tax income**
  - **Employment rate**
  - **Unemployment rate**
  - Owner renter ratio (number of owners / number of renters)
  - Percentage of citizens (citizen / population)
  - **Education percentages (Primary, Secondary and Post-Secondary)**

Combined DATA

- **Asthma (ICD codes from DAD and NACRS)**
- **Hypertension (ICD codes from DAD and Claims)**
- **Chronic Heart Failure (ICD codes from DAD and Claims)**
- **Diabetes (ICD codes from DAD and Claims)**

Reasons for Not Using Lab data

- Test names are not consistent
- Some tests cannot be identified as normal/abnormal
  - Reference ranges are missing
  - Result not being a number
  - Results are missing
  - Reference range is not an actual range (eg. unremarkable, few)
  - Open reference range (eg. <8)
- Some results are actual comments

Skip-Gram, Medical Concept Embedding and Patient Vectors

Sarah Davis, August 7th 2020

Setting up the problem

- Each patient can be represented by a sequence of medical events

Patient 1 = [🔶🟩]

Patient 2 = [🔺🔶🔶🔶🔵💜🔺🔶🔶⭐🟩🟩🟩🔶🟩⭐🟩🔵💜🔺🔶🔶⭐🔶
⭐🟩🟩🔶🔶⭐⭐🔶🔵💜🔺🔶⭐⭐🟩🔵💜🔺⭐🟩🟩🔺🔶🟩
🔵💜🔺🔶⭐🔵💜🔺🔶🟩🔵💜🔺🔶🔶⭐🟩🟩🟩🔶🟩⭐🟩🔵]

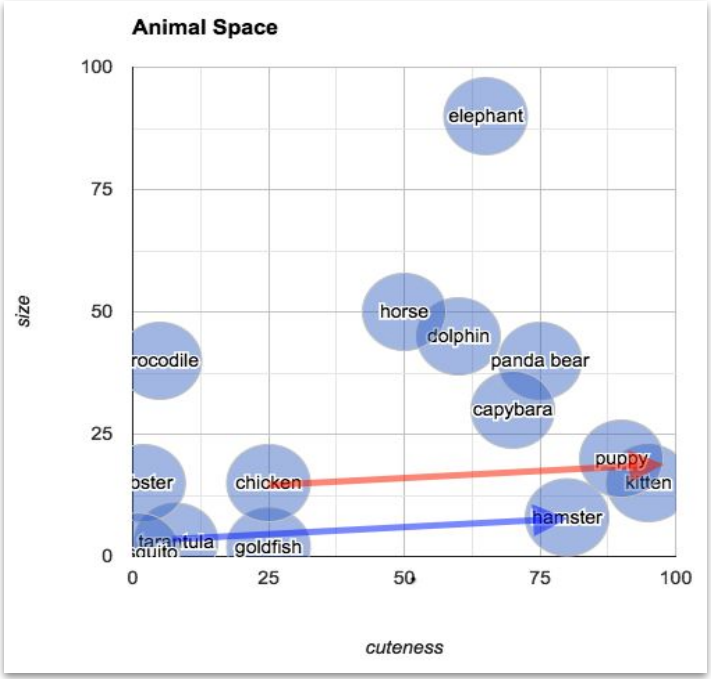
🔺 - Admitted
🔶 - Diagnosis
⭐ - Procedure
🟩 - Prescription
🔵 - Discharged
💜 - Time passes

- This format is problematic for ML, how can we still use all this information?

Question #1: How do we represent medical events?

Skip-Gram Intuition

	cuteness (0–100)	size (0–100)
kitten	95	15
hamster	80	8
tarantula	8	3
puppy	90	20
crocodile	5	40
dolphin	60	45
panda bear	75	40
lobster	2	15
capybara	70	30
elephant	65	90
mosquito	1	1
goldfish	25	2
horse	50	50
chicken	25	15



$$\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$$

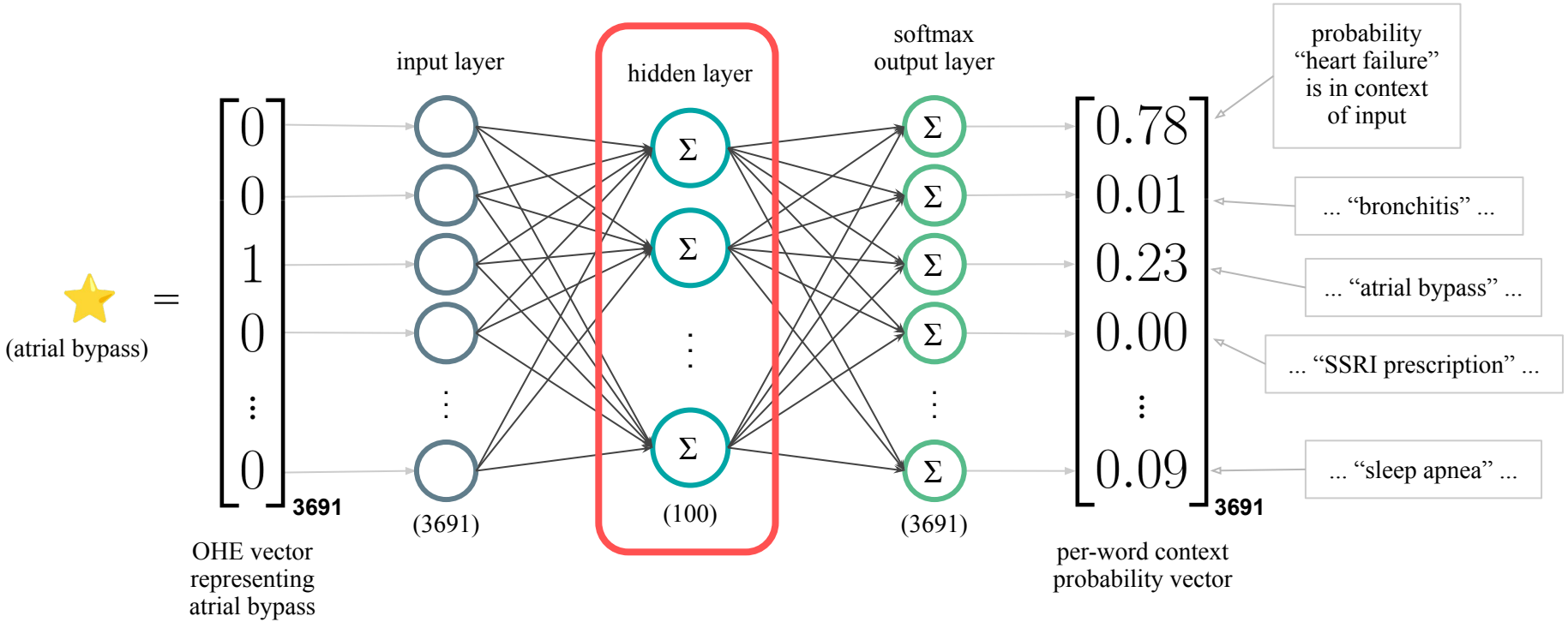
Skip-Gram Preprocessing

Patient 👴 = [🔺🔶⭐🟩🔵]

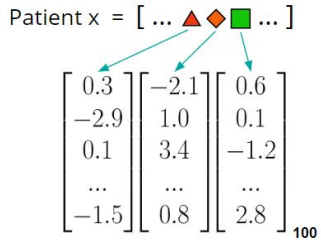
= ADM 848 03.0 C09AA 05

1. Create patient “sentences” like above, includes:
   a. ICD 9 and 10 diagnosis codes, first three characters (Claims, DAD, NACRS)
   b. CCI procedure codes, first five characters (DAD and NACRS)
   c. CCPX procedure code, first four characters (Claims)
   d. ATC drug codes, first five characters (PIN)
   e. Hospital admission event “ADM” (DAD)
   f. Discharge disposition (DAD, NACRS)
2. Replace any word that appears < 100 times in all sentences with “RAREWORD”
3. When # time between codes >= 7 days, insert special “time” word
   a. “0-1m”, “1-3m”, “3-6m”, “6-12m”, “12+m”
4. Collapse long stretches of repeating codes between “time” words, scramble remaining words
5. Over all patient sentences, calculate probability each word appears within the vicinity of every other word

Skip-Gram Training



Final Medical Concept Vectors



	0	1	2	3	4	5	6	...	93	94	95	96	97	98	99
0-1m	-1.157390	0.487504	-0.936831	-0.202482	0.414768	-0.398168	0.625941	...	-0.882061	1.444992	0.694904	-0.495691	0.856951	1.295983	-1.490855
001	-0.289871	0.571144	0.125561	0.122935	0.157949	-0.394100	-0.223712	...	-0.365266	0.055004	0.483395	0.102425	-0.125760	0.272506	-1.208958
003	-0.626011	-0.021549	0.518291	0.202162	0.046956	-0.918956	-0.260235	...	-0.266420	0.206398	0.128862	0.408459	-0.434505	0.121167	-0.069192
005	-0.738970	0.465745	1.293918	0.349703	0.881043	-0.792371	-0.342715	...	-0.336819	0.764527	0.032005	0.900989	-0.713227	0.266496	-0.386682
006	-0.219025	-0.316715	0.043520	0.304523	0.375367	-0.287579	-0.147417	...	-0.123564	-0.099895	0.295174	0.181096	-0.084490	0.289431	-0.242442
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
__12	0.461486	0.152939	0.547748	-1.686489	-0.765570	-1.354261	-0.926514	...	1.451663	-0.858107	1.681526	-0.463042	-1.508841	0.817536	-0.561484
__13	2.396054	-1.156369	-0.309704	-1.263768	-0.340699	-0.630174	0.169734	...	1.148933	-0.365349	1.004886	0.433954	-1.823282	-0.172671	1.353940
__14	0.766981	-1.306420	-1.114545	-1.634153	-0.902243	-0.783598	-0.774667	...	1.303830	-1.924817	1.389760	0.966415	-2.315394	0.386851	1.132255
__15	3.578797	1.130851	1.803574	2.020663	-1.830548	2.896788	0.922845	...	0.060392	-2.145374	-1.032140	-2.529413	2.775403	0.634863	1.115544
adm	5.158481	-1.197036	1.813345	-0.358255	0.018905	3.788475	-2.438269	...	0.464651	-1.704977	-2.039047	-3.790647	-3.922381	-2.721766	2.590863

3691 rows × 100 columns

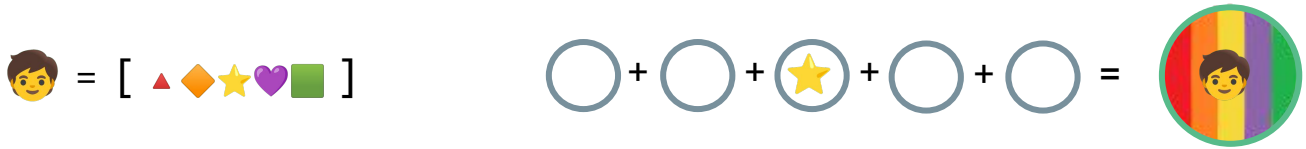
1NV89 - Appendectomy, CIHI
('K35', 0.9434328079223633) Acute appendicitis (ICD 10)
('59.0', 0.9383569359779358) Appendectomy {Appendectomy with or without abscess} (CCPX)
('540', 0.9228229522705078) Accute appendicitis (ICD 9)
('59.1', 0.9102750420570374) Drainage of appendiceal abscess {Appendectomy with or without abscess} (CCPX)
('541', 0.8842280507087708) Appendicitis, unqualified (ICD 9)
('542', 0.8088841438293457) Other appendicitis (ICD 9)
('K38', 0.7951012849807739) Other diseases of appendix (ICD 10)
('K36', 0.7671115398406982) Other appendicitis (ICD 10)
('543', 0.7277620434761047) Hyperplasia of appendix (lymphoid) (ICD 9)
('K37', 0.7113857269287109) Unspecified appendicitis (ICD 10)

adm - Admission of any kind, DAD (all the associated codes are ICD 10)
('Z72', 0.4798443615436554) Problems related to lifestyle
('Z86', 0.4744756817817688) Personal history of certain other diseases
('Z87', 0.46461647748947144) Personal history of other diseases and conditions
('K21', 0.4364264905452728) Gastro-esophageal reflux disease
('E66', 0.4279201328754425) Overweight and obesity
('R52', 0.4227736294269562) Pain, unspecified
('R00', 0.417984902858734l3) Abnormalities of heart beat
('F41', 0.413082718849l8213) Other anxiety disorders
('M79', 0.40182837843894960) Other and unspecified soft tissue disorders, not elsewhere classified
('R79', 0.40095287561416626) Other abnormal findings of blood chemistry

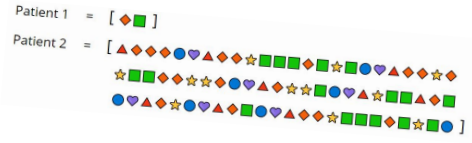
Z33 - Pregnant state, ICD 10
('5AB03', 0.8187727928161621) Obstetrical ultrasound (CIHI)
('O20', 0.7615938782691956) Hemorrhage in early pregnancy (ICD 10)
('O21', 0.7135295271873474) Excessive vomiting in pregnancy (ICD 10)
('O23', 0.6850298643112183) Infections of genitourinary tract in pregnancy (ICD 10)
('5AC20', 0.6587706208229065) Antepartum administration / pharmacotherapy (CCPX)
('K37', 0.6572344303131104) Unspecified appendicitis (ICD 10)
('O47', 0.654293954372406) False labour (ICD 10)
('A74', 0.6359087228775024) Other diseases caused by chlamydiae (ICD 10)
('761', 0.6305762529373169) Other delivery complications (ICD 9)
('651', 0.6242740154266357) Multiple gestation (ICD 9)

Question #2: How do we represent a patient?

- Recall $\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$
- According to this logic, a patient can be represented by sum of all the medical vectors associated with their medical records



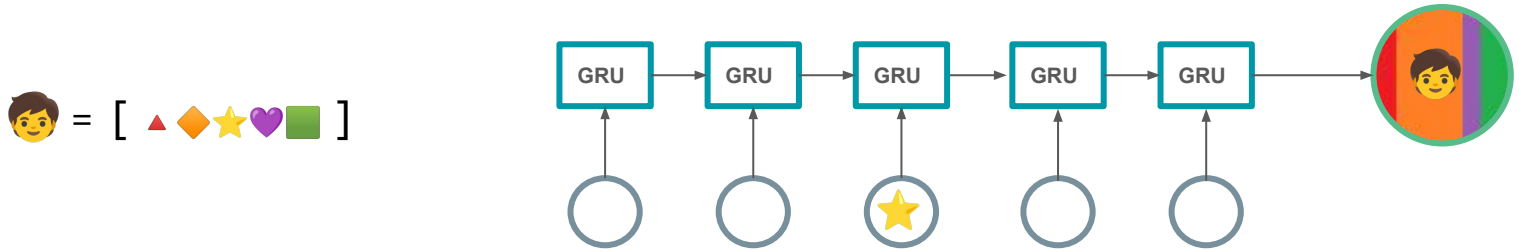
- Leads to a few problems…
  - Some patients have more codes than others
  - If we restrict the number of codes, how many should we use?



- Readmission prediction results using different cutoff points reported in the final section

What are some future methods to explore?

- Some codes in an EHR may not be very important to include for 30-day readmission prediction
  - e.g. repeating insulin prescriptions for a diligent patient with type I diabetes
- Using a GRU-RNN (gated recurrent unit recurrent neural network), certain codes can be “forgotten” when creating the final patient vector

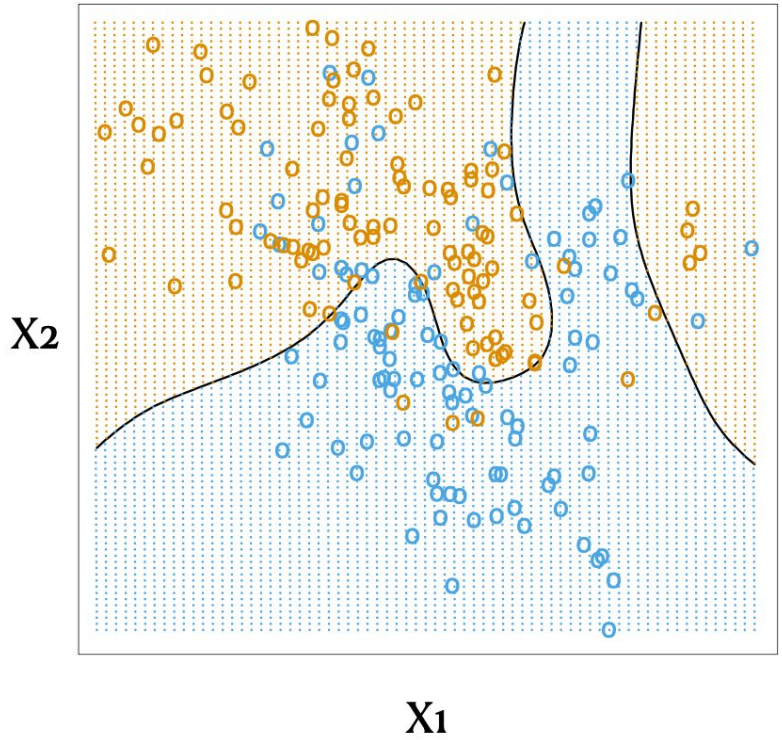


- Allows for “long term memory” of important medical events

Models

Models for predicting readmissions

- We are interested in predicting an outcome variable, Y, that indicates whether a patient will be readmitted within 30 days (Y=1) or not (Y=0).
- Since Y is a binary variable, the problem is a binary classification problem.
- Instead of predicting the value of Y directly, we predict the probability that Y=1.
- This probability will be a function of our predictor variables, X_1 and X_2: $Pr(Y=1) = f(X_1, X_2)$.
- The models that we use differ in how they build the function $f(X_1, X_2)$.



Observation with Y=1

Observation with Y=0

Logistic regression

- One of the oldest models for classification.
- Assumes the logarithm of the odds ratio is a linear function of the predictors.
- Each predictor has a single number (i.e., coefficient).
- $\beta_1 > 0$ means that an increase of X_1 is associated with an increase in Pr(Y=1).

$$\log \frac{\Pr(Y = 1 \mid X)}{\Pr(Y = 0 \mid X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Generalized additive model

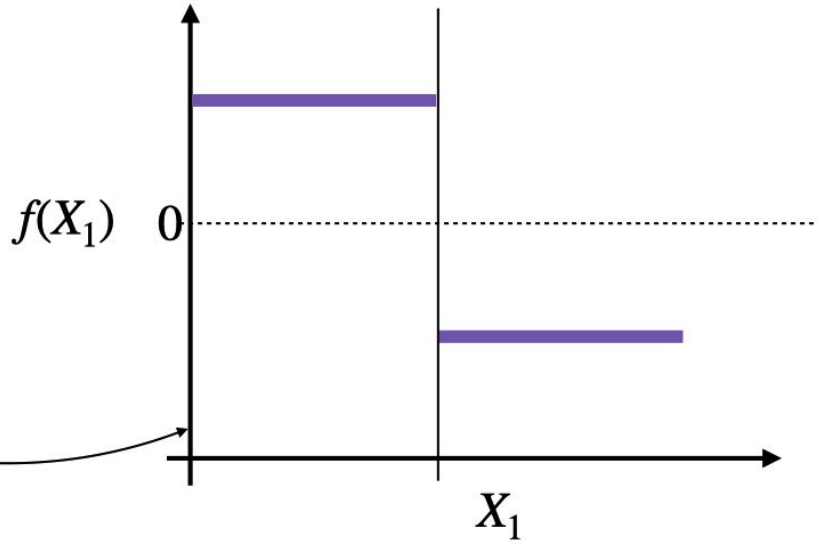
- Similar to logistic regression but now instead of a single number, each predictor has its own function.

$$\log \frac{\Pr(Y=1|X)}{\Pr(Y=0|X)} = \alpha + f_1(X_1) + f_2(X_2)$$

Generalized additive model

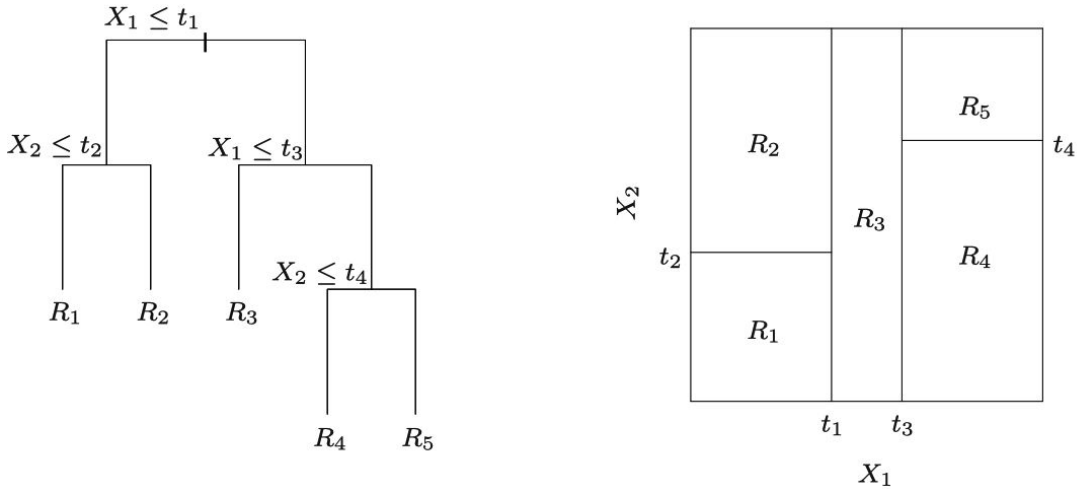
- Similar to logistic regression but now instead of a single number, each predictor has its own function.

$$\log \frac{\Pr(Y=1|X)}{\Pr(Y=0|X)} = \alpha + f_1(X_1) + f_2(X_2)$$



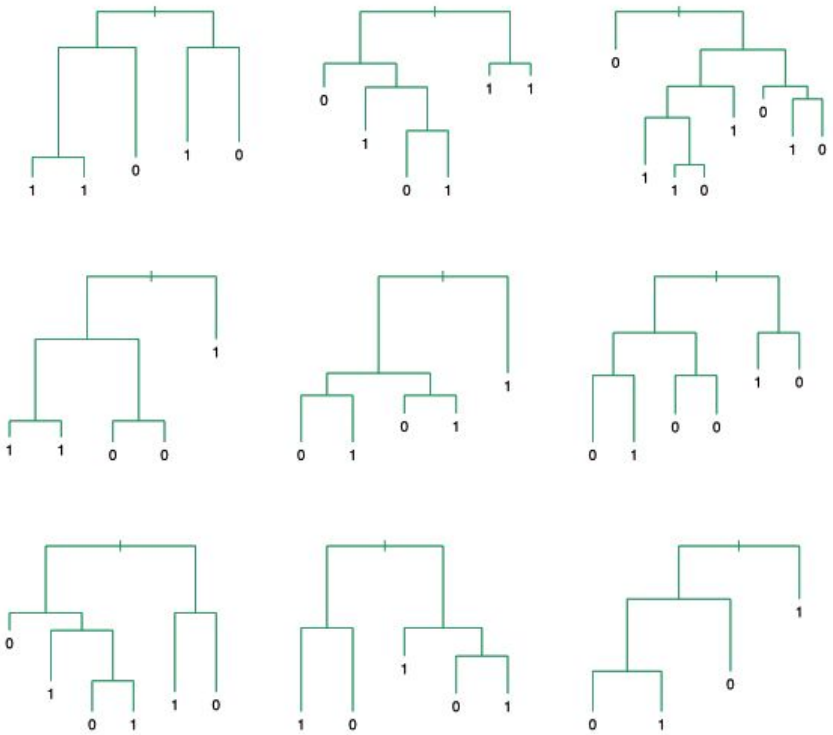
Decision trees

- Another strategy is to segment the (X_1, X_2) space and assign different probabilities to each segment.
- Since the segmentation rule can be represented as a decision tree, this model is called a classification decision tree.



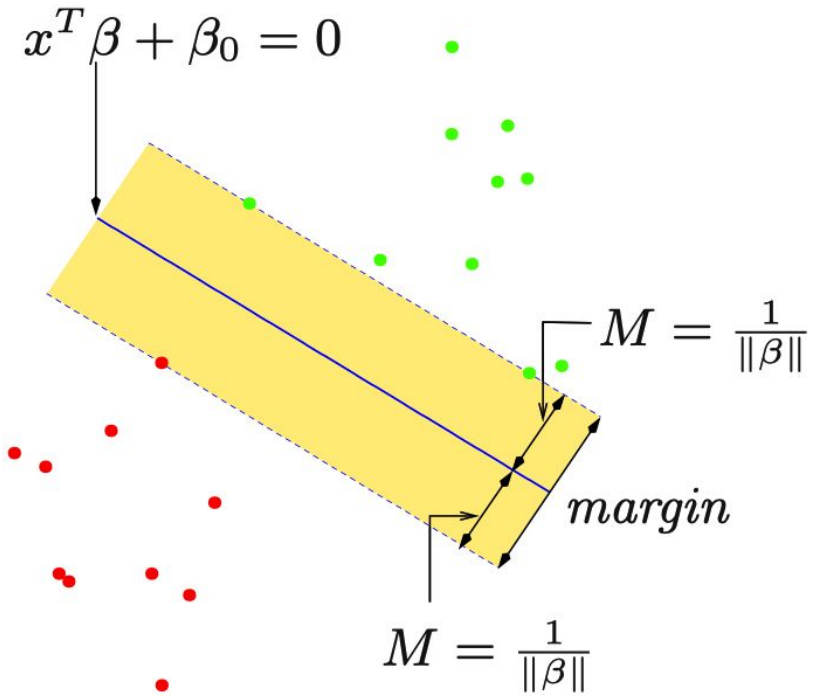
Gradient boosting machine

- Instead of building one decision tree, a gradient boosting machine, build many (e.g., thousands).
- Furthermore, the trees are built sequentially so that the next tree will perform well in places where the previous one did not.



Support vector machine

- A support vector machine tries to find a decision boundary such that there will be a big “gap” between the Y = 1 and Y = 0 cases.
- Furthermore, in order to find non-linear patterns, it maps the data from the (X_1, X_2) space to a higher dimensional space.



Results, Next Steps, and Discussions

Before You See the Results ...

- Performance measure: AUC
- We divided the data into 80% training, 10% validation, and 10% test sets
  - The test set has not been touched
  - All models were evaluated on the validation set
- The current models predict 30-day readmission, not ‘unplanned 30-day readmission or death’
- The baseline is an incomplete version of LACE
  - # of emergent admissions is missing
  - # of NACRS visits was used instead of # of emergency room visits
- Skip-gram features used the last 25 codes of each patient

Current Performance

Features	Model	AUC	Environment
LACE* (baseline)	Logistic regression	.625	R
Only clinical	Logistic regression	.735	
	GBM	.809	
Only skip-gram	Logistic regression	.759	Python
	GBM	.777	
	SVM	.796	
Using both	Logistic regression	.778 / .759	Python / R
	GBM	.810 / .816	
	SVM	.812 / NA	

Varying the Coverage of Patient Sequences

- Using both clinical and skip-gram features
- Tested on Python

# of last codes used	Logistic regression	GBM	SVM
10	.777	.813	.816
25	.778	.810	.812
50	.766	.807	.799
100	.751	.798	.781

Next Steps

- Extracting additional data
- Re-building the models
- Adding more features
- Trying other deep learning approaches
- Parameter tuning
- Writing a paper

Adding More Variables?

1. The admit code in DAD
2. Death records of the target population between 2015 and 2017
3. ED indicator in NACRS
    - NACRS contains all outpatient visits, such as day surgery, outpatient and community-based clinics, and ED visits

(Note: if we get #1 and #3, we can build LACE)

4. There are other optional ones, too ...

Questions about the Data

1. In DAD records, if a patient receives the same procedure twice, does the CCI code appear twice in the DAD record?
2. What should we do about “weird” patient identifiers?

Weird Patient Identifier

Normally: 12 digits : 002552245639

Patient Identifier Length	Number of Patients (admitted 2015&16)	Example / Format	Number of patients (all dad records)
1	2	1 or 0	2
2	-	01 or 99	2
5	1	60870	2
7	59	682131E	79
8	1548	01315597 D2005148	1642
9	280	. .3212 .01950293	315

Questions?