Machine Learning for Healthcare

6.871, HST.956

Lecture 5: Learning with noisy or censored labels

David Sontag

CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Course announcements

- No recitation this Friday, but will be an extra office instead (2pm, 1-390)
- Problem set 1 due Mon Feb 24th 11:59pm

Roadmap

- **Module 1: Overview of clinical care & data** (3 lectures)
- **Module 2: Using ML for risk stratification and diagnosis (9 lectures)**
  - **Supervised learning with noisy and censored labels**
  - NLP, Time-series
  - Interpretability; Methods for detecting dataset shift; Fairness; Uncertainty
- **Module 3: Suggesting treatments** (4 lectures)
  - Causal inference; Off-policy reinforcement learning

QUIZ

- **Module 4: Understanding disease and its progression** (3 lectures)
  - Unsupervised learning on censored time series with substantial missing data
  - Discovery of disease subtypes; Precision medicine
- **Module 5: Human factors** (3 lectures)
  - Differential diagnosis; Utility-theoretic trade-offs
  - Automating clinical workflows
  - Translating technology into the clinic

Outline for today's class

1. **Learning with noisy labels**
   - Two consistent estimators for class-conditional noise (Natarajan et al., NeurIPS '13)
   - Application in health care (Halpern et al., JAMIA '16)
2. Learning with right-censored labels

Labels may be noisy

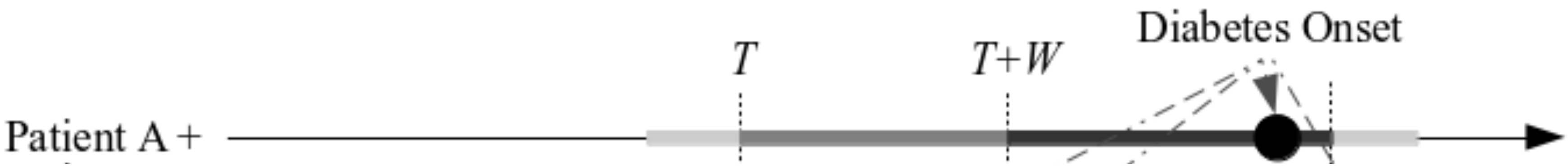
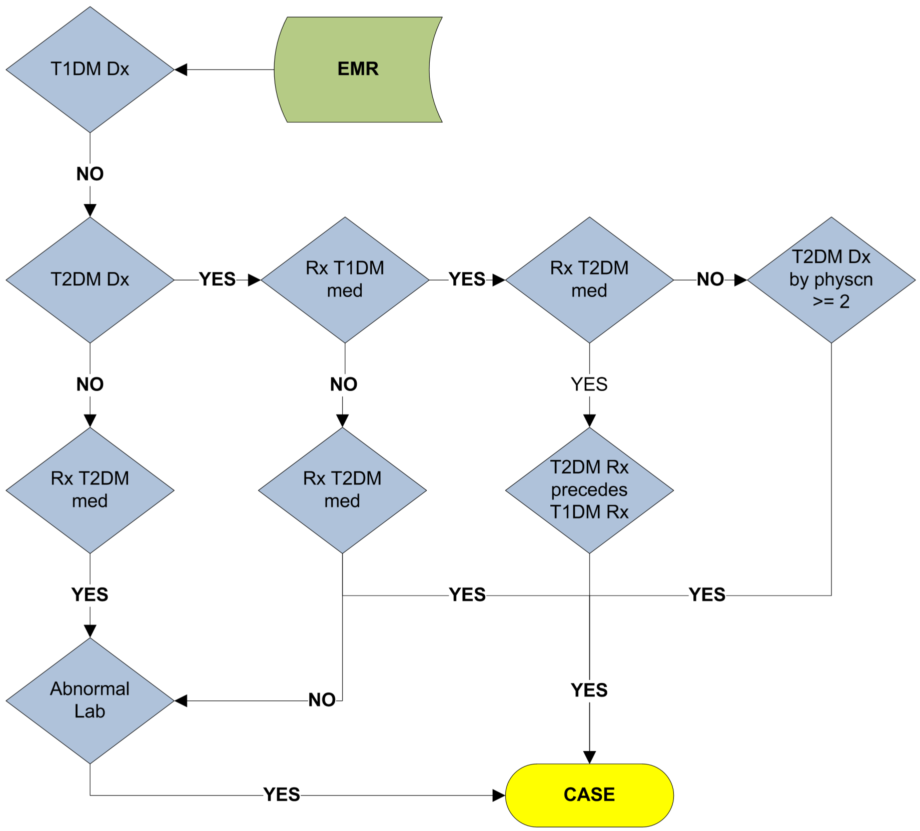
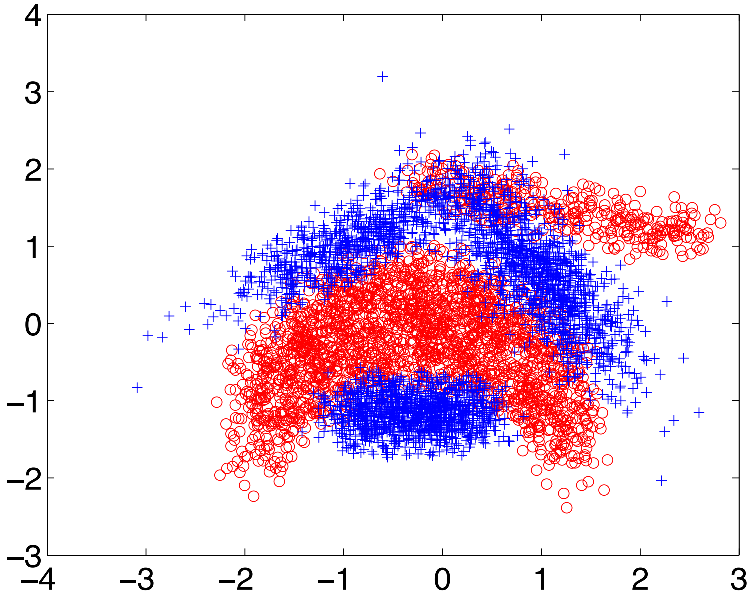


Figure 1: Algorithm for identifying T2DM cases in the EMR.

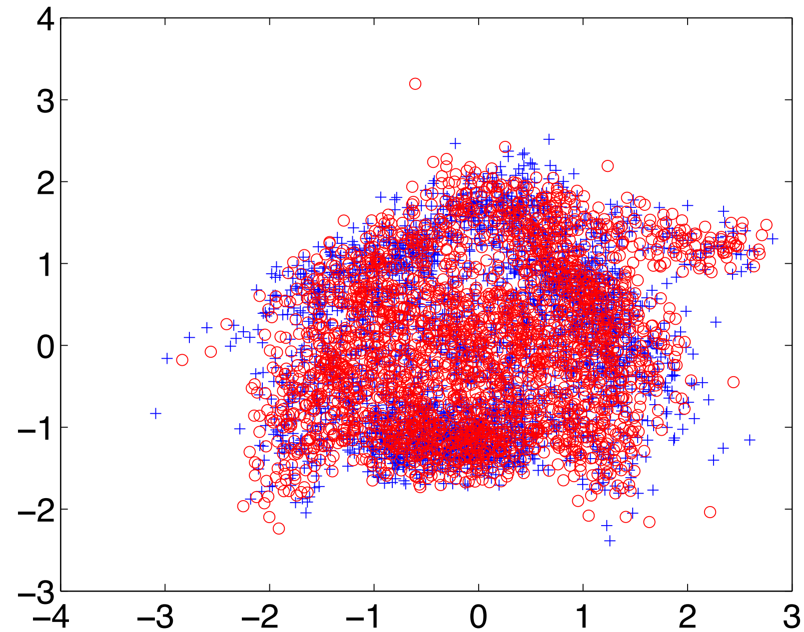


Source: https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf

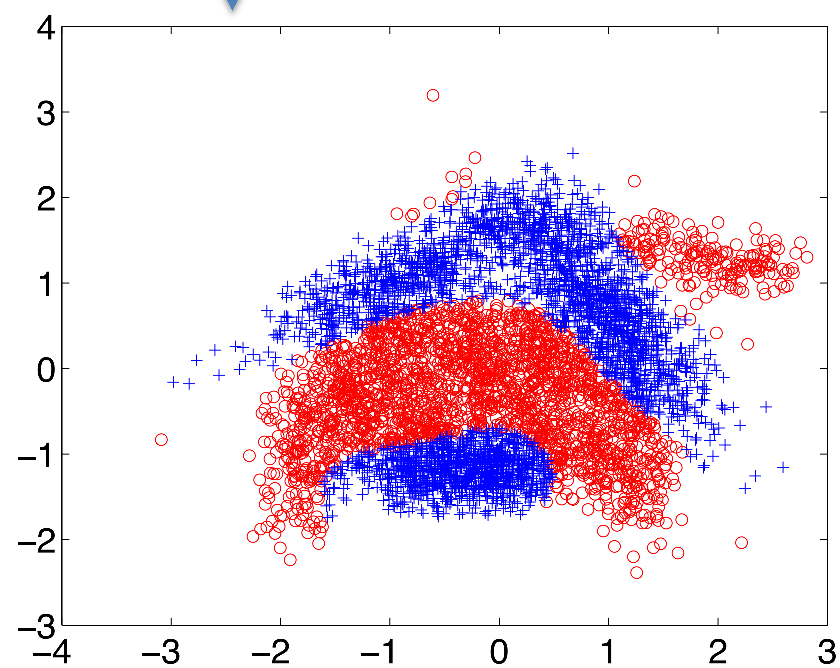
If the derived label is noisy, how does it affect learning?



40% label noise



Machine learning



[Natarajan et al., NeurIPS ’13. Figure 2]

Tl;dr of learning with noisy labels

1. If we are in a world with
   a) class-conditional label noise and
   b) lots of training data,

   learning as usual, substituting noisy labels, works!

2. We can modify learning algorithms to make them work better with label noise.

   Two methods from Natarajan et al. '13:

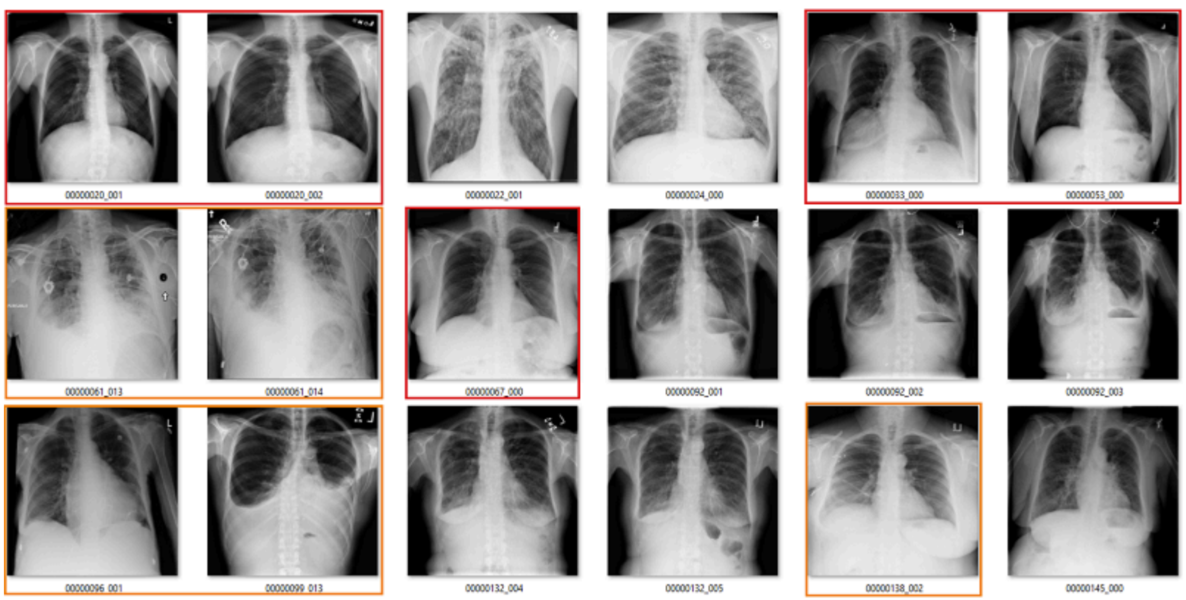
   a) Re-weight the loss functions
   b) Modify (suitably symmetric) loss function

(Natarajan et al., *Learning with Noisy Labels.* NeurIPS '13)

Comments on learning with noisy labels

- Cross-validation to choose parameters uses a separate validation set with *noisy* labels
- What about instance-dependent noise?

Fibrosis



red = mislabeled
orange = maybe mislabeled

Figure source: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

Comments on learning with noisy labels

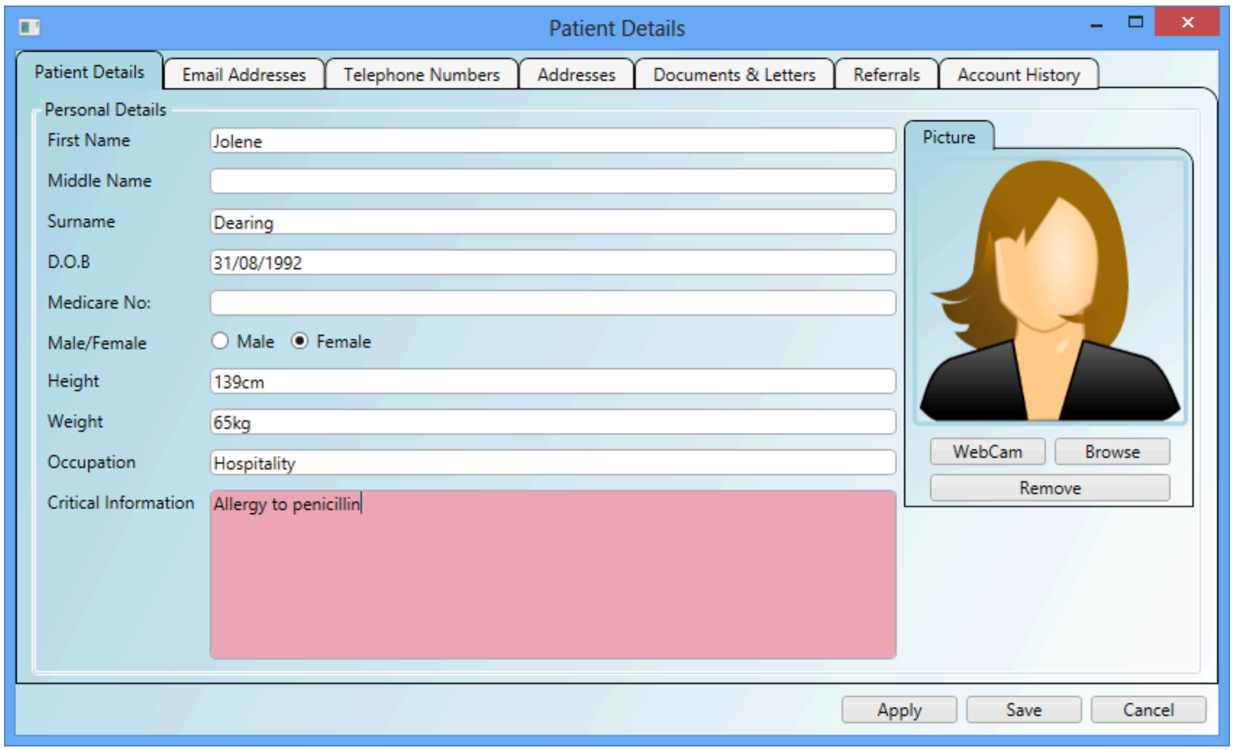
- Cross-validation to choose parameters uses a separate validation set with *noisy* labels
- What about instance-dependent noise?
  - Recent work (Menon et al. ‘18) shows that in general impossible
  - If one makes (reasonable) assumptions about where the noise may be greater, can show that maximizing AUROC with noisy labels is consistent

(Menon, van Rooyen, Natarajan. Learning from binary labels with instance-dependent noise. Machine Learning Journal, 2018)

Outline for today’s class

1. Learning with noisy labels
   - Two consistent estimators for class-conditional noise (Natarajan et al., NeurIPS ‘13)
   - **Application in health care (Halpern et al., JAMIA ‘16)**
2. Learning with right-censored labels

Goal: (continuously predicted) electronic phenotype



Hundreds of relevant clinical variables

Abdominal pain
Active malignancy
Altered mental status
Cardiac etiology
Renal failure
Infection
Urinary tract infection
Shock
Smoker
Pregnant
Lower back pain
Motor Vehicle accident
Psychosis
Anticoagulated
Type II diabetes
...

Simplest approach: rules

- We would like to estimate, for every patient, which clinical tags apply to them
- Common practice is to derive manual rules:

Need to include:
nursing facility
nursing care
facility nursing / rehab
nsg facility
nsg faclty
…

Nursing home?

physician response (gold standard)

text contains: *"nursing home"*	T	F	
T	297	129	PPV 0.70
F	1,319	34511	
	Sensitivity 0.18		

Slow, expensive, poor sensitivity.

Often we can find noisy labels WITHIN the data!

Phenotype	Example of noisy label (anchor)
Diabetic (type I)	gsn:016313 (insulin) in **Medications**
Strep Throat	Positive strep test in **Lab results**
Nursing home	“from nursing home” in **Text**
Pneumonia	“pna” in **Text**
Stroke	ICD9 434.91 in **Billing codes**

How can we use these for machine learning?

Learning with anchors

***Y* is the true label**
***A* is the anchor variable ⚓**
***X* is all features except for the anchor**

- Formal condition:

Conditional Independence

$$A \perp X | Y$$

- Using this, we can do a reduction to learning with noisy labels, thinking of A as the noisy label
- *We may need to modify feature set to (more closely) satisfy this property*

[Halpern, Horng, Choi, Sontag, AMIA ’14; Halpern, Horng, Choi, Sontag, JAMIA ‘16]

Anchor & Learn Algorithm

(special cased for anchors being positive only)

Training

1. Treat the anchors as "true" labels
2. Learn a classifier to predict whether the ***anchor*** appears based on ***all other features***
3. Calibration step: $\frac{1}{|\mathcal{P}|}\sum_{\mathcal{P}} P(A|X)$ P = data points with A=1

Test time

1. If the anchor is present: Predict 1
2. Else: Predict using the learned classifier (with calibration)

Evaluating phenotypes

- Derived anchors and learned phenotypes using 270,000 patients' medical records

History

Alcoholism
Anticoagulated
Asthma/COPD
Cancer
Congestive heart failure
Diabetes
HIV+
Immunosuppressed
Liver malfunction

Acute

Abdominal pain
Allergic reaction
Ankle fracture
Back pain
Bicycle accident
Cardiac etiology
Cellulitis
Chest pain
Cholecystitis
Cerebrovascular accident
Deep vein thrombosis
Employee exposure
Epistaxis
Gastroenteritis
Gastrointestinal bleed
Geriatric fall
Headache
Hematuria
Intracerebral hemorrhage
Infection
Kidney stone
Laceration
Motor vehicle accident
Pancreatitis
Pneumonia
Psych
Obstruction
Septic shock
Severe sepsis
Sexual assault
Suicidal ideation
Syncope
Urinary tract infection

[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

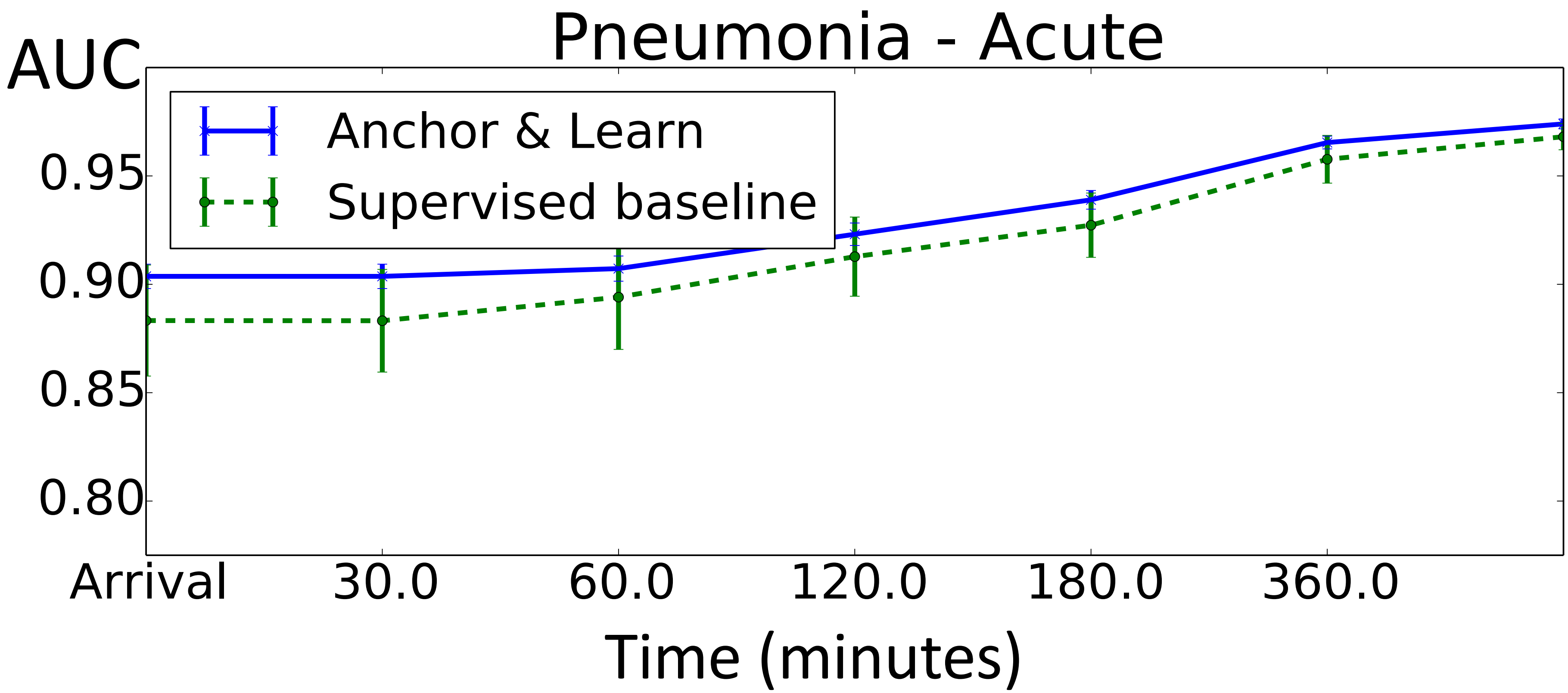
Evaluating phenotypes

- Derived anchors and learned phenotypes using 270,000 patients' medical records
- To obtain ground truth, added a small number of questions to patient discharge procedure, rotated randomly

Deployed in BIDMC Emergency Department

[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

Evaluating phenotypes



Comparison to supervised learning using labels for 5000 patients

Evaluating phenotypes – example model (cardiac etiology)

Anchors

ICD9 codes
410.* acute MI
411.* other acute …
413.* angina pectoris
785.51 card. shock

Pyxis
coron. vasodilators
loop diuretic

Highly weighted terms

Ages
age=80-90
age=70-80
age=90+

Medications
lasix
furosemide

Sex=M

Pyxis
aspirin
clopidogrel
Heparin Sodium
Metoprolol Tartrate
Morphine Sulfate
Integrilin
Labetalol

Unstructured text
nstemi
stemi
ntg
lasix
nitro
cp
chest pain
edema
cmed
chf exacerbation
sob
pedal edema

[Halpern, Horng, Choi, Sontag, AMIA ‘14]
[Halpern, Horng, Choi, Sontag, JAMIA ‘16]

Evaluating phenotypes – example model (cardiac etiology)

Anchors

ICD9 codes
410.* acute MI
411.* other acute …
413.* angina pectoris
785.51 card. shock

Pyxis
coron. vasodilators

cardiac medicine
BIDMC shortform

Highly weighted terms

Ages
age=80-90
age=70-80
age=90+

Medications
lasix
furosemide

Sex=M

Pyxis
aspirin
clopidogrel
Heparin Sodium
Metoprolol Tartrate
Morphine Sulfate
Integrilin
Labetalol

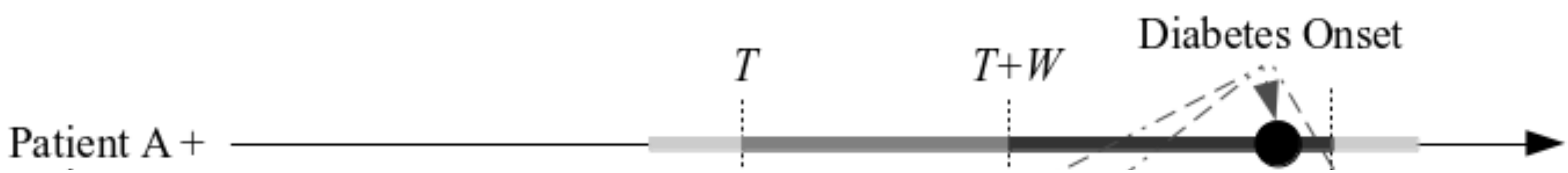
Unstructured text
nstemi
stemi
ntg
lasix
nitro

cp
chest pain
edema
cmed
chf exacerbation
sob
pedal edema

[Halpern, Horng, Choi, Sontag, AMIA '14]
[Halpern, Horng, Choi, Sontag, JAMIA '16]

Outline for today's class

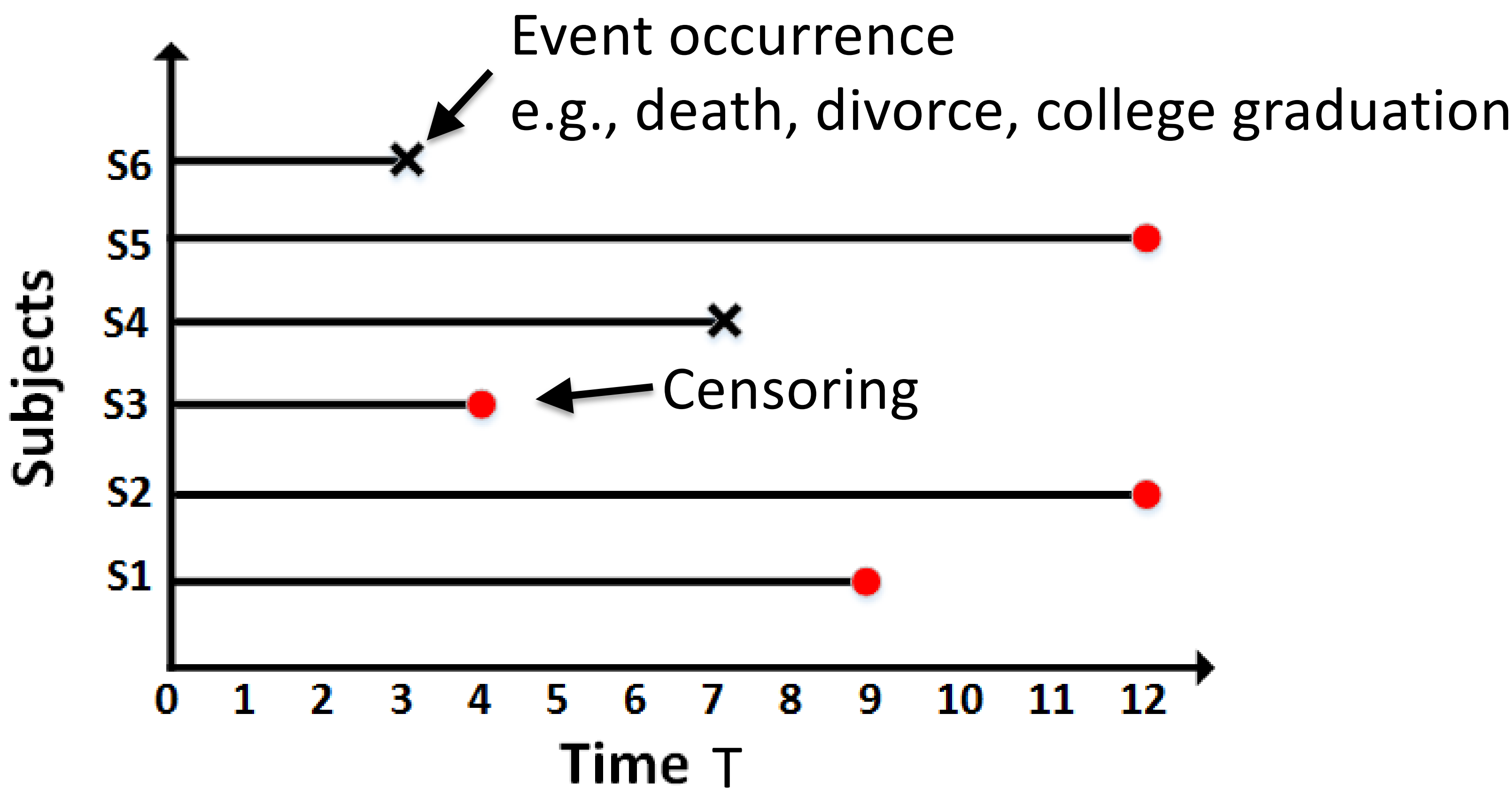
1. Learning with noisy labels
   - Two consistent estimators for class-conditional noise (Natarajan et al., NeurIPS '13)
   - Application in health care (Halpern et al., JAMIA '16)
2. **Learning with right-censored labels**



Instead of reduction to binary classification, let's now predict *when* a patient will develop diabetes

Survival modeling

- How do we learn with right-censored data?



[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

Notation and formalization

- f(t) = P(t) be the probability of death at time t
- Survival function: $S(t) = P(T > t) = \int_t^{\infty} f(x)dx$

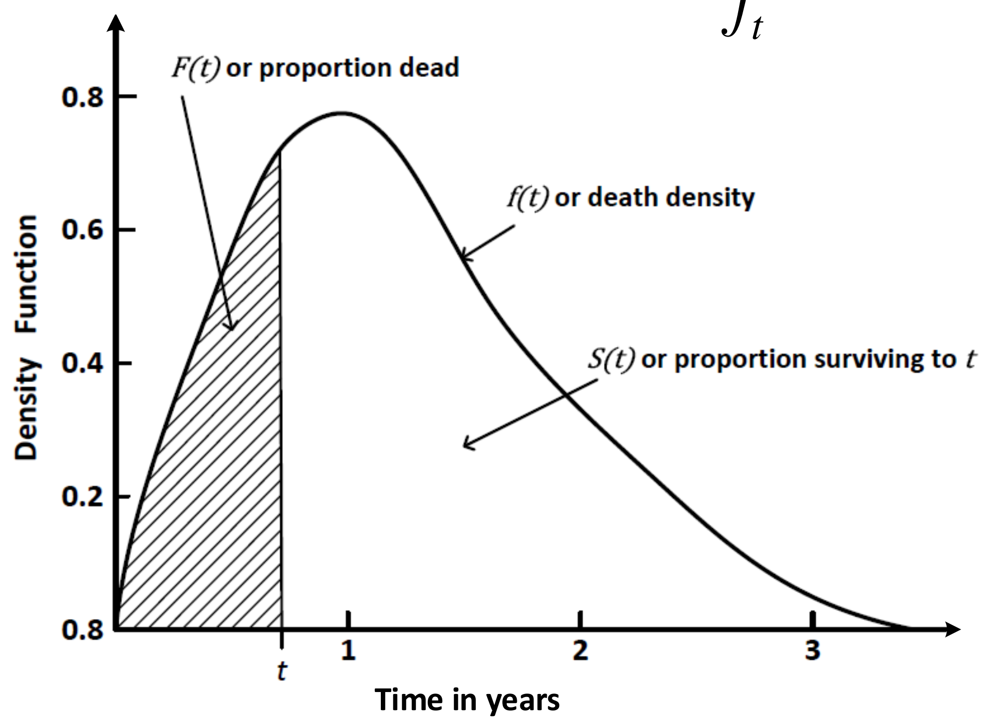


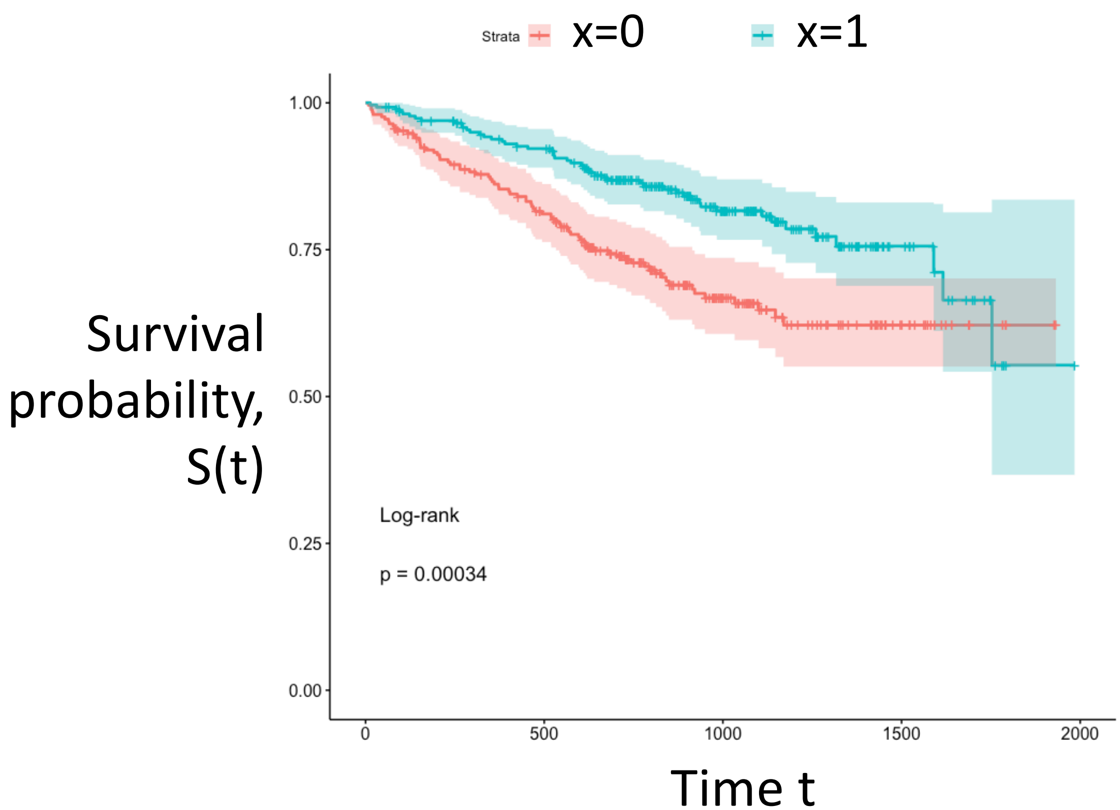
Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

Kaplan-Meier estimator

- Example of a non-parametric method; good for unconditional density estimation



Observed event times

$$y_{(1)} < y_{(2)} < \cdots < y_{(D)}$$

$d_{(k)}$ = # events at this time

$n_{(k)}$ = # of individuals alive and uncensored

$$\widehat{S}_{K-M}(t) = \prod_{k:y_{(k)} \le t} \left\{ 1 - \frac{d_{(k)}}{n_{(k)}} \right\}$$

[Figure credit: Rebecca Peyser]

Maximum likelihood estimation

- Common parametric densities for f(t):

(parameters can be a function of x)

Table 2.1 Useful parametric distributions for survival analysis

Distribution	Survival function $S(t)$	Density function $f(t)$
Exponential ($\lambda > 0$)	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$
Weibull ($\lambda, \phi > 0$)	$\exp(-\lambda t^{\phi})$	$\lambda \phi t^{\phi-1} \exp(-\lambda t^{\phi})$
Log-normal ($\sigma > 0, \mu \in R$)	$1 - \Phi\{(\ln t - \mu)/\sigma\}$	$\varphi\{(\ln t - \mu)/\sigma\}(\sigma t)^{-1}$
Log-logistic ($\lambda > 0, \phi > 0$)	$1/(1 + \lambda t^{\phi})$	$(\lambda \phi t^{\phi-1})/(1 + \lambda t^{\phi})^2$
Gamma ($\lambda, \phi > 0$)	$1 - I(\lambda t, \phi)$	$\{\lambda^{\phi}/\Gamma(\phi)\} t^{\phi-1} \exp(-\lambda t)$
Gompertz ($\lambda, \phi > 0$)	$\exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$	$\lambda e^{\phi t} \exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

Maximum likelihood estimation

- Data are ($\mathbf{x}$, T, b)=(features, time, censoring), where *b=0,1* denotes whether time is of censoring or event occurrence

Maximum likelihood estimation

- Two kinds of observations: censored and uncensored

Uncensored likelihood

$$p_{\boldsymbol{\theta}}(T = t \mid \mathbf{x}) = f(t)$$

Censored likelihood

$$p_{\boldsymbol{\theta}}^{\text{censored}}(t \mid \mathbf{x}) = p_{\boldsymbol{\theta}}(T > t \mid \mathbf{x}) = S(t)$$

- Putting the two together, we get:

$$\sum_{i=1}^{n} b_i \log p_{\boldsymbol{\theta}}^{\text{censored}}(t \mid \mathbf{x}) + (1 - b_i) \log p_{\boldsymbol{\theta}}(t \mid \mathbf{x})$$

Optimize via gradient or stochastic gradient ascent!

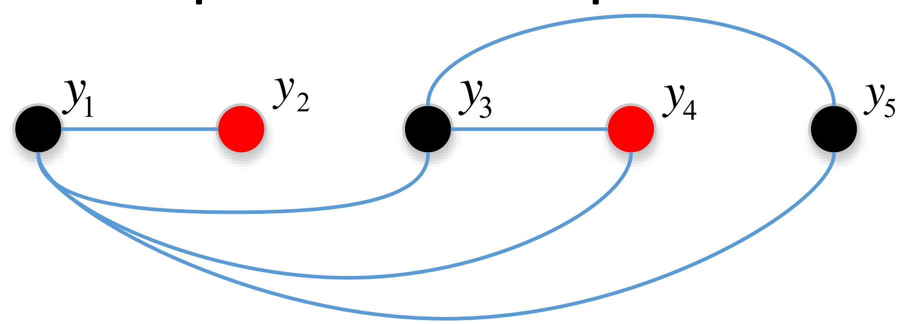
Evaluation for survival modeling

- Concordance-index (also called C-statistic): look at model's ability to predict *relative* survival times:

$$\hat{c} = \frac{1}{num} \sum_{i:b_i=0} \sum_{j:y_i<y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

- Illustration – blue lines denote pairwise comparisons:

Black = uncensored
Red = censored



- Equivalent to AUC for binary variables and no censoring

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

Comments on survival modeling

- Could also evaluate:
  - Mean-squared error for uncensored individuals
  - Held-out (censored) likelihood
  - Derive binary classifier from learned model and check calibration

- Partial likelihood estimators (e.g. for cox-proportional hazards models) can be much more data efficient

Conclusion

- We tackled two challenges that commonly arise in supervised learning in health care
  1. Classification with noisy labels
  2. Regression with censored labels
- Strong assumptions allowed us to develop simple solutions
  - $x \perp \tilde{Y} | Y$ (noise rate constant for all examples)
  - $C \perp T | x$ (censoring time independent of survival time)
- Can we relax these assumptions? Can we do survival modeling with noisy labels?