AI for Medicine Specialization

July 2020

Coursera

deeplearning.ai

Coursera AI for Medicine Specialization: Course #1

AI for Medical Diagnosis

Week 1 - Disease Detection with Computer Vision

Key Challenge #1

Class Imbalance → Use weighted loss

e.g
P1 Normal
P2 Normal
P3 Normal
P4 Mass
P5 Normal
P6 Normal
P7 Mass
P8 Normal

$$L(X,y) = \begin{cases} w_p \times -\log P(Y=1|X) & \text{if } y=1 \\ w_n \times -\log P(Y=0|X) & \text{if } y=0 \end{cases}$$

$w_p = \frac{\#\text{ of negative}}{\#\text{ total}} = \frac{6}{8}$

$w_n = \frac{\#\text{ of positive}}{\#\text{ total}} = \frac{2}{8}$

e.g if loss from each example is 0.3 [$-\log(1-0.5)=0.3$, $-\log 0.5 = 0.3$] if pred prob = 0.5

$\frac{6}{8} \times 0.3 \times 2 = 0.45$ (↑ freq pos)

$\frac{2}{8} \times 0.3 \times 6 = 0.45$ (↑ freq neg)

to have equal contribution to the loss

With weighted loss →

$$loss^{(i)} = loss_{pos}^{(i)} + loss_{neg}^{(i)}$$

$$loss_{pos}^{(i)} = -1 \times weight_{pos}^{(i)} \times y^{(i)} \times \log(\hat{y}^{(i)})$$

$$loss_{neg}^{(i)} = -1 \times weight_{neg}^{(i)} \times (1-y^{(i)}) \times \log(1-\hat{y}^{(i)})$$

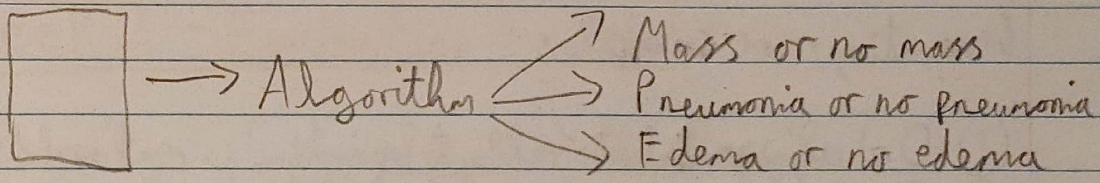
Can resample to achieve balanced classes

e.g
P3 Normal
P6
P1
P8

P7 Mass
P4
P7
P4

Re-sampling methods (undersampling, oversampling)

Key Challenge #2



Multi-task

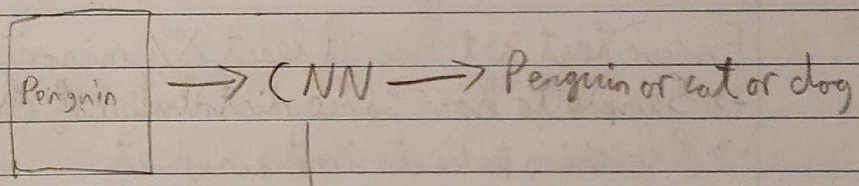
$$L(X, y) = L(X, y_{mass}) + L(X, y_{Pneumonia}) + L(X, y_{edema})$$

Each have their own weighted loss functions

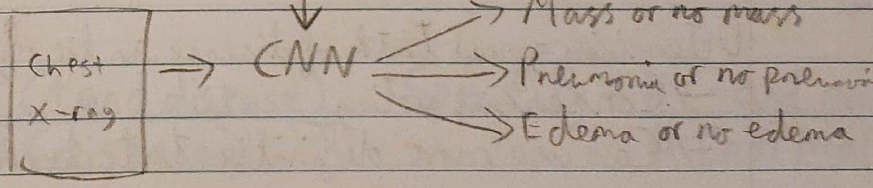
Key Challenge #3

Working with a small training set

1. Pre-training



2. Fine-tuning



Network can learn new task with a better starting point

general features / higher-level features

Fine-tuning — Chest X-ray → Early layers ... Later layers

Option 1	Update	Update
Option 2	— (freeze)	Update

Transfer Learning

Generating more samples

① Do augmentations reflect variations in real world?
② Do augmentations keep the label the same?
e.g augmenting a normal sample can generate an image with a rare disease → Does not preserve

Algo Testing
e.g Rotating, flipping, cropping, color noise.....

Challenge #1
Patient overlap: Make sure patients X-rays only occur in one of the training/validation/test sets, so the model does not memorize rare phenomenas
(Overly optimistic test perf)

Challenge #2
Ensure test set has at least X% minority class e.g 50% so
label
50% 0
50% 1
validation set
→ Sample to have same distribution of classes as test set in
→ Remaining patients in training set

Challenge #3
Determining ground truth of disease in lie of inter-observer disagreement
→ Conservsus voting
→ Can use more definitive tests such as CT scans and Biopsy

Quiz

		P(positive)	
Q)	P1 Normal	0.6	Calculate
	P3 Normal	0.3	loss for
	P5 Mass	0.4	normal

→ P1 $-\ln(1-0.6) = 0.91$
$-\ln(1-0.3) = 0.36$
ln cause natural logarithm

$0.91 + 0.36 = 1.27$ //
ln = $\log_e$

GRAD-CAM: Gradient-weighted class activation mapping
gradient-based
Visual explanations from deep networks via localization

Week 2 - Evaluating models

- $\text{Accuracy} = \frac{\text{Examples correctly classified}}{\text{Total number of examples}}$

$\text{Accuracy} = P(\text{correct} \cap \text{disease}) + P(\text{correct} \cap \text{normal})$

→ Using $P(A \cap B) = P(A|B)\,P(B)$

$\text{Accuracy} = P(\text{correct}|\text{disease})P(\text{disease}) + P(\text{correct}|\text{normal})P(\text{normal})$

(+) Sensitivity TPR — (−) Specifity TNR

→ $\text{Accuracy} = \text{Sensitivity} \times P(\text{disease}) + \text{Specifity} \times P(\text{normal})$

$P(\text{disease})$: Prevalence — proportion that have the disease

$P(\text{normal}) = 1 - P(\text{disease})$

→ $\text{Accuracy} = \text{Sensitivity} \times \text{prevalence} + \text{Specifity} \times (1 - \text{prevalence})$

$\text{Sensitivity} = \frac{\#(+ \text{ and disease})}{\#(\text{disease})}$ $\quad \text{Specifity} = \frac{\#(- \text{ and normal})}{\#(\text{normal})}$

$\text{Prevalence} = \frac{\#(\text{disease})}{\#\text{total}}$

- Positive predictive value (PPV): $P(\text{disease}|+)$
  Negative predictive value (NPV): $P(\text{normal}|-)$

$\text{PPV} = \frac{\#(+ \text{ and disease})}{\#(+)}$ $\quad \text{NPV} = \frac{\#(- \text{ and normal})}{\#(-)}$

*It's specificity not specifity

Confusion Matrix

		Model output: +	Model output: −	
Ground Truth	Disease	TP	FN	→ Sensitivity
	Normal	FP	TN	→ Specifity
		↓ PPV	↓ NPV	

$$PPV = \frac{TP}{TP+FP} \qquad Sensitivity = \frac{TP}{TP+FN}$$

$$NPV = \frac{TN}{TN+FN} \qquad Specifity = \frac{TN}{FP+TN}$$

Extra: $$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1-\text{specifity})(1-\text{prevalence})}$$

Confidence Intervals

Hard to evalnate entire population (p), so we get a sample ($\hat{p}$)

if $\hat{p}=0.80$ for $n=100$

with 95% confidence, p is in the interval $[0.72, 0.88]$

↑ lower bound ↑ higher bound

- Interpretation of 95% confidence: In repeated sampling, this method produces intervals that include the population accuracy in about 95% of samples

if $\hat{p}=0.80$ for $n=500$

with 95% confidence, p is in the interval $[0.76, 0.84]$

- ROC curve is created by plotting TPR against FPR at various threshold settings. The ideal point is at the top left, with a TPR of 1.0 and FPR of 0

Week 3 - MRI Image Segmentation

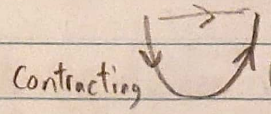
- MRI sequence is a 3D volume as opposed to a single 2D image like X-rays.
  → It's made of multiple sequences and it will consist of multiple 3D volumes.
  → Key idea is to combine the information from different sequences by treating them as different channels.
  → Once we combine, to the machine these are channels stacked in the depth dimension

- Image Registration
  → A pre-processing approach that transforms images so that they are aligned with each other. Eg if a patient moves during a particular scan

- Segmentation
  ① 2D approach: Break up the 3D MRI volume into many 2D slices and pass each slice into the segmentation model to generate a segmentation for each slice. 2D slices then combined to form the 3D output volume.

  ② 3D approach: Break up the 3D MRI volume into many 3D sub volumes

- U-net is a convolutional network architecture for fast and precise segmentation of images



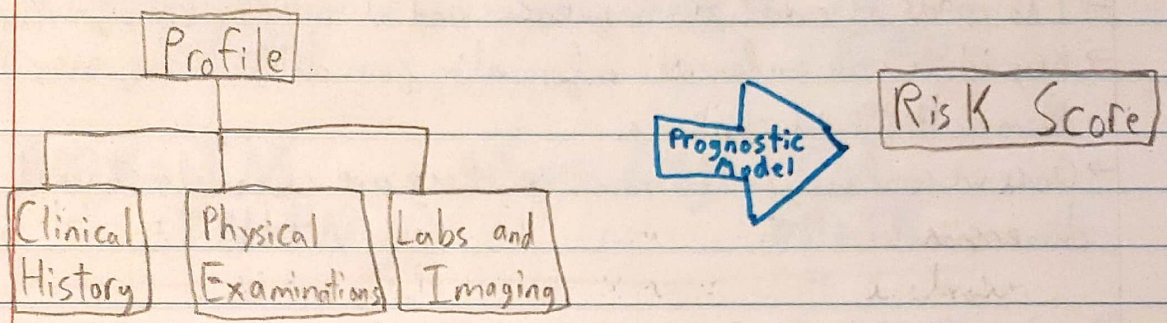
- Soft Dice Loss $L(\text{Prediction, Ground Truth}) = 1 - \frac{2\sum_i^n p_i g_i}{\sum_i^n p_i^2 + \sum_i^n g_i^2}$

Coursera AI for medicine specialization : Course #2

AI for Medical Prognosis

Week 1 - Linear Prognostic Models

- Prognosis is a medical term that refers to predicting the risk of a future event. e.g Death, Heart Attack, Stroke
  → Advantages : Informing patients of risks of illness or survival with illness
  Guiding treatment, who should get drugs, end-of-life care etc



- Concordant Pair • Permissable Pair •

	Patient C	Patient D	
Died within 10 Years?	Yes	No	← Outcomes are different ∴ permissable
Risk Score	0.7	0.6	

↑ Higher Risk Bigger Effect ∴ Concordant

if Risk Scores are equal, then risk ties.

- C-Index

- +1 for a permissible pair that is concordant
- +0.5 for a permissible for risk tie

$$\text{C-Index} = \frac{\#\text{concordant pairs} + 0.5 \times \#\text{risk ties}}{\#\text{permissible pairs}}$$

Interpretation: $P(\text{score}(A) > \text{score}(B) \mid Y_A > Y_B)$

- Random model score = 0.5
- Perfect model score = 1.0

e.g

Patient	Event	Risk
A	Yes	0.8
B	No	0.43
C	Yes	0.62
D	Yes	0.58
E	No	0.62

- Permissible Pairs: AB AE BC BD CE DE
  Concordant Pairs: AB AE BC BD
  Risk Ties: CE

$$C\text{-index} = \frac{4 + 0.5 \times 1}{6} = \frac{3}{4} //$$

- Age, Cholestrol, Blood Pressure, Age_Cholestrol ....

↑ Interaction [Age x Choles]

→ Allows us to have further insights into feature relationships and target, as features are often dependent & not independent

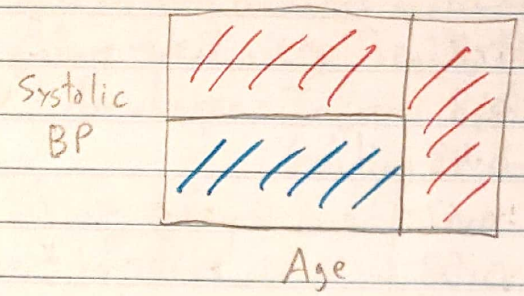
e.g output with interaction : 0.9448
output without interaction : 0.9965

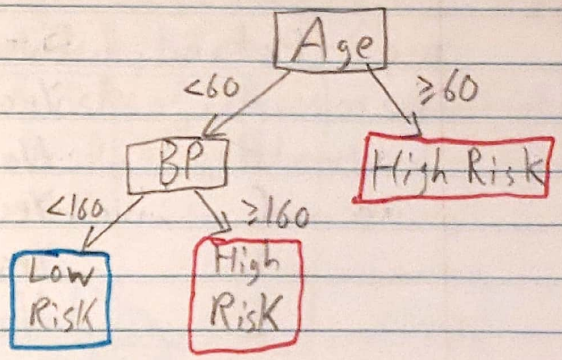
→ Model adjusted for the fact that the effect of high cholestrol becomes less important for older patients compared to younger patients.

- Ln(feature): Common when there's reason to believe that the relationship between the risk and the features is linear in the natural log of features. In addition, it removes skew from data transforming it closer to normal distribution.

Week 2- Prognosis with Tree-based models

- Decision trees divide the input space into regions of high-risk and low-risk using vertical and horizontal boundaries (non-linear association)



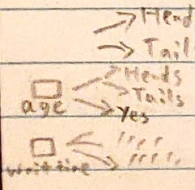


- How to fix overfitting: • Decide on a max depth
  • Random Forest

- Random forests construct multiple decision trees and average their risk predictions. Different patients, features etc
  → Usually higher accuracy than a single tree
  → Also called ensemble learning method

- Eliminating missing data from training and test set can cause a difference in distribution that would prove a model having a low performance on a new test set with no missing value.
  → Graph distribution of old test set before and after dropping missing values

Why is data missing?



1) Missing Completely at Random (Missing not dependent on anything)
2) Missing at Random (Missing dependent only on available information)
3) Missing not at Random (Missingness dependent on unavailable info)

- Imputation replaces missing data with an estimated value based on other available information.

- Mean Imputation: → Split Train/Test
  → Calculate mean for a column of missing values for Train
  → Fill both Train and Test column with that missing value
  * Test set might be small, so its mean value might not be representative ∴ use train mean
  → Mean imputation does not preserve relationships between variables

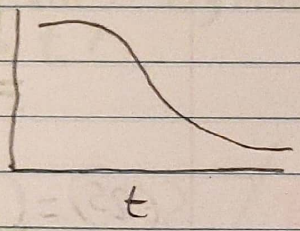
- Regression Imputation: → Preserves linear relationship between variables
  e.g $BP = Coefficient_{age} \times age + offset$ $(y = mx + c)$

Week 3 - Survival Models and Time

- Properties of a survival function

$$S(t) = \begin{cases} 1 \text{ if } t = 0 \\ 0 \text{ if } t = \infty \end{cases}$$

e.g



- Treatment Given — Stroke Event

		T_i	
e.g Sep 2018	Sep 2019	12 months	
Aug 2018	Oct 2019 Study ends	14+ months	[censoring]
Aug 2018	Nov 2018 Withdraws	3+ months	[censoring]

Right censoring: The time to event is only known to exceed a certain value
- → End-of-study censoring
- → Loss-to-follow-up censoring

Estimating Survival Example

i	T_i
1	10
2	8+
3	60
4	20
5	12+
6	30
7	15+

Probability of survival to t months?
t = 25 months

$$S(25) = Pr(T > 25) = \frac{\#\text{ survived to 25 months}}{\#\text{ patients}}$$

if censoring patients
die immediately: $\frac{2}{7} = 0.29$
never die: $\frac{5}{7} = 0.71$

→ We can call and know real value for these patients

$$S(25) = Pr(T \geq 26 \mid T \geq 25)\, Pr(T \geq 25 \mid T \geq 24) \ldots P(T \geq 1 \mid T \geq 0)$$

$$Pr(T > 25 \mid T \geq 25) = 1 - Pr(T = 25 \mid T \geq 25)$$

$$S(25) = (1 - Pr(T = 25 \mid T \geq 25))(1 - Pr(T = 24 \mid T \geq 24)) \ldots (1 - Pr(T = 0 \mid T \geq 0))$$

$$\frac{\#\text{ died at 25}}{\#\text{ Known to survive to 25}} = \frac{0}{2}$$

Post simplification we end up with

$$S(25) = (1 - Pr(T = 20 \mid T \geq 20))(1 - Pr(T = 10 \mid T \geq 10))$$

$$\left(1 - \frac{1}{3}\right) \qquad \left(1 - \frac{1}{6}\right)$$

$$= \frac{2}{3} \times \frac{5}{6} = \frac{10}{18} = \frac{5}{9} = 0.56 //$$

Expression Generalization $S(t) = \prod_{i=0}^{t} 1 - \Pr(T=i \mid T \geq i)$

Kaplan Meier Estimate Model

$$\frac{\#\text{died at } i}{\#\text{Known to survive to } i}$$

$$S(t) = \prod_{i=0}^{t} 1 - \frac{d_i}{n_i}$$

Week 4 - Build a risk model using linear and tree-based models

From survival to hazard

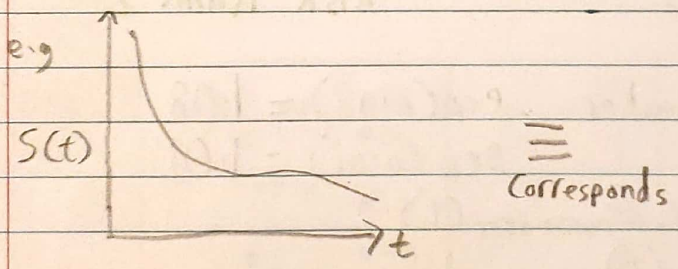
$S(t) = \Pr(T > t)$

Probability of survival past any time t?

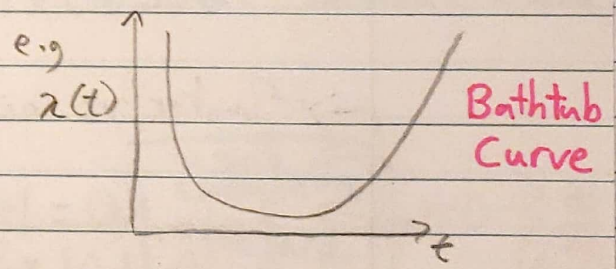
Hazard

$\lambda(t) = \Pr(T = t \mid T \geq t)$

What is a patient's immediate risk of death if they make it to time t?

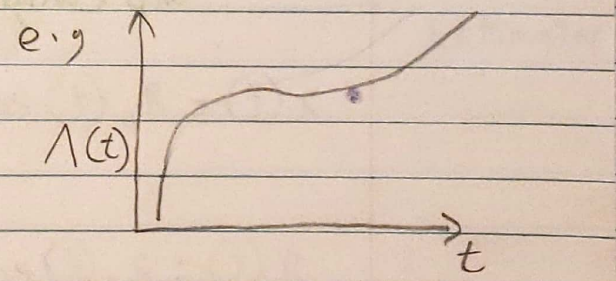


≡ Corresponds



$$\lambda(t) = -\frac{S'(t)}{S(t)}$$

Rate of death if aged t



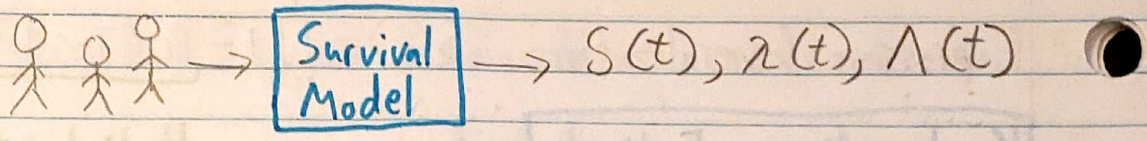
$\Lambda(t)$ → Accumulated Hazard

What's the patient's accumulated hazard upto time t?

$\Lambda(3) = \lambda(0) + \lambda(1) + \lambda(2) + \lambda(3)$

$\Lambda(t) = \sum_{i=0}^{t} \lambda(i)$ $t = 0, 1, 2, 3$ $t = [0 \ldots 3]$

$\Lambda(t) = \int_0^t \lambda(u)\, du$



Individualized Predictions $\lambda_{individual}(t) = \underbrace{\lambda_0}_{\text{Baseline Hazard}}(t) \times factor$

Cox Proportional Hazard Model

e.g $\lambda(t) = \lambda_0(t) \exp(0.08 \times \text{is\_smoker} + 0.01 \times \text{age})$

50 years Smoker $\lambda_1(t)$ 30 years non-smoker $\lambda_2(t)$

$\lambda_1(t) = \lambda_0(t) \exp(0.08 \times 1 + 0.01 \times 50)$
$= \lambda_0(t) \exp(0.58)$
$= \lambda_0(t) \times 1.79$ Risk Rank 1

$\lambda_2(t) = \lambda_0(t) \exp(0.08 \times 0 + 0.01 \times 30)$
$= \lambda_0(t) \exp(0.3)$
$= \lambda_0(t) \times 1.35$ Risk Rank 2

→ Smoker vs non-smoker $\exp(0.08) = 1.08$
→ Age $\exp(0.01) = 1.01$
→ $\lambda_{smoker}(t) = 1.08 \times \lambda_{nonsmoker}(t)$
→ $\lambda_{51}(t) = 1.01 \times \lambda_{50}(t)$

exp (weight) risk increase for factor until increase in variable

- $\lambda(t) = \lambda_0(t) \exp(\underbrace{0.08}_{B_1} \times \underbrace{\text{is\_smoker}}_{x_1} + \underbrace{0.01}_{B_2} \times \underbrace{\text{age}}_{x_2})$

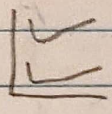
$\lambda(t) = \lambda_0(t) \exp(B_1 x_1 + B_2 x_2 + \dots)$
$= \lambda_0(t) \exp(\sum_i B_i x_i)$

x_i B_i $\exp(B_i)$ → factor risk increase

$x_i \to x_i + 1$

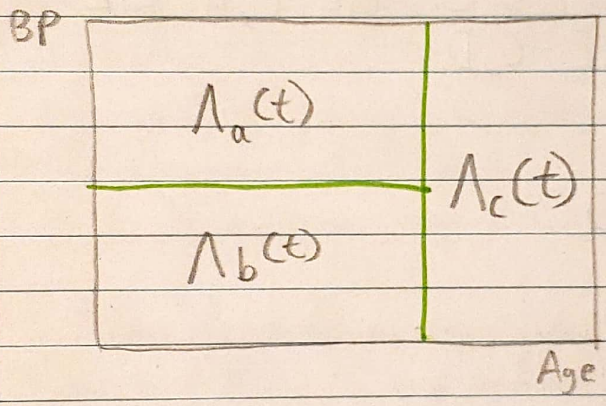
if $\exp(B_i) > 1$ increase risk
$\exp(B_i) < 1$ decrease risk

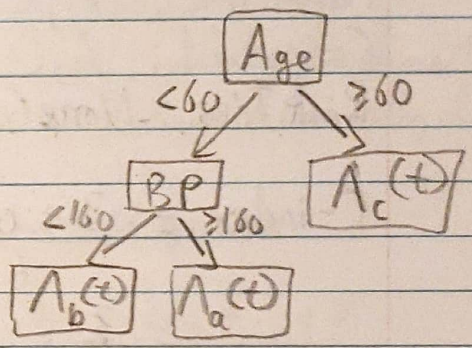
Disadvantages:
- Can't model non-linear relationships
- Hazard function between 2 patients always proportional to each other



→ We need hazard functions that look different for different types of ppl.

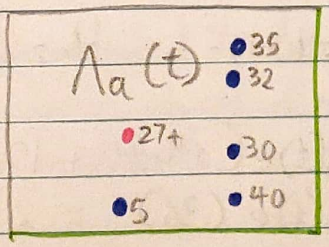
Survival Tree





Time to Event Survival Data

e.g



i	T_i
1	35
2	32
3	27+
4	30
5	40
6	5

→ Cumulative Hazard Estimation → $H(t) = \sum_{i=0}^{t} \frac{d_i}{n_i}$

Nelson-Aalen Estimator

e.g $t=33$ $\quad H(33) = \sum_{i=0}^{33} \frac{d_i}{n_i}$

$$= \frac{d_5}{n_5} + \frac{d_{30}}{n_{30}} + \frac{d_{32}}{n_{32}}$$

$$= \frac{1}{6} + \frac{1}{4} + \frac{1}{3}$$

$$= 0.75$$

Example of Harrel's C-Index

Patient	T	Risk
A	15+	0·65
B	20	1·73
C	5	0·70
D	5+	0·54
E	10	0·83

Permissible : A,C A,E B,C B,E C,D C,E
Concordant : A,C A,E C,D
Risk Ties : None

$$C\text{-index} = \frac{3 + 0.5 \times 0}{6}$$
$$= \frac{3}{6}$$
$$= 0.5$$

Quiz $\beta_{age} = 0.9$ $\beta_{smoker} = 10.0$ Hazard ratio? Cox Model

40 year old non-smoker $\lambda_1(t) = \lambda_0(t) e^{(0.9 \times 40 + 10 \times 0)}$
$= \lambda_0(t) e^{(36)} = \lambda_0(t)\, 4.31123e{+}15$

30 year old smoker $\lambda_2(t) = \lambda_0(t) e^{(0.9 \times 30 + 10 \times 1)}$
$= \lambda_0(t) e^{(37)} = \lambda_0(t)\, 1.1719e{+}16$

Coursera AI for Medicine Specialization : Course #3

AI for Medical Treatment

Week 1 - Treatment Effect Estimation

- e.g Group gets treatment : 2% get a heart attack, 0.02 absolute risk (AR)
  Group is control/placebo: 5% get a heart attack, 0.05 absolute risk (AR)
  In 1 year
  = 0.03 absolute risk reduction (ARR)

- But how do we select assignment to groups?
  Selection bias will result in biased results. e.g treatment group are mostly young & healthy, where control group are mostly old & sick
  → Randomized control trial! (RCT)
  e.g equal mean age & BP between both groups

- But how do we know significance of effect/treatment?
  → Pair a ARR with a p value. The more patients in the trials, the lower the p value, and the more statistically significant is the ARR.

- Number needed to treat (NNT): $1/0.03 = 33.3$
  → Number of people who need to receive the treatment in order to benefit one of them

- Causal inference

Neyman-Rubin Causal model

i	$Y_i(1)$	$Y_i(0)$	$Y_i(1) - Y_i(0)$
Unit	Outcome w/ Treatment	Outcome w/out Treatment	Effect
1	0	1	Benefit -1
2	1	1	No effect 0
3	0	0	No effect 0
4	1	0	Harm 1

Average Treatment Effect (ATE): Mean value of Effect column

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

↑ Expectation

- In reality, the challenge is that we don't get to observe what happens to a patient with treatment and without treatment. This is the fundamental problem of causal inference.

- So, how do we actually estimate the average treatment effect which is given by the previous expression when we don't know any one of the individual differences in outcome?
  - Can also match those from each group with similar
  - → Randomized control trials. Group those who took the treatment and group those that don't take the treatment, and compute mean, then take difference.

Individualized Treatment Effect

ARR = -ve ATE

- These functions will be learned by models called base learners.
  e.g Two-tree method, T-learner
  Single-tree method, S-learner

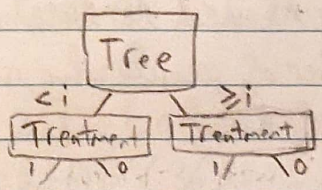
Two-tree method

$\hat{\mu}_1(x) - \hat{\mu}_0(x)$ = treatment

Tree (treatment) Tree (control)

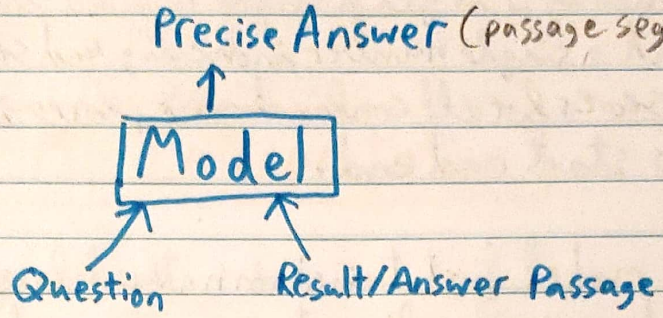
Single-tree method

$\hat{\mu}(x, w)$



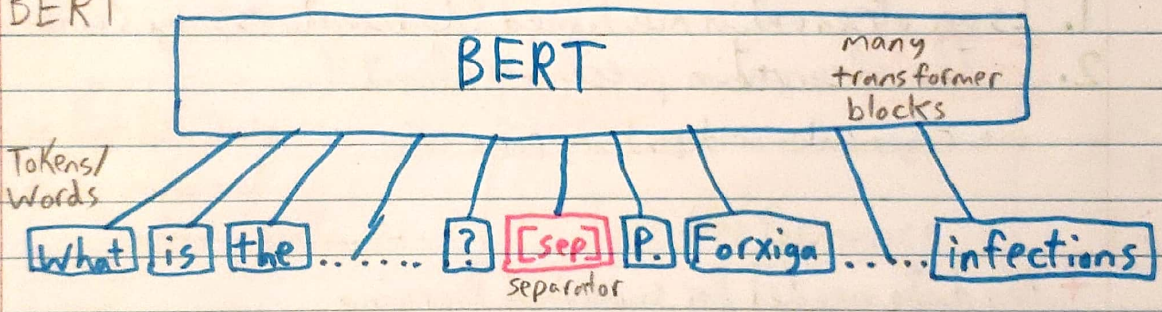
Week 2 - Medical Question Answering

- Important task in NLP systems, to output the answer to a question



Model: ELMo, BERT, XLNet.....
↑ our focus this week

- BERT



Q: "What is the drug Forxiga used for?"

P: "Forxiga is a tablet medication ..... treat infections"

→ We get word representations at the end. Close meaning words have closer 768 dimension values than further words. t-SNE can be used to reduce to 2 dimensions & visualize

- Handling words with multiple meanings
  - Non-Contextualized Word Representations: Word2Vec, Glove
  - Contextualized Word Representations: ELMo, BERT

→ BERT does so by [MASK] a random word & predicting what it might be by the other word representations.
→ BioBERT trained in the context of Medicine

Defining the answer in a text

- There is a start and end token to figure out answer from passage
- A start and end token vector is calculated from each word representation & multiplied by S to get a single number indicating start score & E to get a single number indicating end score for each word in passage
- Compute scores for all combinations & place in grid. Highest value indicates start and end.

→ Usually model first trained on natural language questions from general domain then fine-tune further on biomedical datasets.

Automatic label extraction for medical imaging (unsupervised learning) from radiology reports

1. Is an observation mentioned? (Observation synonyms condensed ← or Is-A relationships & their synonyms)
2. Is the observation present or absent?

↓ via regex rules or dependency parse rules

via Radiologist or terminology such as SNOMED CT ←

\+ No data needed for supervised learning
− Lots of manual work to refine rules and test + requires expertise

Evaluate automatic labelling

- via Human annotation from radiology report
- via Human annotation from Image

Precision: Among the positive labels, what is the fraction of positive ground truths

Recall: Among the positive ground truths, what is the fraction of positive labels

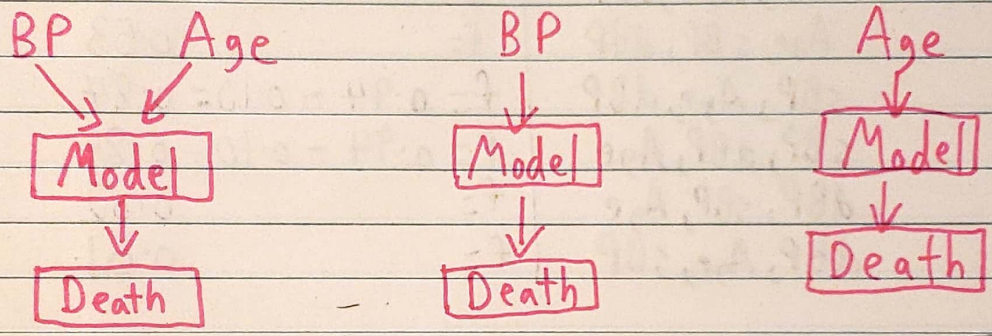
(PPV) **Precision:** $\frac{TP}{TP+FP}$ **Recall:** $\frac{TP}{TP+FN}$ (Sensitivity)

Harmonic mean of precision & recall **F1-score** = $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (Dice Geff score)

- Evaluating on multiple disease categories
  1) Compute for each class & average → Macro-Average
  2) Compute globally → Micro-Average

Week 3 - ML Interpretation

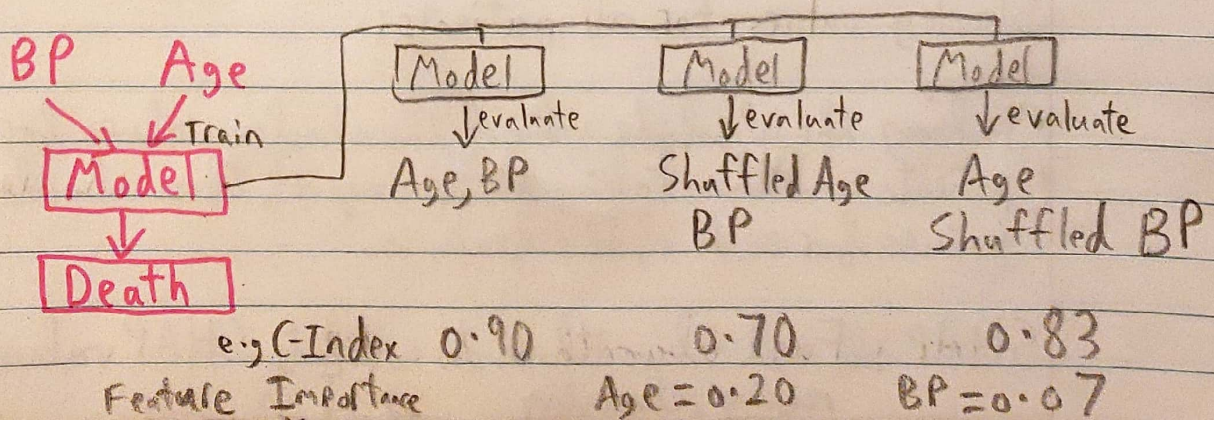
- Drop-Column method



Model Features	Test Performance	Feature Importance
{BP, Age}	0.90	
{Age}	0.85	0.08
{BP}	0.82	0.05

- We need to build as many extra models as features, which can get computationally expensive

- Permutation method



- Individual feature importance
Shapley Values:

f = 0.95	Age, dBP, sBP vs dBP, Age	f = 0.94
f = 0.98	dBP, sBP vs dBP	f = 0.96
f = 0.93	Age, sBP vs Age	f = 0.40
f = 0.94	sBP vs {}	f = 0.10

→ 3! = 6 way the feature set can be formed

Age, dBP, sBP	f = 0.01
Age, sBP, dBP	f = 0.53
sBP, Age, dBP	f = 0.94 − 0.10 = 0.84
sBP, dBP, Age	f = 0.94 − 0.10 = 0.84
dBP, sBP, Age	f = 0.02
dBP, Age, sBP	f = 0.01

(This is all computed for a single patient) Avg = 0.38

Importance for Patient A	Shapley Values (I)
Age	0.10
sBP	0.38
dBP	0.37

$f(\{Age, dBP, sBP\}) - f(\{\}) = 0.95 - 0.10 = 0.85$

↑ (0.10 + 0.38 + 0.37)

- Shapley values for all patients:

- Taking the absolute value of the shapley values for all patients and then taking the average of those shapley values for each of the features.

SHAP: Fast algorithms to compute SHAP values, don't require retraining