AI for Medicine Notes (Based on the Coursera Specialization)

Created by Aldo Zaimi

April 21, 2020

- **The three main challenges of AI in medicine during training**
  - **Class imbalance:**
    - One of the classes (typically the pathology) may be underrepresented in the dataset (minority class).
    - Consequences: misclassification of the minority class will have less contribution to the loss function (i.e. the algorithm will prioritize the majority class).
    - Solutions: (i) Use weighted loss function (use the frequencies of the classes to weight the loss function so that all classes will have the same contribution on the loss function), (ii) Use sampling (use undersampling for the majority class and oversampling for the minority class when preparing your train/val/test sets).
  - **Multi-task applications:**
    - Many diagnosis applications require identification of more than one disease/condition (i.e. we need models that can perform multi-label classification).
    - Solution: Use multi-label loss function (sum of individual losses when considering each label as a binary cross entropy loss).
  - **Small dataset size:**
    - Most medical datasets are not large enough and there is a high risk of overfitting when training complex models.
    - Solutions: (i) Use transfer learning to finetune a pre-trained model on your task (you can update all layers if your dataset is relatively large, or only update the last layers is your dataset is small), (ii) Use data augmentation: (a) pick transformations that reflect variations in the real world (e.g. small contrast changes in x-ray images), (b) pick transformations that preserve the label (e.g. horizontal flipping on a chest x-ray is not adequate as it changes the heart orientation/position and may be identified as an anomaly).
- **The three main challenges of AI in medicine during testing**
  - **Patient overlap:**
    - Having the same patient (even if different samples) in different sets can add bias to the evaluation of the model performance (i.e. same patient on both train and test sets).
    - Solution: split by patient and not by image (i.e. all samples of the same patient should be on the same set).

  - **Set sampling:**
    * Random sampling between train/dev/test sets can lead to absence of the minority class in one of the sets.
    * Solution: minority class sampling (first sample the test set by fixing a minimum percentage (e.g. 50%) of minority class samples that should appear, then do the same for the validation/dev set, and then put all remaining samples on the train set).
  - **Ground truth:**
    * It can sometimes be difficult to obtain valid ground truth labels for medical data
    * Solutions: (i) consensus voting (i.e. the final ground truth label could be a majority vote by a group of experts), (ii) more definitive test (i.e. validation of the diagnosis with another modality/technique)

- **Evaluation metrics for medical applications**
  - **Confusion matrix:**
    * True positives (TP): Sick patients declared positive.
    * True negatives (TN): Healthy patients declared negative.
    * False positives (FP): Healthy patients declared positive (false alarm, type I error).
    * False negatives (FN): Sick patients declared negative (miss, type II error).
  - **Sensitivity:**
    * Also called recall or true positive rate.
    * Proportion of sick patients that are declared positive.
    * Sensitivity = $TP/(TP+FN)$.
  - **Specificity:**
    * Also called true negative rate.
    * Proportion of healthy patients that are declared negative.
    * Specificity = $TN/(FP+TN)$.
  - **Accuracy:**
    * Accuracy is not enough for most tasks (many sources of bias).
    * Accuracy = sensitivity $\times$ prevalence + specificity $\times$ (1-prevalence).
    * Accuracy = $(TP+TN)/(TP+TN+FP+FN)$.
  - **Positive predictive value (PPV):**
    * Proportion that has the disease given a positive test result.
    * PPV = $TP/(TP+FP)$.
    * PPV = (sensitivity $\times$ prevalence) / (sensitivity $\times$ prevalence + (1-specificity) $\times$ (1-prevalence)).
  - **Negative predictive value (NPV):**
    * Proportion that is healthy given a negative test result.
    * NPV = $TN/(TN+FN)$.
  - **ROC curve:**
    * Plot of sensitivity vs specificity for different decision thresholds.

    - Useful when deciding the right threshold for a given task depending on the sensitivity and/or specificity needed.
  - **Confidence intervals:**
    - Useful when testing models on a sample from the total population.
    - Metrics typically given with a 95% confidence interval (i.e. with 95% confidence, the metric is in the interval [a,b]).

- **Image segmentation considerations (MRI tumor use case)**
  - **3D MRI segmentation training:**
    - Input: multiple 3D volumes of different MRI sequences.
    - Preprocessing: combine all sequences (may need image registration techniques) to create final 3D volume with several channels (one sequence = one channel).
    - 2D segmentation approach: segment 2D slices of the 3D volume and combine 2D segmentation maps to obtain 3D segmented volume. Easier and less memory saturation (+) but loss of spatial context for segmentation (-).
    - 3D segmentation approach: segmentation of the 3D volume. Needs to be divided into 3D subvolumes of fixed size to avoid memory saturation (-) but better spatiel context for segmentation task (+).
    - Data augmentation: same transformations applied to 3D images and 3D ground truths (labels).
    - Loss function: soft Dice loss.
  - **Model generalization:**
    - Generalization issue: MRI scanners around different countries/centres have different parameters.
    - Internal validation: using test set of population the model was developed on.
    - External validation: using a test set from a new population to determine generalization potential on different data.
    - Another solution: get a small train/validation dataset from the new population and perform fine-tuning.
  - **Measuring patient outcomes (clinical application):**
    - During model evaluation: access to ground truths.
    - In clinical applications: how to determine if our model improves patient health outcomes?
    - Approach 1: decision curve analysis: quantify the net benefit of using the model to guide patient care.
    - Approach 2: randomized controlled trials: comparing patient outcomes for patients on whom the model is applied vs those on whom the model is not applied.

- **Prognosis models**
  - **Prognosis:**
    - Definition: predicting the risk of a future event.
    - Useful for informing patients about (i) risk of illness, (ii) survival with illness.

    - Useful for guiding treatment (e.g. risk of heart attack to decide who should get drugs, 6-month mortality risk to decide who should receive end-of-life care).
    - Prognosis model: model that takes the patient profile as input (i.e. clinical history, physical examination, labs and imaging) and gives a risk score as output (a probability or an arbitrary number).
    - Prognosis models have (i) profile features with values and weights (i.e. importance of each feature for the model) for each, (ii) can include interaction terms to capture dependence between variables and (iii) the final risk score is typically the sum of the weighted values of the features.
  - **Examples of prognosis models:**
    - Atrial fibrillation: chads vasc score that uses input features (congestive heart failure, hypertension, age 75 or older, diabetes, stroke/TIA/TE, vascular disease, age 65 to 74, sex category) to predict the one year risk of stroke.
    - Liver disease mortality: MELD score that uses input features (ln(creatinine), ln(bilirubin), ln(INR), intercept (fixed to 1)) to compare between patients on liver transplant waiting list and determine who needs it more.
    - Risk of heart disease: ASCVD risk estimator plus that uses input features (ln(age), ln(total cholesterol), ln(HDL-C), current smoker (0/1), diabetes (0/1)) to output a probability of getting heart disease after 10 years.
  - **Evaluation of prognosis models:**
    - Concordant pairs: patient with worst outcome has the highest risk score.
    - Risk ties: patients with same risk score, but different outcomes.
    - Patient pairs that have the same risk score and the same outcome are not considered in the evaluation (called non permissive pairs).
    - C-index: (number of concordant pairs + 0.5 $\times$ number of risk ties) / number of permissive pairs.
    - Interpretation of C-index as a probability between patients A and B with scores S and outcomes Y: $P(S(A) > S(B)|Y(A) > Y(B))$.
    - Perfect model C-index = 1.0.
    - Random model C-index = 0.5.

- **Prognosis with tree-based models**
  - **Decision trees:**
    - Use case: given age and systolic blood pressure, predict 10 year mortality risk.
    - Objective: divide the input space into regions of high-risk and low-risk using vertical and horizontal boundaries.
    - Output binarization: setting a threshold (typically 50%) to determine whether a region will be labelled as low or high risk.
    - How to avoid overfitting in decision trees: (i) monitor train vs test accuracy, (ii) limit the max depth parameter of the tree to avoid having a complex model.
  - **Random forests:**
    - Concept: constructing multiple decision trees using (i) random sampling with replacement and (ii) a subset of features, and averaging their risk predictions.

    * Random forests are an ensemble method and typically perform better than decision trees (with lower risk of overfitting).

- **Missing data**

  - **Identifying missing data:**
    * Dropping samples with missing input variables can lead to model bias and poor performance on new test data that has no missing values.
    * Different distributions before/after dropping missing data: one part of the distribution may be different after dropping missing data (e.g. patients younger than 40 do not get their blood pressure systematically taken during a routine exam).
    * (1) Missing data completely at random: missingness not dependent on anything (i.e. no bias).
    * (2) Missing data at random: missingness only dependent on available information (e.g. patients younger than 40 get their blood pressure randomly checked, but older patients always get their blood pressure checked).
    * (3) Missing data not at random: missingness dependent on unavailable information (e.g. the probability of a patient getting his blood pressure checked depends on the time the doctor has and number of other patients waiting after him).
  - **Handling missing data:**
    * Imputation: replacing missing data by using estimated value based on available information.
    * (1) Mean imputation: (i) average all available values of the missing variable on the training set, (ii) replace missing values with that mean on both the train and test sets.
    * (2) Regression imputation: (i) learn a linear model to estimate the missing values from other available variables on the training set, (ii) use that model to estimate missing values on both the train and test sets.

- **Survival and hazard estimates**

  - **Survival models:**
    * Survival models: focusing on the time to the occurrence of an event (i.e. time from treatment to recurrence, time from diagnosis to death).
    * We are interested in the probability of survival past any time t.
    * Survival function $S(t) = P(T > t)$ (i.e. probability of survival past time t).
    * Survival functions conditions: (1) $S(u) \leq S(v)$ if $u \geq v$, (2) typically $S(t) = 1$ for $t = 0$ and $S(t) = 0$ for $t = \infty$.
  - **Estimating survival with censored data:**
    * For each patient, get time passed from start (e.g. treatment) to event (e.g. stroke).
    * Data censoring happens when (i) the patient does not have the event during the study timeline (end-of-study censoring) or (ii) the patient withdraws from the study (loss-to-follow-up censoring).
    * Right censoring: for patients with data censoring, the event always happens after, if any.

* Chain rule of conditional probability: $P(A, B) = P(A|B)P(B)$ and $P(A, B, C) = P(A|B, C)P(B|C)P(C)$.
* Estimating survival at time t using the Kaplan-Meier estimate:

$$\prod_{i=0}^{t}(1 - P(T = i|T \geq i))$$

where $P(T = i|T \geq i))$ is the ratio between the number of patients who have died at time i and the number of patients known to have survived at time i.

– **Hazard models:**

* Survival function: probability of survival past any time t ($S(t) = P(T > t)$).
* Hazard function: immediate risk of death if the patient makes it to time t ($\lambda(t) = P(T = t|T \geq t)$) (typically bathtub curve).
* From survival function to hazard function: $\lambda(t) = -S'(t)/S(t)$ (i.e. rate of death at time t).
* From hazard function to survival function: $exp(-\int_0^t \lambda(u)du)$.
* Cumulative hazard function: $\Lambda(t) = \int_0^t \lambda(u)du$ (i.e. the accumulated hazard up to time t).
* Customizing risk models to individual patients: $\lambda_i(t) = \lambda(t) \times factor$.