Explanatory AI in Treatment Planning

Data-Driven Automation and Decision Making and Explanatory AI

Tom Purdie, PhD, MCCPM
Princess Margaret Cancer Centre
Disclosures

Method and System for Automated Planning of Radiation Therapy technology patented in PCT/CA2011/001130

Automated Quality Assurance (QA) and Planning technology patented in WO2014197994 A1

Receive royalties from RaySearch Laboratories for license of technology for Automated Breast Treatment Planning and Machine Learning-based Automated Treatment Planning

Have an equity interest in bridge7, licensee of technology for Machine Learning-based for Automated Quality Assurance in Radiation Oncology
Acknowledgements

L. Conroy
C. McIntosh
Outline

AI as a Black Box
How to Achieve Explainability
Similar Patients ➔ Source for Explainability
Atlas Selection Learning
Human | AI Collaboration
Recognizing Digits

Can think of this as a Black Box ... But should we?
Recognizing Digits

Can think of this as a Black Box ... But should we?

Deep Neural Network

Input Layer → Hidden Layers → Output Layer

Zero

Courtesy of Chris McIntosh
Recognizing Digits

Can think of this as a Black Box ... But should we?

Recognizing Patients

Typically patient features are concatenated into a vector
e.g. raw image or image features

Vector is used to quantify the patient

Correspondence between Patient Anatomy

Courtesy of Chris McIntosh
Recognizing Patients

Typically patient features are concatenated into a vector  
  e.g. raw image or image features  
Vector is used to quantify the patient
Recognizing Patients

Typically patient features are concatenated into a vector

e.g. raw image or image features

Vector is used to quantify the patient
How to Achieve Explainability

Intrinsic Interpretable Models
➔ Examines Classifier Structure

Post-hoc Interpretable Models
➔ Examines Feature Prediction Interactions

Data

Interpretable Classifier

Explanations

Prediction

Data

Complex Classifier

Explanatory Models

Prediction

Explanations

Courtesy of Yunsheng Chen
Intrinsic Interpretable Model
Bayesian Network
Joint probability distributions to define probability of one event given set of other known information

Kalet et al., PMB 2015
Post-hoc Interpretable Model

Local Interpretable Model-agnostic Explanations (LIME)

Learned Decision Boundary
Linear Approximation
Data point
Perturbed samples (data point +/- class)

Changes in feature values

Changes in classifier prediction

Features:
- Eccentricity
- Orientation
- Volume
- Thickness
- Shape
- Mean intensity
- Std. intensity
- Mode intensity
- Intensity Histogram
- Max intensity
- Min intensity

Changes in feature values:
- Average Effect in %
- Feature Groups

Data points and perturbed samples are shown around the learned decision boundary, illustrating how changes in feature values and classifier predictions are analyzed with LIME.
Derived Features Can Provide Insight

- **x Distance to Prostate**
- **Distance to Femur**
- **y Distance to External**
Patient Similarity
What does similar mean between patients?
Patient Similarity

What does similar mean between patients?

Courtesy of Chris McIntosh
Learning Similarity

A

B

C

A

B

C
Features for Similarity using Atlas Regression Forests

Image

Leaf Segmentation (Homogeneous Dose)

Encoding (Density Estimation)
Learning Similarity

Bhattacharyya Distance between learned and observed density models at each leaf
Learning Similarity

A

B

C

A

B

C

McIntosh, Purdie IEEE TMI 2016
How many samples do I need in my training dataset?

Accuracy | Number of Training Atlases
Accuracy is the % of voxels with similar radiation dose to ground truth

- Whole Breast
- Breast Cavity (Boost)
- Prostate

Courtesy of Chris McIntosh
Explainable AI for Human | AI Collaboration

Improve Adoption of AI for Clinical Use

Courtesy of Leigh Conroy
Thanks for your attention