Predictive Modeling

Lecture I

Michael C Sachs

Agenda

Learning objectives

1. Understand how predictions differ from associations
2. Understand the term “machine learning” and its role in medical research
3. Describe a predictive model, comparing and contrasting single-variable and multi-variable models
4. Name and recognize common measures of predictive accuracy
5. Understand the difference between in-sample and out-of-sample prediction
6. Be aware of terminology related to prediction models in medical research
7. Identify problems related to overfitting in the development of a predictive model
8. List statistical strategies to avoid overfitting and overoptimisim bias

Resources

See handout for bibliography

Agenda

1. Introduction
2. Prediction models and their role in the clinic
3. The ROC curve and AUC
4. Study designs

Break

5. Predictive signatures
6. Tools for avoiding overfitting and bias
7. Signature development methods
8. Case study

Lunch

- Lab session
- Epilogue

Lab session

- Review paper on prediction model for CORUS
- Analyze and interpret results and conclusions
- Discussion on alternative strategies and limitations based on what we've learned

Introduction

Review

Types of research questions (in statistical terms):

1. Description/quantifying distributions, e.g., what is the distribution of X in the population?
2. Comparing distributions/estimating associations, e.g., what is the relationship between X and Y? Does the relationship between X and Y differ by Z?
3. Clustering of observations
4. Clustering of variables
5. Prediction, e.g., can I predict future/unobserved values of Y using X?

Why prediction?

- Wish to know an unknown or future event
  - Underlying disease state (diagnosis/classification)
  - Future disease outcome (prognosis)
  - Response to treatment
- Form predictions based on observations
  - Medical tests
  - Questionnaires
  - Genetic mutations/expression
  - Predictions from a multivariable model = signature
- Statistical applications
  - Propensity scores, probability of receiving treatment
  - Imputation of missing data
  - Forecasting

Are the predictions “good”?

Answer: It depends

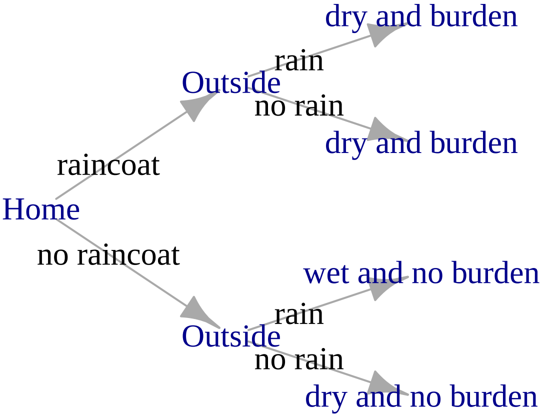
What is the intended use of the predictions?

- Implement a new policy or screening program on a population level
- Guide treatments for individual patients
- Allocate funds for further research and development?

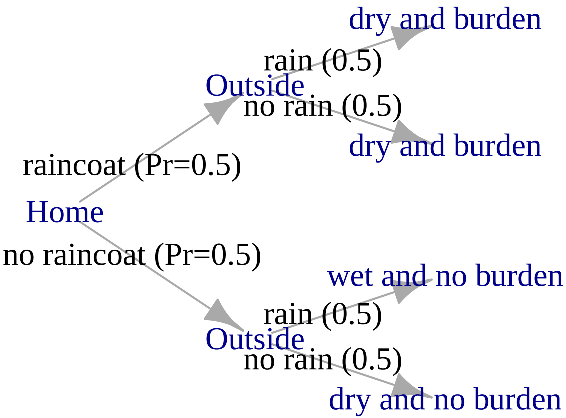
A prediction by itself is not clinically useful unless it leads to an action

The role of predictions in the clinic

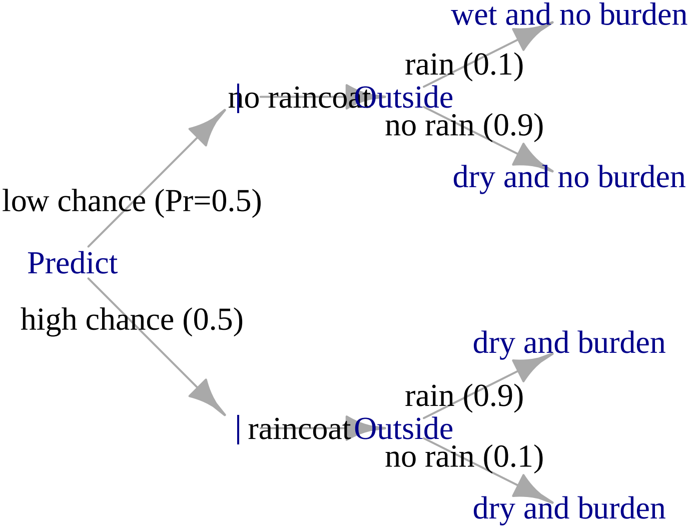
Decision trees



Incorporating probabilities and utilities



Incorporating a prediction



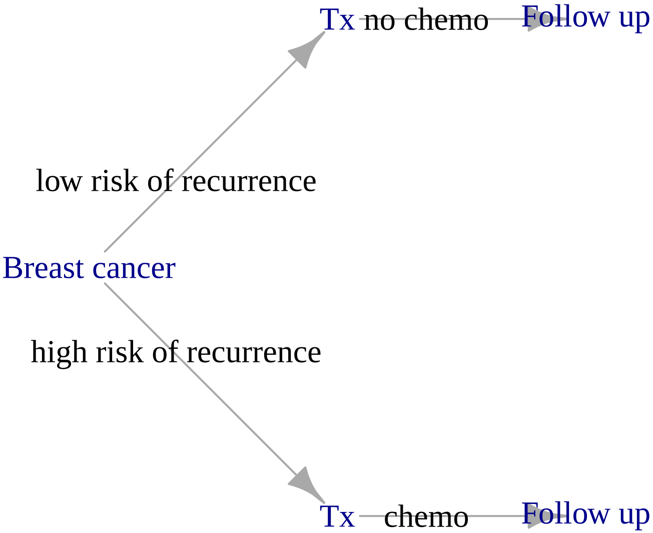
Analysis

Does using the prediction lead to reduced suffering, on average?

- Utility values are personal and not easy to define, and may vary over time
- As with drugs, we don't expect predictions to benefit everyone, every time, but on averge, yield benefit

A biomedical example

Oncotype DX


Tx
no chemo
Follow up
low risk of recurrence
Breast cancer
high risk of recurrence
Tx
chemo
Follow up

Case studies

Oncotype DX. A gene expression signature used to predict breast cancer recurrence.

Corus CAD. A gene expression signature used to predict existing obstructive coronary artery disease.

- What are the predictions used for?
- How was the signature developed?
- What observations go into the signature?
- How useful are the predictions?

Demystifying some jargon

Machine learning, artificial intelligence, deep learning, . . . , oh my

- All of these refer to classes of algorithms that take training data and output models for generating predictions
- They may be distinguised from statistical models because they do not necessarily imply probability models for the data
  - Do not model the data generating process but rather attempts to learn from the dataset at hand
  - Do not yield interpretable estimates of associations (which variables are important)
- The algorithms are generally flexible, allowing interactions and non-linearities

So you want to use machine learning

"I've heard about machine learning and I don't want to miss out, can you help me apply it to my data?"

Wrong!

The goal of medical studies is to produce the evidence that can be used to

- Identify methods to diagnose disease
- Identify risk factors for disease
- Identify treatments for disease
- Identify methods for disease prognosis
- Identify strategies for prevention of disease
- Improve understanding of basic science

Start with a hypothesis!