Predictive Modeling

K Arnold

A Whimsical Take on AI

slides

An Ethics of Artificial Intelligence Curriculum for Middle School Students

Bingo Activity

- In *cohorts* (groups of 4 or 5)...
- *Teams* -> *your cohort's channel* -> *Files* tab -> "Prediction Bingo Activity"

Predictive Modeling

- A powerful tool to turn data into action.
- It works because God made the universe predictable (and successful prediction rewarding)
- **Need for wisdom**: It can be used for great good and great harm

Power of Predictive Modeling

- **Medicine**: wearable monitor for seizures or falls, detect malaria from blood smears, find effective drug regimens from medical records
- **Drug Discovery**: predict the efficacy of a synthesis plan for a drug
- **Precision Agriculture**: predict effect of micro-climate on plant growth
- **Urban Planning**: forecast resource needs, extreme weather risks, ...
- **Government**: classify feedback from constituents
- **Retail**: predict items in a grocery order
- **Recommendation systems**: Amazon, Netflix, YouTube, ...
- **User interfaces**: gesture typing, autocomplete / autocorrect

and so much more...

The universe is surprisingly predictable

- God created the world with actionable structure
  - We gradually learn how to perceive that structure and act within it.
  - The better our perceptions align with how the universe is structured, the better our actions
  - We can discover that structure by learning to be less surprised by what we see (= predicting our perceptions)
- Perceptions are thus both accurate and fallable.

Predictive modeling technology: Need for wisdom

- Potential for great good
- But also great harm:
  - Lack of **fairness** in facial recognition, sentencing, lending, job applicant scoring, ...
  - Lack of **transparency** in how "Big Data" systems make conclusions
  - Lack of **privacy** as data is increasingly collected and aggregated
  - Amplification of extreme positions in social media, YouTube, etc.
  - Oversimplification of human experience
  - Hidden human labor
  - Illusion of objectivity
  - ...!

Stating and refining the question

Six types of questions

1. **Descriptive:** summarize a characteristic of a set of data
2. **Exploratory:** analyze to see if there are patterns, trends, or relationships between variables (hypothesis generating)
3. **Inferential:** analyze patterns, trends, or relationships in representative data from a population
4. **Predictive:** make predictions for individuals or groups of individuals
5. **Causal:** whether changing one factor will change another factor, on average, in a population
6. **Mechanistic:** explore "how" as opposed to whether

Leek, Jeffery T., and Roger D. Peng. "What is the question?." Science 347.6228 (2015): 1314-1315.

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and viral illnesses

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and viral illnesses
3. **Inferential:** examine whether any relationship between dietary factors and viral illnesses found in the sample hold for the population at large

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and viral illnesses
3. **Inferential:** examine whether any relationship between dietary factors and viral illnesses found in the sample hold for the population at large
4. **Predictive:** given a person's demographics and diet, predict the severity of illness

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and viral illnesses
3. **Inferential:** examine whether any relationship between dietary factors and viral illnesses found in the sample hold for the population at large
4. **Predictive:** given a person's demographics and diet, predict the severity of illness
5. **Causal:** whether people who were randomly assigned to eat a diet high in fresh fruits and vegetables or one that was low in fresh fruits and vegetables contract more severe viral illnesses

Ex: Viral illnesses

1. **Descriptive:** severity of viral illnesses in a set of data collected from a group of individuals
2. **Exploratory:** examine relationships between a range of dietary factors and viral illnesses
3. **Inferential:** examine whether any relationship between dietary factors and viral illnesses found in the sample hold for the population at large
4. **Predictive:** given a person's demographics and diet, predict the severity of illness
5. **Causal:** whether people who were randomly assigned to eat a diet high in fresh fruits and vegetables or one that was low in fresh fruits and vegetables contract more severe viral illnesses
6. **Mechanistic:** how a diet high in fresh fruits and vegetables leads to a reduction in the severity of viral illnesses

Our focus: Prediction

- **Why** are we doing this?
  - Predictions enable decision-making
  - trying to predict helps us understand.
- **What** are we doing?
  - Predict something unknown from something known. Specifically: complete-the-table model
  - Mostly we'll *assume independent observations* (e.g., generalize across people)
  - Sometimes we'll *forecast* (e.g., predict how a time series will continue)
- **How**: We'll look at...
  - methods that consider similar examples (Nearest Neighbors)
  - methods that look at overall trends (linear/logistic regression)
  - more advanced methods, time permitting

Example: Home Sale Prices

From Ames, Iowa home sales, 2006-2010. (De Cock, 2011)

Lot_Area	Total_Bsmt_SF	Gr_Liv_Area	Garage_Cars	Sale_Price
31770	1080	1656	2	215000
11622	882	896	1	105000
14267	1329	1329	1	172000
11160	2110	2110	2	244000
13830	928	1629	2	189900

(2930 total rows)

- *Y*: *response* variable (aka *outcome*, *dependent variable*): `Sale_Price`
- *X*: *features* (aka *predictors*, *covariates*, etc.): everything else

Note: X is much easier to measure than Y