Introduction to Statistical Machine Learning
CSC/DSCC 265/465

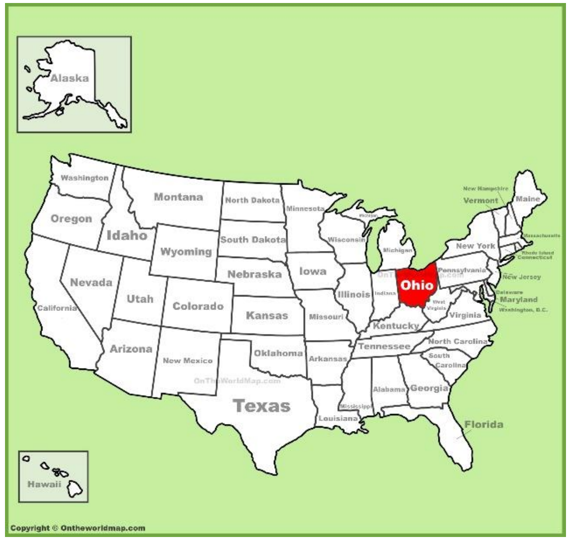
Kaggle Challenge I

Cantay Caliskan

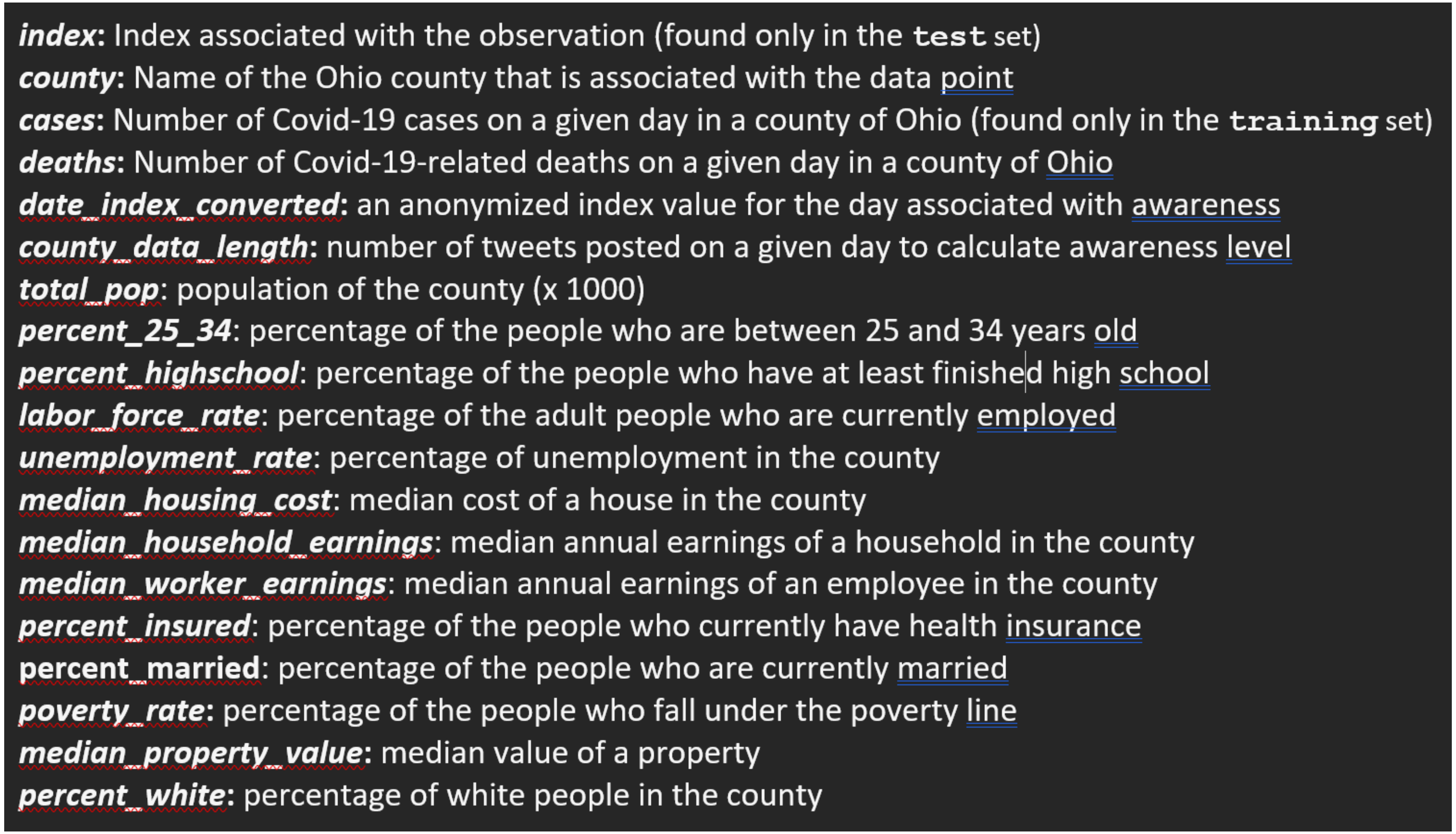
Kaggle Challenge I

Kaggle Challenge I

- A prediction challenge
- You will be asked to predict the number of Covid-19 cases by day and county
- **Input (X)** – all of the variables in the training dataset (and more…)
- **Output (Y)** – number of Covid-19 cases



Information about the dataset



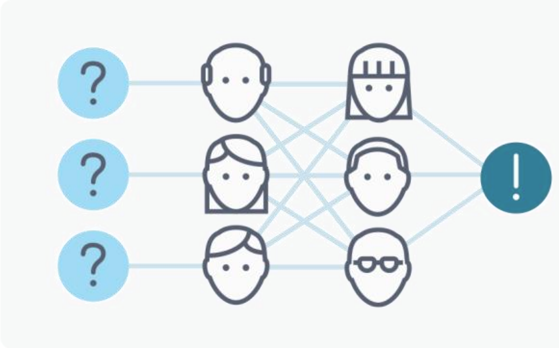
index: Index associated with the observation (found only in the **test** set)
county: Name of the Ohio county that is associated with the data point
cases: Number of Covid-19 cases on a given day in a county of Ohio (found only in the **training** set)
deaths: Number of Covid-19-related deaths on a given day in a county of Ohio
date_index_converted: an anonymized index value for the day associated with awareness
county_data_length: number of tweets posted on a given day to calculate awareness level
total_pop: population of the county (x 1000)
percent_25_34: percentage of the people who are between 25 and 34 years old
percent_highschool: percentage of the people who have at least finished high school
labor_force_rate: percentage of the adult people who are currently employed
unemployment_rate: percentage of unemployment in the county
median_housing_cost: median cost of a house in the county
median_household_earnings: median annual earnings of a household in the county
median_worker_earnings: median annual earnings of an employee in the county
percent_insured: percentage of the people who currently have health insurance
percent_married: percentage of the people who are currently married
poverty_rate: percentage of the people who fall under the poverty line
median_property_value: median value of a property
percent_white: percentage of white people in the county

There are also many 'topic awareness' variables…

Tasks

Tasks

- You will work on three (3) groups of tasks:
  - ***Descriptive Analysis (20 points)***
    - Summary of the Covid-19 experience of Ohio, descriptive visuals/graphs
  - ***Kaggle Competition (50 points)***
    - You will create a model that provides the highest ***R2*** value for predicting correct **'cases'** by using the awareness scores and social/economic variables
  - ***Lab Report (30 points)***
    - Provide your findings in a traditional IEEE format
      - *Abstract*, *Introduction*, Data, *Methods*, *Results*



Prediction: Steps

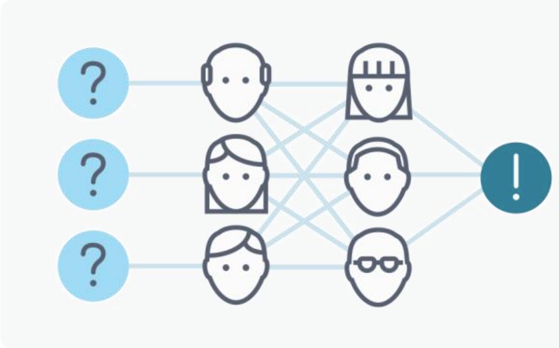
1) Develop a prediction model using the **`training`** dataset

2) Using the model, classify the observations in the **`test`** dataset

3) Use the **`sample submission`** file (a smaller version of the **`test`** dataset) to submit your solutions [solutions submitted according to the Index variable]

4) If not happy with the results, repeat the Steps 1), 2), and 3)

kaggle



UNIVERSITY of ROCHESTER

Online Competition

- Online competition you can enter on ***Kaggle***:
  - https://www.kaggle.com/t/e235def16ffb4591a6e295c769be024b
- Goal: Develop a prediction model that classifies the observations with the highest ***R2*** possible
  - **No model restrictions!**
  - You can:
    - Use any prediction algorithm that you think will give the highest accuracy
    - Perform any type of feature engineering
    - Perform weighting, dimensionality reduction etc.
    - Use any external dataset to enrich your training and test datasets
    - Note: You cannot use external datasets that report #cases or #deaths.

Important:

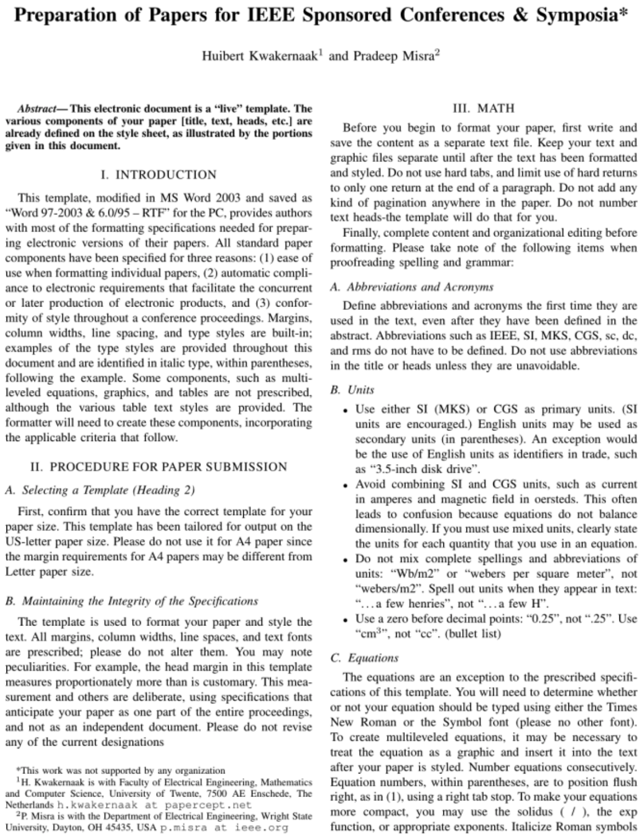
- Use **training_data.csv** to *train your model*
- Use **sample_submission.csv** to submit your answers
- You can send up to 10 submissions every day (competition is currently open!)
- Provide the *R2* score in your report

Online Competition: Further Do’s and Don’ts

- Code:
  - **You cannot post your solution / code online.**
  - You can use `Python` (only)
  - Your code should be ***executable***, i.e.:
    - We should be able to run your code by running the cells ***consecutively***
    - We should also be able to run your code on a ***laptop*** (for instance, a new MacBook Pro) in a reasonable amount of time (in max. a few hours)
    - We should be able to ***understand*** what your code is doing. So, please make sure that:
      - you write **a lot of comments** describing your code
      - you only include the code that works
      - you only include your best solution
      - you name your variables mutually intelligibly (i.e. **case_data**, not **td123** etc.)
- Model:
  - Your classification must give a number as the prediction

Lab Report

- You will be using a ***LATEX*** template (*IEEE*) to produce our reports.
    - Link to the template found in the instructions.



Preparation of Papers for IEEE Sponsored Conferences & Symposia*

Huibert Kwakernaak[1] and Pradeep Misra[2]

Abstract— **This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

I. INTRODUCTION

This template, modified in MS Word 2003 and saved as "Word 97-2003 & 6.0/95 – RTF" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multileveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. PROCEDURE FOR PAPER SUBMISSION

A. Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. Please do not use it for A4 paper since the margin requirements for A4 papers may be different from Letter paper size.

B. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations

III. MATH

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter", not "webers/m2". Spell out units when they appear in text: "…a few henries", not "…a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc". (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols

*This work was not supported by any organization

[1] H. Kwakernaak is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands `h.kwakernaak at papercept.net`

[2] P. Misra is with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA `p.misra at ieee.org`

1) Abstract
2) Introduction
3) Data
4) Methods
5) Results

Lab Report: Do’s and Don’ts

- A few important notes:
  - Think of the report as an ***essay***!
  - The report should have the following sections:
    - *Abstract*, *Introduction*, *Data*, *Methods*, *Results*
  - Try to create a good flow, and a ‘story-like’ report
  - Spend enough time on explaining:
    - Your data
    - Your goals
    - What you did to achieve these goals
    - What you think you could have done to achieve better results
      - Criticize yourself!

Deliverables

Deliverables

- Your code in ***.ipynb*** format
  - Add a lot of comments to your code!

- Your ranking in ***Kaggle*** system

- The lab report in ***IEEE*** format published as a ***.pdf*** file
  - Lab report should include all of the visuals, tables, etc.

- Submit everything through *BlackBoard*

Grading

Kaggle Competition: Grading

- You will be graded based on the following criteria:
  - **Code**
    - Cleanliness/understandability (i), executability (ii), format (iii)
  - **Ranking**
    - Ranking in the **`Kaggle`** competition
  - **Lab Report**
    - *Introduction* (i), *Data* (ii), *Methods* (iii), *Results* (iv)
    - Flow, readability, level of detail, quality of visuals/tables, adherence to the guidelines

More about Grading

- Other important information about Kaggle competition:
  - The lowest grade you can get from the **ranking** component will be **60/100**.
  - The highest ranked project will get **100/100** for the **ranking** component.
  - However:
    - If your accuracy is close to the benchmark reported in the guidelines, your grade may be lower (and it may be zero, as well).

Deadlines

- Please submit your code, solution submission, and report by:
  - **Deadline: Sunday, April 9, 11:59 PM**
  - You can send the code and report up to two days after the deadline (however, a late penalty will be applied)
  - **You *must* send the solution on Kaggle by the deadline (no late submission possible for Kaggle (Part II))**

And one last reminder…

- Let's say you have achieved a really good (or maybe a really bad) R2 and you are done with model training:
  - **Please do not post the solutions online!**
  - **Or, simply said, please do not post any related code online ☺**