About the Course & Intro to Machine Learning

Dr. Jomana Bashatah

About the Course

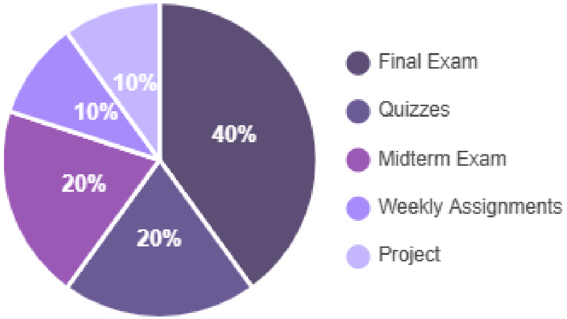
- Machine Learning (ML) is a very hot topic
- Difficulty of the course:
  - For the course:
    - Matrix Algebra is not required and will not be used
    - Calculus is not required and will not be used
    - Programming knowledge is not required. The basic R knowledge you need is introduced in class and in the book.
  - Support
    - Everybody can get a good grade! Do the work and ask me for help if needed (I will spend all the time needed with you)
- Materials
  - We will use the following textbook:
    - Practical Machine Learning with R: Tutorials and Case Studies by Carsten Lange
    - Open Access web version of the book:
      - https://ai.lange-analytics.com/htmlbook/index.html#how-the-book-is-organized

Reward/Return of Good Work Inside and Outside of Class

- Good grade
- You get an understanding of machine learning that enables you:
  - To talk about machine learning algorithms during an interview
  - To use machine learning on your senior project/master thesis
  - To build on the knowledge from the course by reading more advanced articles and watching videos to become an expert.

The Assessment Algorithm

- Learning Quest Progress
  - Quizzes
    - 20%
  - Weekly Assignments
    - 10%
  - Midterm Exam
    - 20%
  - Project
    - 10%
  - Final Exam
    - 40%



Course Assessment Breakdown

My Policies

- • Late submission will incur a 25% penalty
- • Any assignment late more than 2 weeks will receive a zero
- • Submissions with >25% similarity will receive a zero (no re-submissions)
- • Attendance is mandatory

What is What?

- Artificial Intelligence (AI)
- Machine Learning
- Deep Learning
- Big Data

What is What?

- Categorizing AI, Machine Learning, and Deep Learning

- Categorizing AI, Machine Learning, and Deep Learning

What is What?

- Categorizing AI, Machine Learning, and Deep Learning
- Categorizing AI, Machine Learning, and Deep Learning

What is What?

- Categorizing AI, Machine Learning, and Deep Learning

- Categorizing AI, Machine Learning, and Deep Learning

What About Big Data

- Big Data is not a category of learning. It is a category of data!!!
- Two common definitions
  - Laymen: Many records (thousands?, millions?, billions?)
  - Experts: So many records that they do not fit in the memory of one computer.
    - At least billions of records.
    - Requires distributed computing.

Three Applications of Machine Learning

- Regression
- Classification
- Cluster

Three Applications of Machine Learning

- Regression
  - Outcome variable is continuous
  - We try to predict a numerical value
- Classification
- Cluster

Three Applications of Machine Learning

- Regression
- Classification
- Cluster

Three Applications of Machine Learning

- Regression
- Classification
  - Outcome variable is categorial
  - Most of the times 2 categories such as:
    - Yes/No
    - True/False
    - often represented as dummies: 1/0
  - Sometimes more than two catogories (ordered or unordered):
    - good, fair, bad (ordered)
    - red, blue, green (unordered)
    - strongly agree, agree, disagree, strongly disagree (ordered)
- Cluster

Three Applications of Machine Learning

- Regression
- Classification
- Cluster

Three Applications of Machine Learning

- Regression
- Classification
- Cluster
  - Sorting observations into a number of groups based on feature variables.
  - Groups are as homogenous inside as possible.
  - Groups are as diverse between groups (when comparing groups)

Three Applications of Machine Learning

- Regression
- Classification
- Cluster

Terminolgy

- First 3 Observations (records) of the Housing Dataset (to predict house prices)

-
  ```
  library(rio);library(tidyverse)
  DataHousing =
    import("Data/HousingData.csv") %>%
    select(Price=price, Sqft=sqft_living, Bedrooms=bedrooms,Waterfront=waterfront)
  print(DataHousing[1:3,])
  ```

```
   Price Sqft Bedrooms Waterfront
1 221900 1180        3         no
2 538000 2570        3         no
3 180000  770        2         no
```

- Tidy data:
- Observations (synonym: records) are in the rows.
- Variables (synonym: features) are in the columns.
- Variable names (column names) are in the first row.
- Data are in individual cells (and they form vectors; column names can be interpreted as vector names).

Terminolgy

- Main
- First 3 Observations (records) of the Housing Dataset (predict house prices)
-
```
library(rio);library(tidyverse)
DataHousing =
  import("Data/HousingData.csv") %>%
  select(Price=price, Sqft=sqft_living, Bedrooms=bedrooms,Waterfront=waterfront)
print(DataHousing[1:3,])
```
-
```
   Price Sqft Bedrooms Waterfront
1 221900 1180        3         no
2 538000 2570        3         no
3 180000  770        2         no
```
- Outcome Variable: The variables that is the outcome of the prediction ($Price$)
- Predictor Variables: The variables that predict an outcome ($Sqft$, $Bedrooms$, $Waterfront$)
- Example linear regression:
- $Price = \beta_1 \cdot Sqft + \beta_2 \cdot Bedrooms + \beta_3 \cdot Waterfront + \beta_4$
- Synonyms
- Synonyms for Outcome Variable:
  - Response variable (it responds to the predictors)
  - Dependent variable (it is dependent on the predictors)
  - Endogenous variable (it is dependent on the predictors)
- Synonyms for Predictor Variables:
  - Explanatory variables variables (they explain the outcome)
  - Independent variables (they are chosen independently to see how the impact the outcome)
  - Exogenous variable (they are chosen independently to see how the impact the outcome)

Prediction

- Predicting means that we use the values for one or more known variables to estimate an outcome. Predictions can be forecasts or for the same time period.
- Predict tomorrow’s weather based on today’s barometric change of pressure.
- Predict the price of a house (today) based on it’s square footage (today).
- Variables that are based on a prediction are marked with a hat (e.g., $\widehat{Price_i}$.(

Model

- A model is what we use for predicting an outcome variable based on values of predictor variables — given certain assumptions.
- $\widehat{Price}_i = \beta_1 Sqft_i + \beta_2$

Fitted Model

- Can we use the model from the previous slide to predict the price of a house, if we know the value for the house's predictor variable (e.g., $Sft = 1000$)?
- Only if we know the values for the parameters (the $\beta's$)!
- Suppose OLS based on data determines that $\beta_1 = 300$ and $\beta_2 = 500{,}000$:
- $\widehat{Price}_i = 300Sqft_i + 500000$
- A model where the parameters (the $\beta's$) have been determined by a machine learning algorithm is called a fitted model.
- A fitted model can be used for predictions. E.g., a house with a square footage of 1,000 sqft is predicted to cost $8000,000.
- In our case:
  - $\widehat{Price}_i = 300 \cdot 1{,}000 + 500{,}000 = 800{,}000$

Parameters

- The βs of a model are the parameters. The parameters are determined by the optimizer of a machine learning algorithm.
- Machine learning can be (over)simplified to the following steps:
- Determine the model including the βs.
- Use machine learning to determine the βs and therefore create a fitted model.
- Use the fitted model to predict based on predictor variables.

Training vs. Testing Data

- Training Dataset
- When using data to calibrate the parameters minimizing some type of prediction error (training the model), most but not all of the observations are used.
- . . .
- Only about 60% – 90% of the total observations are usually used to calibrate the parameters (the βs of the model). These observations are randomly chosen and the resulting dataset is called the training dataset.

Training vs. Testing Data

- Testing Dataset
- Observations not randomly chosen for training makeup the testing dataset. Testing data are never used to optimize model performance in any way! Instead, they form a hold-out dataset used to assess the predictive quality of a model.
- . . .
- Using the training dataset for this purpose is not an option because we would measure how well the model approximates the training data rather than assessing the predictive quality on new data — data that the model never has seen before

Why Using R for Machine Learning?

- Machine Learning Software
- R (free, advanced, timely delivery of new algorithms, easy to use with the tidyverse and tidymodels packages)
- Phython (free, advanced, often first delivery of new algorithms, not as easy to use because it is a programming language rather then a statistical language)
- SAS (not free, somehow advanced, slow in delivering new algorithms, easy to use)
- Stata (not free, somehow advanced, slow in delivering new algorithms, easy to use)
- SPSS (not free, not advanced, slow in delivering new algorithms, optimized for survey processing, extremely easy to use)

Time to Meet Your ML Partner

- You’ve learned the concepts… now meet the tool!
- Coming in Lecture 2:
  - Installing R & RStudio (your new best friends)
  - Basic R syntax and commands
  - Loading and exploring datasets
  - Creating your first visualizations
- Bring your laptop and let’s start coding!
- The journey of 1,000 algorithms begins with a single model.