

Introduction to Machine Learning

Samuel Carton (adapted from Marek Petrik)

8/2025

What is machine learning?

Arthur Samuel (1959, IBM, computer checkers):

Field of study that gives computers the ability to learn without being explicitly programmed



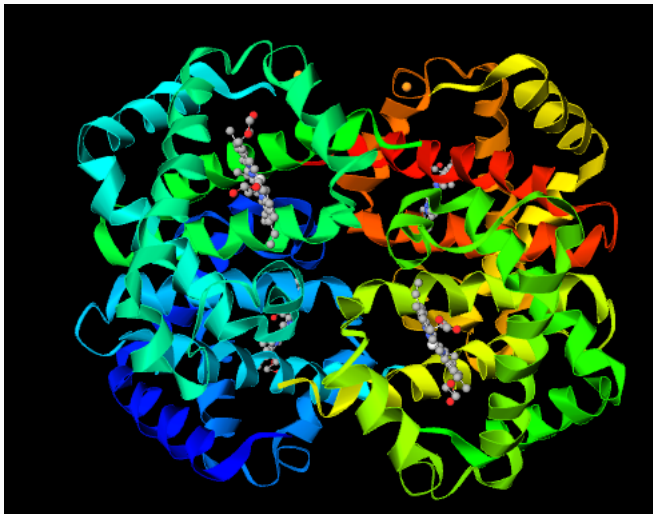
Creative commons license, Source: <http://flickr.com>

IBM Watson: Computers Beat Humans in Jeopardy!

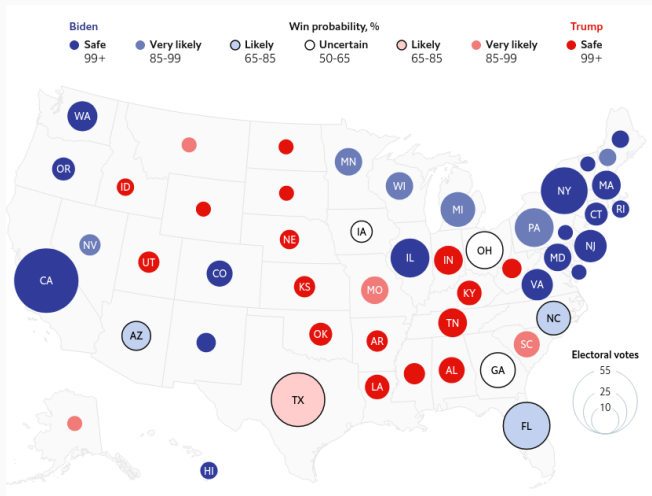


Fair use, <https://en.wikipedia.org/w/index.php?curid=31142331>

AlphaFold: Machine Learning for Protein Folding



Predicting Elections



<https://economist.com>

Personalized Product Recommendations

Online retailers mine purchase history to make recommendations

https://www.amazon.com/Fischer-Hannibal-94-Ski-177cm/dp/B013H2X3G5/ref=as_li_tf_l11447786gq-d-1-1

Share

Fischer Hannibal 94 Ski
Be the first to review this item

Price: **\$649.95** + \$12.95 shipping

Item is eligible. **No interest if paid in full within 12 months** with the Amazon.com Store Card. [Apply now](#)

Note: Not eligible for Amazon Prime.

Only 1 left in stock.
Get it as soon as Friday, Jan. 27 when you choose **Two-Day Shipping** at checkout.

Ships from and sold by [Backcountry](#).

Color: **One Color**

Size: **177cm**

- Length: 170 cm, 177 cm, 184 cm
- Dimensions: 126 / 94 / 112 mm
- Turn Radius: (177 cm) 23 m
- Profile: Tour Rocker (rockerized tip)
- Construction: sandwich (ABS sidewall)







New (1) from \$649.95 + \$12.95 shipping

[Report incorrect product information.](#)

Roll over image to zoom in







Customers Also Shopped For

Page 3 of 4 [Start over](#)

| | | | | | |
|---|---|---|---|---|---|
|  |  |  |  |  |  |
| K2 Mens Wayback 88 Skis 274 \$450.95 | Faction Men's Ski Candidate 1.0 Downhill Skis \$503.20 - \$712.95 | Fischer Hannibal 100 Ski \$699.00 - \$699.95 | Faction Men's Ski Candidate 4.0 Downhill Skis \$719.20 - \$934.95 | | Volkl Nunaq Ski \$699.00 |

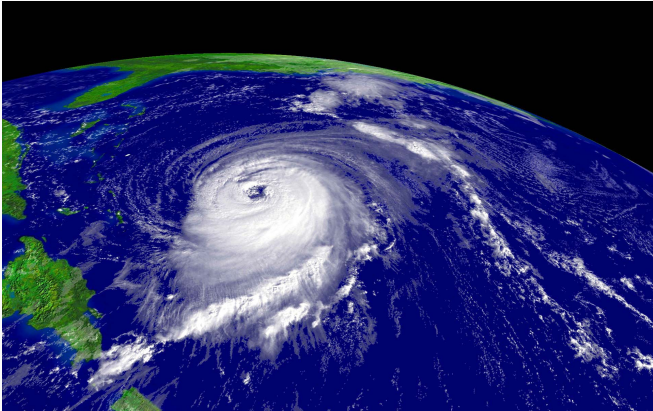
Customers Who Viewed This Item Also Viewed

Page 1 of 7

| | | | | | |
|---|---|---|---|---|--|
|  |  |  |  |  |  |
|---|---|---|---|---|--|

Predicting Strength of Hurricanes

NOAA Models: SHIFOR, SHIPS, DSHIPS, LGEM



Hurricane Isabel, 2003, Source: NOAA.gov

Other Applications

1. Health-care: Identify risks of getting a disease
2. Detect spam in emails
3. Recognize hand-written text
4. Create a fake video (<https://www.youtube.com/watch?v=cQ54GDm1eL0>)

Any other applications?

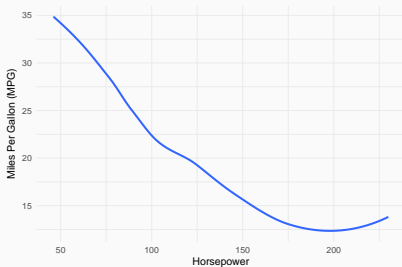
What is Machine Learning

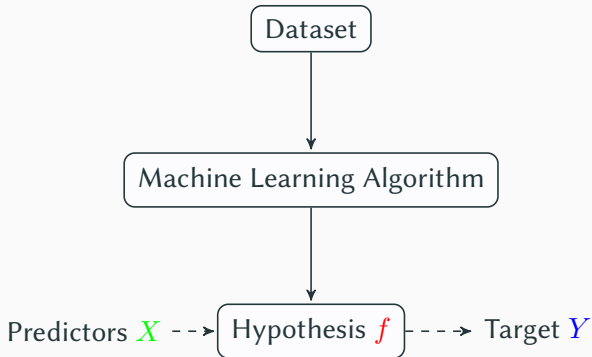
- Discover unknown function f :

$$Y = f(X)$$

- f = **hypothesis**, or model
- X = **features**, or predictors, or inputs
- Y = **response**, or target

$$\text{MPG} = f(\text{Horsepower})$$



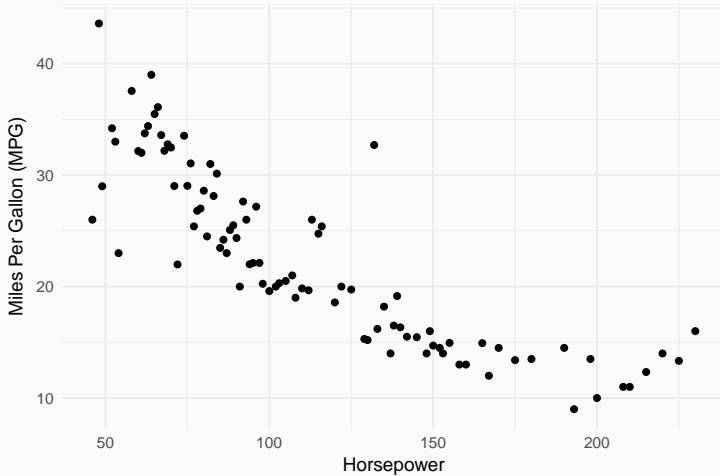


Purpose: Inference or prediction

Auto Dataset

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |
| | ... | ... | ... |

Auto Dataset



Bayes Classifier

What is the **MPG** of a car with **horsepower** = 150?

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |

$$f(x) = \mathbb{E}[Y \mid X = x]$$

Bayes Classifier

What is the **MPG** of a car with **horsepower** = 150?

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |

$$f(x) = \mathbb{E}[Y \mid X = x]$$

Bayes Classifier

What is the **MPG** of a car with **horsepower** = 150?

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |

$$f(x) = \mathbb{E}[Y \mid X = x]$$

Limitation of Bayes Classifier

What is the MPG of a car with horsepower = 200?

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |

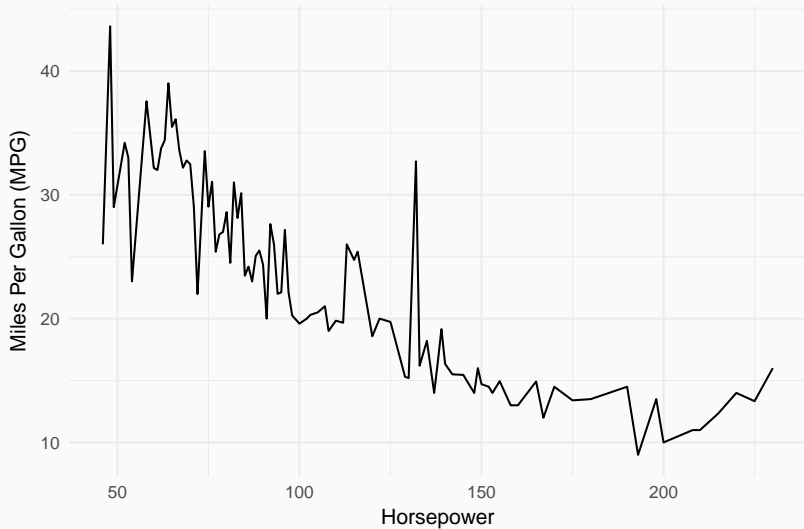
Limitation of Bayes Classifier

What is the **MPG** of a car with **horsepower** = 200?

| | mpg | horsepower | name |
|----|-------|------------|---------------------------|
| 1 | 18.00 | 130.00 | chevrolet chevelle malibu |
| 2 | 15.00 | 165.00 | buick skylark 320 |
| 3 | 18.00 | 150.00 | plymouth satellite |
| 4 | 16.00 | 150.00 | amc rebel sst |
| 5 | 17.00 | 140.00 | ford torino |
| 6 | 15.00 | 198.00 | ford galaxie 500 |
| 7 | 14.00 | 220.00 | chevrolet impala |
| 8 | 14.00 | 215.00 | plymouth fury iii |
| 9 | 14.00 | 225.00 | pontiac catalina |
| 10 | 15.00 | 190.00 | amc ambassador dpl |

Return the **nearest neighbor**.

Nearest Neighbor Hypothesis



Must allow for errors ϵ :

$$Y = f(X) + \epsilon$$

1. World is too complex to model precisely
2. Many features are not captured in data sets
3. Datasets are limited

Must allow for errors ϵ :

$$Y = f(X) + \epsilon$$

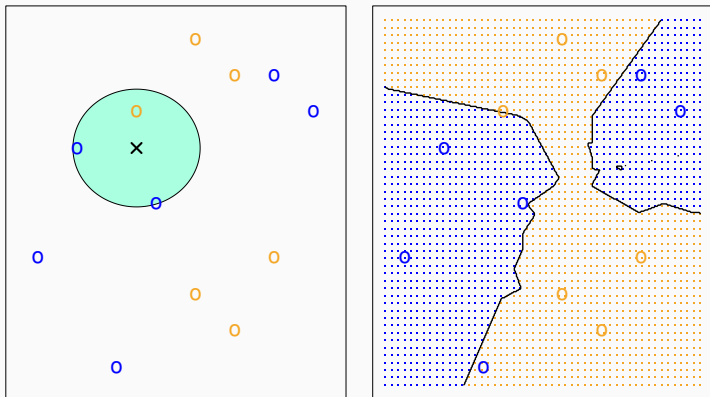
1. World is too complex to model precisely
2. Many features are not captured in data sets
3. Datasets are limited

How to reduce prediction errors?

KNN: K-Nearest Neighbors

Idea: Use several similar training points when making predictions. Errors will average out.

Example with 2 features (horsepower, model year)



KNN: Effect of different k

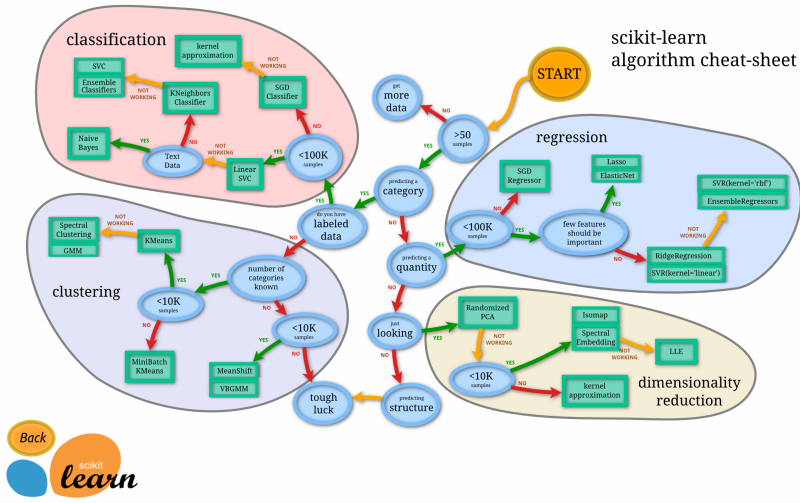


Questions ...

- How to choose k ?
- Are there better methods?

Machine Learning Choices ...

scikit-learn
algorithm cheat-sheet



Source: <http://scikit-learn.org/stable/tutorial/machinelearningmap/index.html>

End of the semester: Know what, when, and why to use, know

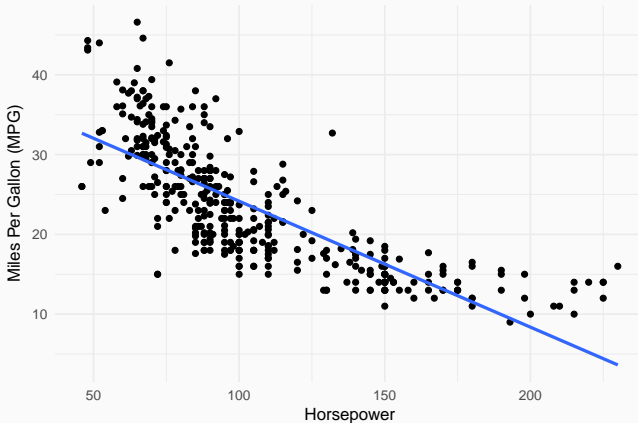
This Course: First Steps in Machine Learning

1. **Foundations:** Mathematics necessary to use and understand ML algorithms
2. **Algorithms:** Use common ML algorithms
3. **Principles:** Understand when different algorithms are appropriate
- 4.

Parametric Prediction Methods

Linear regression:

$$\text{MPG} = f(\text{horsepower}) = \beta_0 + \beta_1 \times \text{horsepower}$$



- Download and install R: <http://cran.r-project.org>
- Try using RStudio as an R IDE
- *See class website for more info*
- All my R code is in the class git repo
- You can use Python or another language, but at your own risk

Why R (vs Python)

1. Language syntax particularly suitable for manipulating tabular data
2. Better-quality packages at `cran.r-project.org` than `pypi.org`
3. Excellent data manipulation and visualization tools: `dplyr`, `ggplot`, `tidyverse`, better than python versions like `pandas`, `matplotlib`
4. R-notebooks more flexible and git-friendly than Jupyter
5. Shiny: a neat web framework to create simple web data interface
6. Rcpp: convenient interface with C++
7. Easier to install and use particularly on Windows/Mac

Why Python (vs R)

1. It is popular
2. Better general programming language, good support for data structures
3. Better support for numerical algebra: numpy, scipy,
4. Much better support for deep learning: tensor flow, keras
5. Clean syntax

- **Website:** (get there through mycourses)
`https://gitlab.cs.unh.edu/carton/ml-fall2025`
- **Grading:** See website
- **Assignments:** posted on myCourses at least a week in advance
- **Questions:** myCourses Discussion
- **Programming language:** R, Python