Lecture 10
Introduction to Machine Learning

Dennis Sun
Stanford University
DATASCI 112

February 2, 2024

1. What is Machine Learning?

2. A Controversy in the Wine World

Classic Artificial Intelligence

Classic AI attempts to codify the rules that a human would use to make decisions.

Example: If you are trying to build a system that finds all the proper nouns in a text document, you might hard-code the following rules:

- If a word is a proper noun, then the first letter of the word is capitalized.
- The first letter of a sentence is always capitalized.
- …

The system can deduce new rules from existing rules, e.g.,

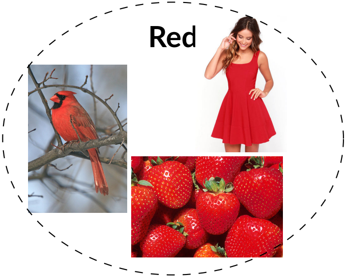
- It is impossible to tell whether the first word of a sentence is a proper noun just from the capitalization.

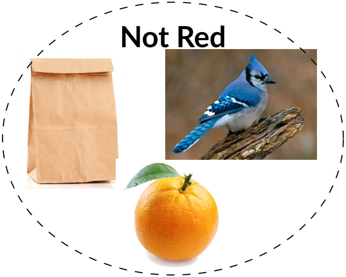
Pros: The model is super interpretable!
Cons: For complex tasks, there are too many rules, and we can't anticipate them all.

What is Machine Learning?

Exercise: Pair up with the person sitting next to you. One of you will be an Earthling, the other a Martian.
The Earthling should explain to the Martian what the word "red" means. The Martian should try to be obtuse.

Moral: We often learn by seeing examples.

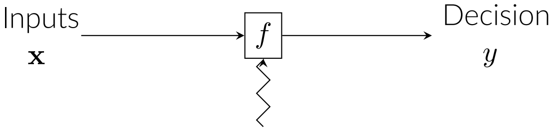




Rather than trying to come up with the rules ourselves, we can learn the rules from data. This is the essence of **machine learning**.

What is Machine Learning?

Learning refers to the act of coming up with a rule for making decisions based on a set of inputs.



Inputs $\mathbf{x}$

f

Decision y

Goal of Machine Learning:
Come up with a rule f
from **training data** $(\mathbf{x}_i, y_i)$.

The decision y is typically called the **target** or the **label**.

1 What is Machine Learning?

2 A Controversy in the Wine World

Orley Ashenfelter
Economics Professor

In 1991, Orley Ashenfelter predicted that the 1986 vintage of Bordeaux wines would be disappointing.

He did this without tasting a drop of the wine.

Robert Parker
The Wine Advocate

Wine critics were outraged.

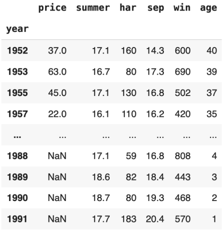
Robert Parker had predicted that the 1986 vintage would be "very good and sometimes exceptional" based on tasting an early sample.

How did Ashenfelter make this prediction?

Ashenfelter collected data on summer temperature and winter rainfall in Bordeaux from 1950 to 1991.

The quality of wines becomes apparent after 10 years. So for vintages up to 1980, he also collected their price.

```python
import pandas as pd

df = pd.read_csv("https://dlsun.github.io/pods/data/bordeaux.csv",
                 index_col="year")
df
```



	price	summer	har	sep	win	age
year						
1952	37.0	17.1	160	14.3	600	40
1953	63.0	16.7	80	17.3	690	39
1955	45.0	17.1	130	16.8	502	37
1957	22.0	16.1	110	16.2	420	35
...	...	...	...	...	...	...
1988	NaN	17.1	59	16.8	808	4
1989	NaN	18.6	82	18.4	443	3
1990	NaN	18.7	80	19.3	468	2
1991	NaN	17.7	183	20.4	570	1

38 rows × 6 columns

$$\mathbf{x} \longrightarrow \boxed{f} \longrightarrow y$$

summer temp,

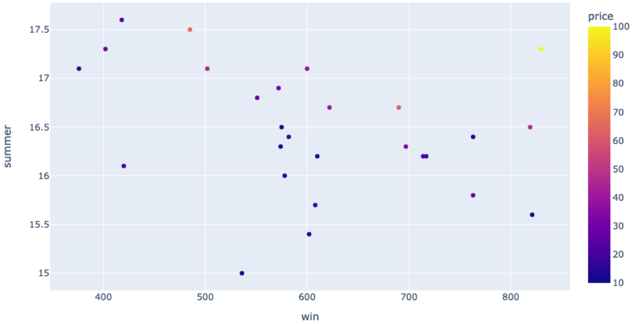
winter rainfall,

harvest rainfall, age

price

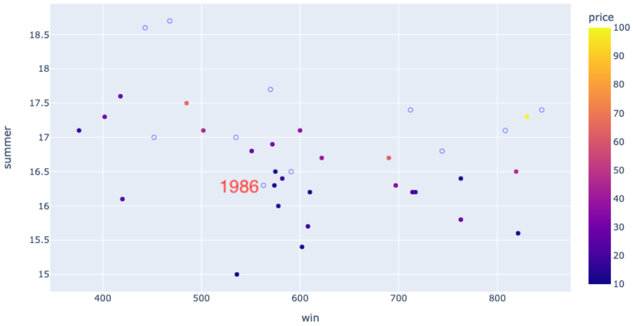
Visualizing the Data

```
import plotly.express as px

px.scatter(df[~df["price"].isnull()],
           x="win", y="summer", color="price")
```



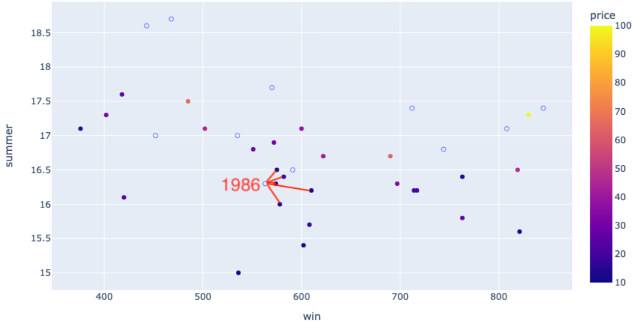
Visualizing the Data

```
import plotly.graph_objects as go

fig1 = px.scatter(df[~df["price"].isnull()],
                  x="win", y="summer", color="price")
fig2 = px.scatter(df[df["price"].isnull()],
                  x="win", y="summer", symbol_sequence=["circle-open"])

go.Figure(data=fig1.data + fig2.data, layout=fig1.layout)
```



What would you predict is the quality of the 1986 wine?

Visualizing the Data



Insight: The “closest” wines are low quality, so the 1986 vintage is probably low quality as well.

This is the intuition behind k-**nearest neighbors**.

Types of Machine Learning Problems

Machine learning problems are grouped into two types, based on the type of y:

Regression: The label y is quantitative.

Classification: The label y is categorical.

Was Ashenfelter's wine problem a regression or a classification problem?

Note that the input features $\mathbf{x}$ may be categorical, quantitative, textual, ..., or any combination of these.