Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 2

Slides adapted from Noah Smith and Chenhao Tan

Administrivia

- Make sure that you enroll in Canvas and have access to Piazza
- Temporary schedule is released `https://github.com/akkikiki/CSCI-4622-Machine-Learning-sp21/blob/main/info/schedule.md`
- Office hours are on Thursdays and Fridays, 4-5pm

Learning Objectives

- Understand feature extraction
- Understand the basics of decision tree

Outline

Ice-Breaking

Features

Decision tree

Information gain as splitting criteria

Outline

Canonical Learning Problems

Outputs being discrete vs. continuous

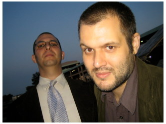
- Regression
- Classificaiton
  - Binary Classificaiton
  - Multi-class Classificaiton

What would the following tasks be? Regression or classification?

- Predict the value of a house
- Predict whether Bob will play tennis or not.

Outline

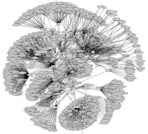
Features



$\longrightarrow \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle$

Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$



$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Features

Let ϕ be a function that maps from inputs ($\boldsymbol{x}$) to values.

Features

Let ϕ be a function that maps from inputs ($\boldsymbol{x}$) to values.

- If ϕ maps to $\{0, 1\}$, we call it a "binary feature."

Features

Let ϕ be a function that maps from inputs ($\boldsymbol{x}$) to values.

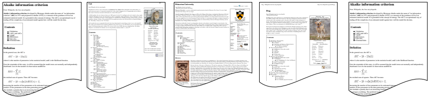
- If ϕ maps to $\{0, 1\}$, we call it a "binary feature."
- If ϕ maps to $\mathbb{R}$, we call it a "real-valued feature."

Features

Let ϕ be a function that maps from inputs ($\boldsymbol{x}$) to values.

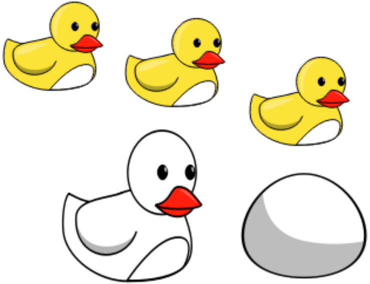
- If ϕ maps to $\{0, 1\}$, we call it a "binary feature."
- If ϕ maps to $\mathbb{R}$, we call it a "real-valued feature."
- Features can be categorical values, ordinal values, integers, and more.

Understanding assumptions in features



- When/why are they appropriate?
- Much of this is an art, and it is inherently dynamic
  - Documents can be analyzed as a sequence of words;
    - E.g., “dogs like cats.” vs. “cats like dogs.”
  - or, as a “bag” of words.
    - E.g., dogs: 1, like: 1, cats:1

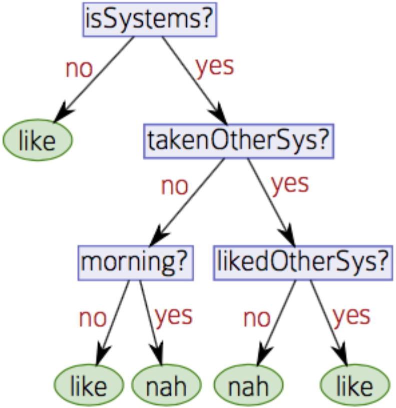
(Extended Reading) Ugly Duckling Theorem (Watanabe 1969)



"...any two entities can be arbitrarily similar or dissimilar by changing the criterion of what counts as a relevant attribute. Unless one can specify such criteria, then the claim that categorization is based on attribute matching is almost entirely vacuous" (Murphy and Edin 1985)

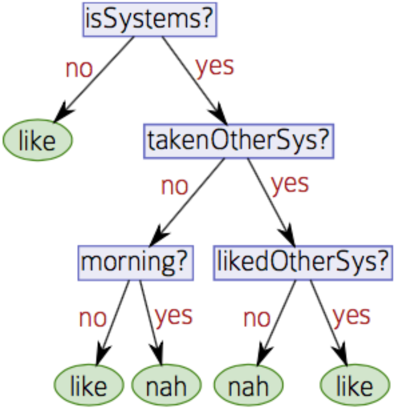
Outline

Overview



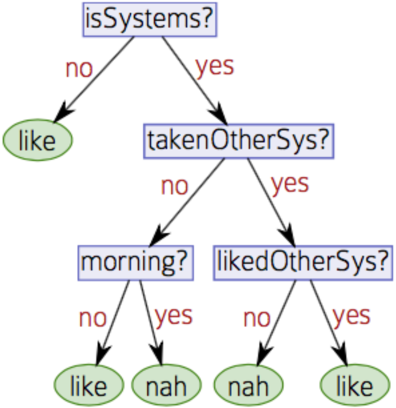
- Task: Will Alice enjoy taking some unknown class x?

Overview



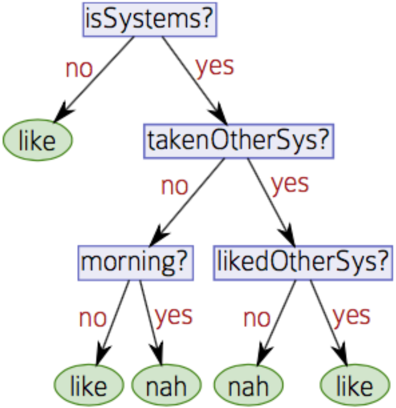
- Task: Will Alice enjoy taking some unknown class x?
- Ask questions about that unknown class, or about Alice

Overview



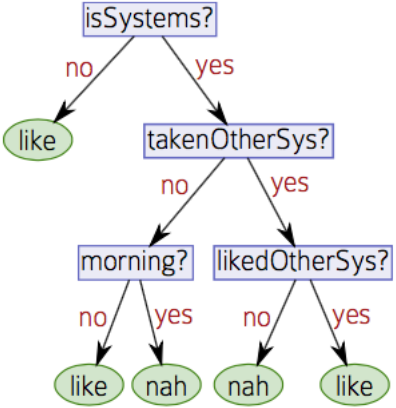
- Task: Will Alice enjoy taking some unknown class x?
- Ask questions about that unknown class, or about Alice
- Is the class classified as systems class?

Overview



- Task: Will Alice enjoy taking some unknown class x?
- Ask questions about that unknown class, or about Alice
- Is the class classified as systems class?
- Did Alice took a systems class?

Overview using Machine Learning Terminology



- Task: Will Alice enjoy taking some unknown class x?
- Which features should we use?
- $\phi_{\text{systems}}(x) = \{0, 1\}$?
- $\phi_{\text{systems, Alice}}(x) = \{0, 1\}$?

Splitting

Example: Predict whether Bob will play tennis on a given day.

When does Bob play tennis?

Splitting

- Bob's tennis log is provied as follows (i.e., training data)
- Consider the tennis problem now with binary features

X			Y
sun	wind	humidity	tennis
sunny	windy	not humid	tennis
sunny	not windy	not humid	tennis
not sunny	not windy	humid	no tennis
sunny	windy	humid	no tennis

Splitting

Converting to binary features and labels

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

Splitting

Converting to binary features and labels

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What would be a good feature to use to “split” these data into two groups?

Splitting

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

Let's use $\phi_{\text{sun}}(X)$ to split

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

Let's use $\phi_{\text{sun}}(X)$ to split

- $X_{\text{left},\text{sun}} : \{0\}$
- $X_{\text{right},\text{sun}} : \{1, 1, 0\}$

Splitting

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

How about $\phi_{\text{wind}}(X)$?

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

How about $\phi_{\text{wind}}(X)$?

- $X_{\text{left},\text{wind}} : \{1, 0\}$
- $X_{\text{right},\text{wind}} : \{1, 0\}$

Splitting

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

How about $\phi_{\text{humid}}(X)$?

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

How about $\phi_{\text{humid}}(X)$?

- $X_{\text{left},\text{humid}} : \{1, 1\}$
- $X_{\text{right},\text{humid}} : \{0, 0\}$

Can we formalize this?

Outline

Information gain as splitting criteria

- Inspired by information theory
- Entropy: measure of **impurity** of set of examples

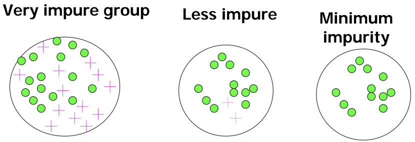


Image from `https://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf`

Entropy

$$H(X) = -\sum_{c} p_c \log_2(p_c),$$

where p_c is the fraction of examples in class (label) c.

Entropy

$$H(X) = -\sum_c p_c \log_2(p_c),$$

where p_c is the fraction of examples in class (label) c. Note that for binary classification, let p be the fraction in the positive class, then

$$H(X) = -p \log_2 p - (1-p) \log_2(1-p)$$

Entropy

$$H(X) = -\sum_c p_c \log_2(p_c),$$

where p_c is the fraction of examples in class (label) c. Note that for binary classification, let p be the fraction in the positive class, then

$$H(X) = -p \log_2 p - (1-p) \log_2(1-p)$$

What is the largest/smallest entropy?

Entropy

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p)$$

What is the largest/smallest entropy?

$$0 \leq p \leq 1$$

Entropy

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p)$$

What is the largest/smallest entropy?

$$0 \leq p \leq 1$$

- When all examples are in the same class, entropy is 0
- When samples are equally balanced, entropy is 1

Information gain

The higher entropy is, the lower the information is.
Information gain is defined as the **difference between impurity at the parent and (weighted average) of impurity at the children**

Information gain

The higher entropy is, the lower the information is.
Information gain is defined as the **difference between impurity at the parent and (weighted average) of impurity at the children**
Splitting based on feature i

- X_{parent}: training subset of the parent node
- $X_{i,\text{left}}$: training subset of the left node
- $X_{i,\text{right}}$: training subset of the right node

Information gain

The higher entropy is, the lower the information is.
Information gain is defined as the **difference between impurity at the parent and (weighted average) of impurity at the children**
Splitting based on feature i

- X_{parent}: training subset of the parent node
- $X_{i,\text{left}}$: training subset of the left node
- $X_{i,\text{right}}$: training subset of the right node

$$IG(X_{\text{parent}}, i) = H(X_{\text{parent}}) - \frac{|X_{i,\text{left}}|}{|X_{\text{parent}}|} H(X_{\text{left}}) - \frac{|X_{i,\text{right}}|}{|X_{\text{parent}}|} H(X_{\text{right}})$$

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{sun})$?

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{sun})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left,sun}} : \{0\}$
- $X_{\text{right,sun}} : \{1, 1, 0\}$

Splitting

X			Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{sun})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left,sun}} : \{0\}$
- $X_{\text{right,sun}} : \{1, 1, 0\}$
- $H(X_{\text{parent}}) = 1$
- $H(X_{\text{left,sun}}) = 0$
- $H(X_{\text{right,sun}}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

$$IG(X, \text{sun}) = 1 - \frac{1}{4} * 0 - \frac{3}{4} * 0.918 = 0.3112$$

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{wind})$?

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{wind})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left,wind}} : \{1, 0\}$
- $X_{\text{right,wind}} : \{1, 0\}$

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{wind})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left},\text{wind}} : \{1, 0\}$
- $X_{\text{right},\text{wind}} : \{1, 0\}$
- $H(X_{\text{parent}}) = 1$
- $H(X_{\text{left},\text{wind}}) = 1$
- $H(X_{\text{right},\text{wind}}) = 1$

$$IG(X, \text{wind}) = 1 - \frac{1}{2} * 1 - \frac{1}{2} * 1 = 0$$

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{humid})$?

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{humid})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left,humid}} : \{1, 1\}$
- $X_{\text{right,humid}} : \{0, 0\}$

Splitting

	X		Y
sun	wind	humidity	tennis
1	1	0	1
1	0	0	1
0	0	1	0
1	1	1	0

What is $IG(X, \text{humid})$?

- $X_{\text{parent}} : \{1, 1, 0, 0\}$
- $X_{\text{left,humid}} : \{1, 1\}$
- $X_{\text{right,humid}} : \{0, 0\}$
- $H(X_{\text{parent}}) = 1$
- $H(X_{\text{left,humid}}) = 0$
- $H(X_{\text{right,humid}}) = 0$

$$IG(X, \text{humid}) = 1 - \frac{1}{2} * 0 - \frac{1}{2} * 0 = 1$$

Splitting

- $IG(X, \text{sun}) = 0.3112$
- $IG(X, \text{wind}) = 0$
- $IG(X, \text{humid}) = 1$

Which feature should we split on?

Splitting

- $IG(X, \text{sun}) = 0.3112$
- $IG(X, \text{wind}) = 0$
- $IG(X, \text{humid}) = 1$

Which feature should we split on?

humid, since it brings the greatest information gain.