Logic and machine learning review

CS 540

Yingyu Liang

Propositional logic

Logic

- If the rules of the world are presented formally, then a decision maker can use logical reasoning to make rational decisions.

- Several types of logic:
  - propositional logic (Boolean logic)
  - first order logic (first order predicate calculus)

- A logic includes:
  - syntax: what is a correctly formed sentence
  - semantics: what is the meaning of a sentence
  - Inference procedure (reasoning, entailment): what sentence logically follows given knowledge

Propositional logic syntax

Sentence	$\rightarrow$ *AtomicSentence* | *ComplexSentence*
AtomicSentence	$\rightarrow$ True | False | *Symbol*
Symbol	$\rightarrow$ P | Q | R | . . .
ComplexSentence	$\rightarrow \neg$*Sentence*
|	(*Sentence* $\wedge$ *Sentence*)
|	(*Sentence* $\vee$ *Sentence*)
|	(*Sentence* $\Rightarrow$ *Sentence*)
|	(*Sentence* $\Leftrightarrow$ *Sentence*)

BNF (Backus-Naur Form) grammar in propositional logic

$((\neg P \vee ((True \wedge R) \Leftrightarrow Q)) \Rightarrow S$ well formed

$(\neg(P \vee Q) \wedge \ \Rightarrow S)$ not well formed

Summary

- Interpretation, semantics, knowledge base
- Entailment
  - model checking
- Inference, soundness, completeness
- Inference methods
  - Sound inference, proof
  - Resolution, CNF
  - Chaining with Horn clauses, forward/backward chaining

Example

8. (Resolution) Given knowledge base

   (a) $P \Leftrightarrow Q$

   (b) P

   use resolution to prove query Q.

First order logic

FOL syntax Summary

- Short summary so far:
  - **Constants:** Bob, 2, Madison, …
  - **Variables:** *x, y, a, b, c*, …
  - **Functions:** Income, Address, Sqrt, …
  - **Predicates:** Teacher, Sisters, Even, Prime…
  - **Connectives:** $\wedge \vee \neg \Rightarrow \Leftrightarrow$
  - **Equality:** $=$
  - **Quantifiers:** $\forall \exists$

More summary

- **Term**: constant, variable, function. Denotes an object. (A ground term has no variables)
- **Atom**: the smallest expression assigned a truth value. Predicate and =
- **Sentence**: an atom, sentence with connectives, sentence with quantifiers. Assigned a truth value
- **Well-formed formula** (wff): a sentence in which all variables are quantified

Example

9. (FOL) Which one is the translation of "Frodo has exactly one ring"?

(A) $\exists x, y \; HasRing(Frodo, x) \wedge HasRing(Frodo, y) \wedge x = y$
(B) $\forall x \; HasRing(Frodo, x) \Rightarrow \exists y(HasRing(Frodo, y) \wedge x = y)$
(C) $\exists x \; HasRing(Frodo, x) \Rightarrow \forall y(HasRing(Frodo, y) \wedge x = y)$
(D) $\exists x \; HasRing(Frodo, x) \wedge \forall y(HasRing(Frodo, y) \Rightarrow x = y)$
(E) none of the above

Example

9. (FOL) Which one is the translation of "Frodo has exactly one ring"?

(A) $\exists x, y\ HasRing(Frodo, x) \land HasRing(Frodo, y) \land x = y$
(B) $\forall x\ HasRing(Frodo, x) \Rightarrow \exists y(HasRing(Frodo, y) \land x = y)$
(C) $\exists x\ HasRing(Frodo, x) \Rightarrow \forall y(HasRing(Frodo, y) \land x = y)$
(D) $\exists x\ HasRing(Frodo, x) \land \forall y(HasRing(Frodo, y) \Rightarrow x = y)$
(E) none of the above

A: D

Machine learning basics

What is machine learning?

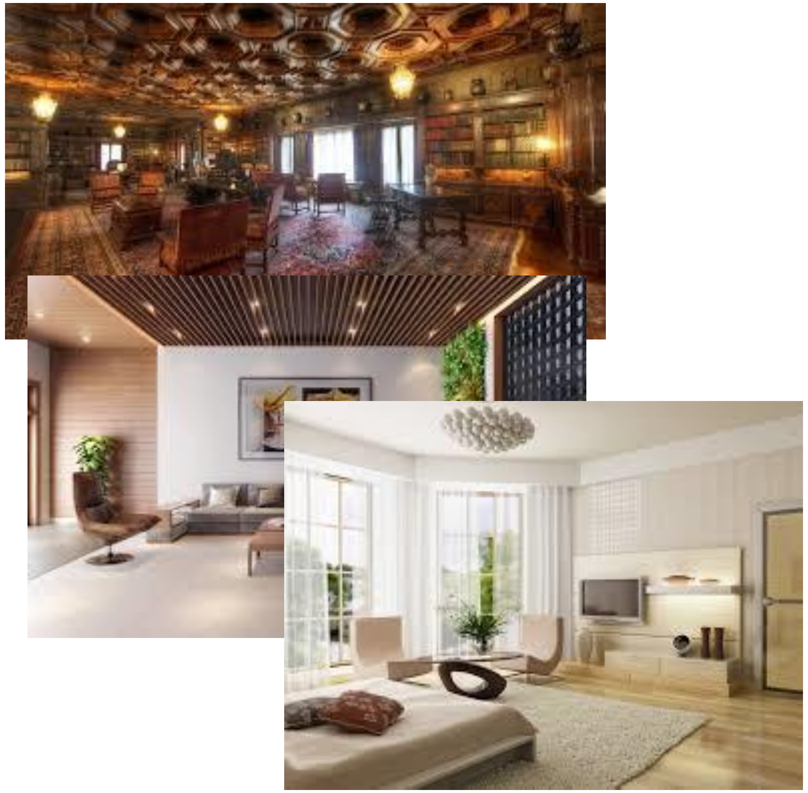
- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience E.”

------- *Machine Learning*, Tom Mitchell, 1997

Example 1: image classification

Task: determine if the image is indoor or outdoor
Performance measure: probability of misclassification

Example 1: image classification



Experience/Data:
images with labels

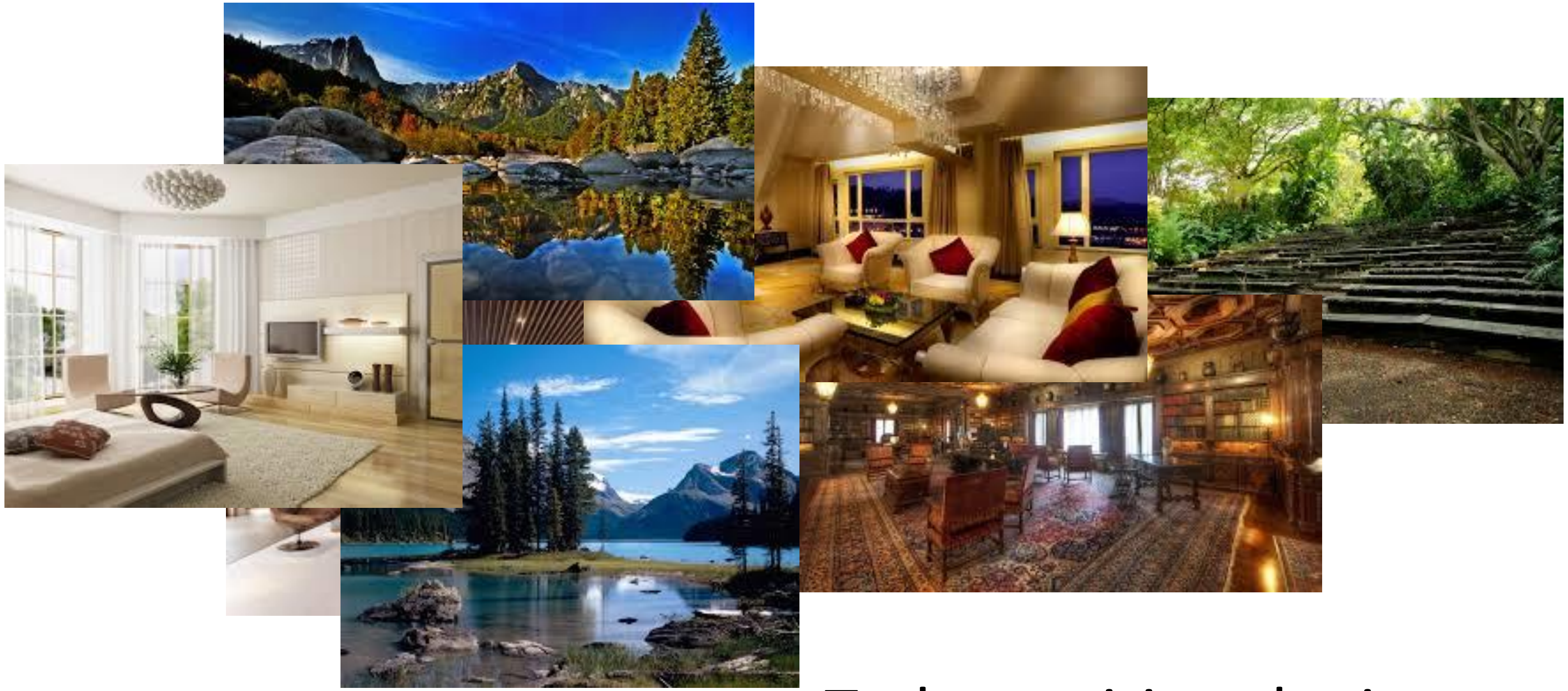
Indoor

outdoor

Example 1: image classification

- A few terminologies
  - Training data: the images given for learning
  - Test data: the images to be classified
  - Binary classification: classify into two classes

Example 2: clustering images



Task: partition the images into 2 groups
Performance: similarities within groups
Data: a set of images

Example 2: clustering images

- A few terminologies
  - Unlabeled data vs labeled data
  - Supervised learning vs unsupervised learning

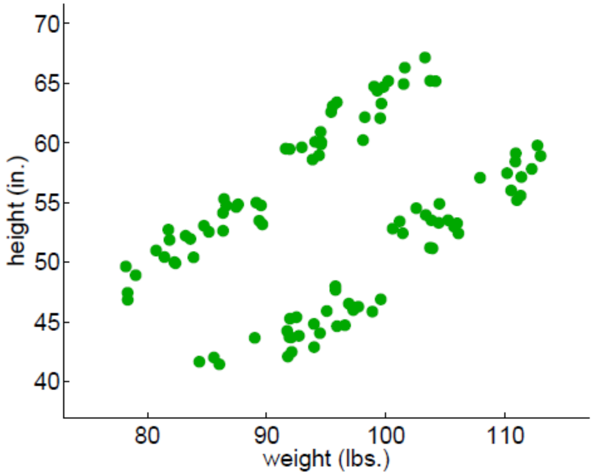
Unsupervised learning

Unsupervised learning

- Training sample $x_1, x_2, \dots, x_n$
- No teacher providing supervision as to how individual instances should be handled
- Common tasks:
  - clustering, separate the *n* instances into groups
  - novelty detection, find instances that are very different from the rest
  - dimensionality reduction, represent each instance with a lower dimensional feature vector while maintaining key characteristics of the training samples

Clustering

- Group training sample into *k* clusters
- How many clusters do you see?
- Many clustering algorithms
  - HAC (Hierarchical Agglomerative Clustering)
  - k-means
  - …



Hierarchical Agglomerative Clustering

Input: *a training sample* $\{\mathbf{x}_i\}_{i=1}^{n}$; *a distance function* $d()$.
1. Initially, place each instance in its own cluster (called a singleton cluster).
2. ***while*** *(number of clusters* > 1*)* ***do:***
3. Find the closest cluster pair A, B, *i.e., they minimize* $d(A, B)$.
4. Merge A, B *to form a new cluster.*
Output: *a binary tree showing how clusters are gradually merged from singletons to a root cluster, which contains the whole training sample.*

- Euclidean (L2) distance

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\sum_{s=1}^{D} (x_{is} - x_{js})^2}.$$

K-means algorithm

- Input: $x_1 \ldots x_n$, k
- **Step 1**: select k cluster centers $c_1 \ldots c_k$
- **Step 2**: for each point x, determine its cluster: find the closest center in Euclidean space
- **Step 3**: update all cluster centers as the centroids

$$c_i = \sum_{\{x \text{ in cluster } i\}} x \Big/ \text{SizeOf(cluster } i)$$

- Repeat step 2, 3 until cluster centers no longer change

Example

17. (Clustering) There are five points in one-dimensional space: $a = 0, b = 1, c = 3, d = 7, e = 9$. Perform Hierarchical Agglomerative Clustering with complete linkage. Complete the resulting clustering tree diagram (i.e., the dendrogram).

$a = 0$ $\quad b = 1$ $\quad c = 3$ $\quad d = 7$ $\quad e = 9$

Example

17. (Clustering) There are five points in one-dimensional space: $a = 0, b = 1, c = 3, d = 7, e = 9$. Perform Hierarchical Agglomerative Clustering with complete linkage. Complete the resulting clustering tree diagram (i.e., the dendrogram).

$$a = 0 \qquad b = 1 \qquad c = 3 \qquad d = 7 \qquad e = 9$$

```
(ab) c d e
(ab) c (de)
(ab) c: 3, c (de): 6
((ab) c) (de)
(((ab) c) (de))
```

Supervised learning

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from some unknown distribution D
- Find $y = f(x) \in \boldsymbol{\mathcal{H}}$ using training data
- s.t. f correct on test data i.i.d. from distribution D

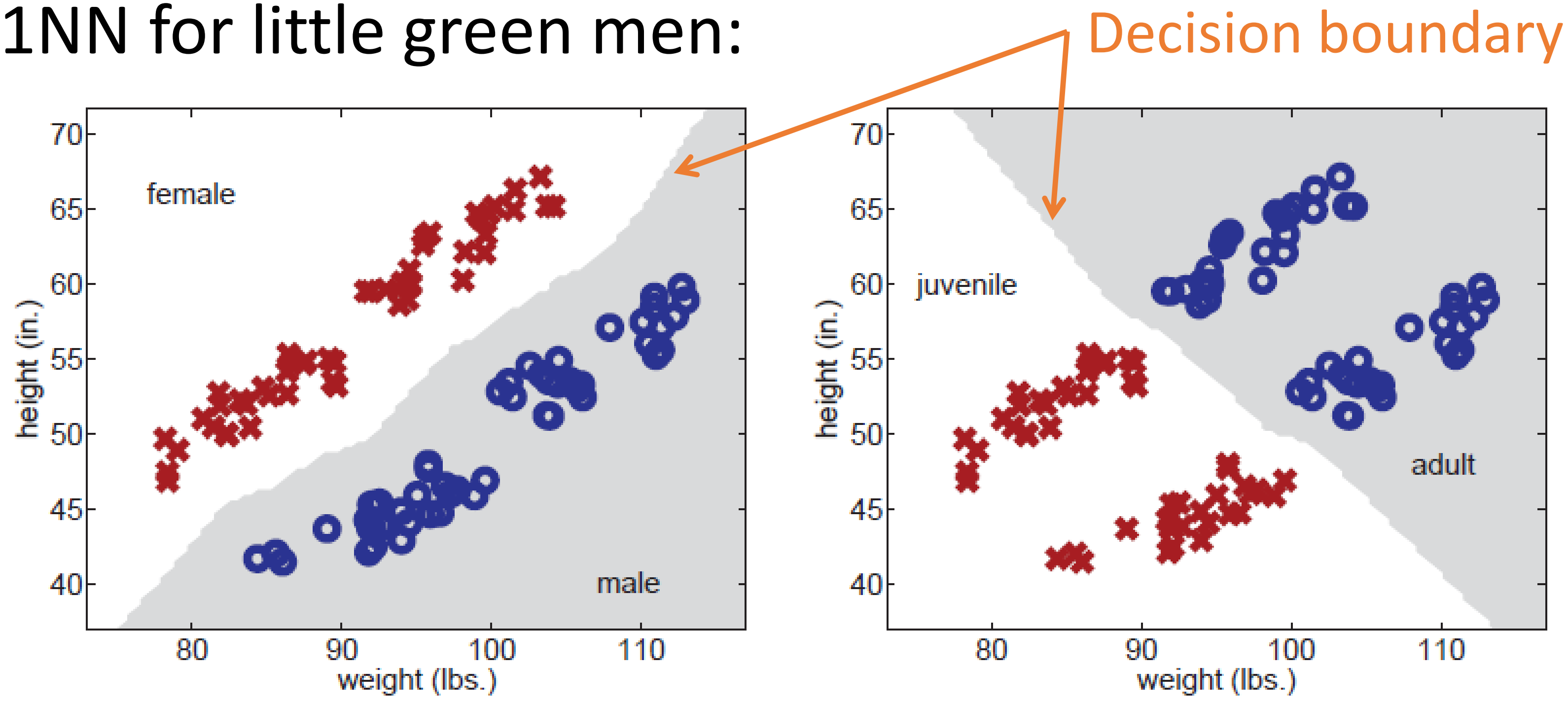
- If label y discrete: classification
- If label y continuous: regression

k-nearest-neighbor (kNN)

Input: Training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$*; distance function* $d()$*;*
number of neighbors k*; test instance* $\mathbf{x}^*$

1. *Find the* k *training instances* $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ *closest to* $\mathbf{x}^*$ *under distance* $d()$.
2. *Output* y^* *as the majority class of* $y_{i_1}, \ldots, y_{i_k}$. *Break ties randomly.*

- 1NN for little green men:



(a) classification by gender

(b) classification by age

Math formulation

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \mathcal{H}$ that minimizes $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f, x_i, y_i)$
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

- Examples of loss functions:
  - 0-1 loss for classification: $l(f, x, y) = \mathbb{I}[f(x) \ne y]$ and $L(f) = \Pr[f(x) \ne y]$
  - l_2 loss for regression: $l(f, x, y) = [f(x) - y]^2$ and $L(f) = \mathbb{E}[f(x) - y]^2$

Maximum likelihood Estimation (MLE)

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Let $\{P_\theta(x, y): \theta \in \Theta\}$ be a family of distributions indexed by θ

- MLE: negative log-likelihood loss

$$\theta_{ML} = \mathrm{argmax}_{\theta \in \Theta} \sum_i \log(P_\theta(x_i, y_i))$$

$$l(P_\theta, x_i, y_i) = -\log(P_\theta(x_i, y_i))$$
$$\hat{L}(P_\theta) = -\sum_i \log(P_\theta(x_i, y_i))$$

MLE: conditional log-likelihood

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Let $\{P_\theta(y|x): \theta \in \Theta\}$ be a family of distributions indexed by θ

Only care about predicting y from x; do not care about p(x)

- MLE: negative conditional log-likelihood loss

$$\theta_{ML} = \mathrm{argmax}_{\theta \in \Theta} \sum_i \log(P_\theta(y_i|x_i))$$

$$l(P_\theta, x_i, y_i) = -\log(P_\theta(y_i|x_i))$$

$$\hat{L}(P_\theta) = -\sum_i \log(P_\theta(y_i|x_i))$$

Linear regression with regularization: Ridge regression

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $f_w(x) = w^T x$ that minimizes $\widehat{L_R}(f_w) = \frac{1}{n}\|Xw - y\|_2^2$

- By setting the gradient to be zero, we have

$$\mathrm{w} = (X^T X)^{-1} X^T y$$

l_2 loss: Normal + MLE

Linear classification: logistic regression

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$ i.i.d. from distribution D

- Assume $P_w(y = 1|x) = \sigma(w^T x) = \frac{1}{1+\exp(-w^T x)}$

$$P_w(y = 0|x) = 1 - P_w(y = 1|x) = 1 - \sigma(w^T x)$$

- Find w that minimizes

$$\hat{L}(w) = -\frac{1}{n}\sum_{i=1}^{n} \log P_w(y_i|x_i)$$

Example

11. (Linear regression) Long time ago, a primate researcher gave you a data set to predict the label y (monkey daily diet weight) from a number of features $x_1, \ldots, x_d$. You built a linear regression model for him:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

Yesterday, the monkey researcher realized that his RA mixed things up: x_d was actually the label, while y was the d-th feature! Alas, neither him nor you have the data set anymore, and the RA was long gone to start up a monkey intelligence company and does not respond to emails. All you have are the coefficients $\beta_0 \ldots \beta_d$, all non-zero. How do you fix the linear regression model? (One line math)

Example

11. (Linear regression) Long time ago, a primate researcher gave you a data set to predict the label y (monkey daily diet weight) from a number of features $x_1, \ldots, x_d$. You built a linear regression model for him:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

Yesterday, the monkey researcher realized that his RA mixed things up: x_d was actually the label, while y was the d-th feature! Alas, neither him nor you have the data set anymore, and the RA was long gone to start up a monkey intelligence company and does not respond to emails. All you have are the coefficients $\beta_0 \ldots \beta_d$, all non-zero. How do you fix the linear regression model? (One line math)

A:

$$(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_{d-1} x_{d-1}) / \beta_d = x_d.$$