CSE 471: MACHINE LEARNING

Learning from Examples

Outline

- Forms of learning
- Model Selection and Optimization
- The Theory of Learning
- Linear Regression and Classification
- Nonparametric Models
- Ensemble Learning
- Developing Machine Learning Systems

Machine Learning

- A computer program is said to learn from
  - experience E (not in traditional algorithm e.g. Dijkstra algorithm)
- with respect to some class of
  - task T and performance measure P if its
  - performance at tasks in T
  - as measured by P
  - improves with experience E (not in traditional algorithm)

Material to cover in next 7 weeks

- Book
  - Artificial Intelligence A Modern Approach (4th Edition)
    - Stuart J. Russell and Peter Norvig
  - Misc. notes/blog posts
- Content
  - See course outline

Learning from Examples

- Supervised Learning
  - Labeled observations
  - Classification vs. Regression
- Unsupervised Learning
  - Unlabeled observations
  - Clustering

Learning from Examples

- Reinforcement Learning
  - Reward & punishment scheme
  - Example: Learning from a chess game
    - Win → Reward
    - Loss → Punishment
    - Agent itself determines what actions were fruitful!

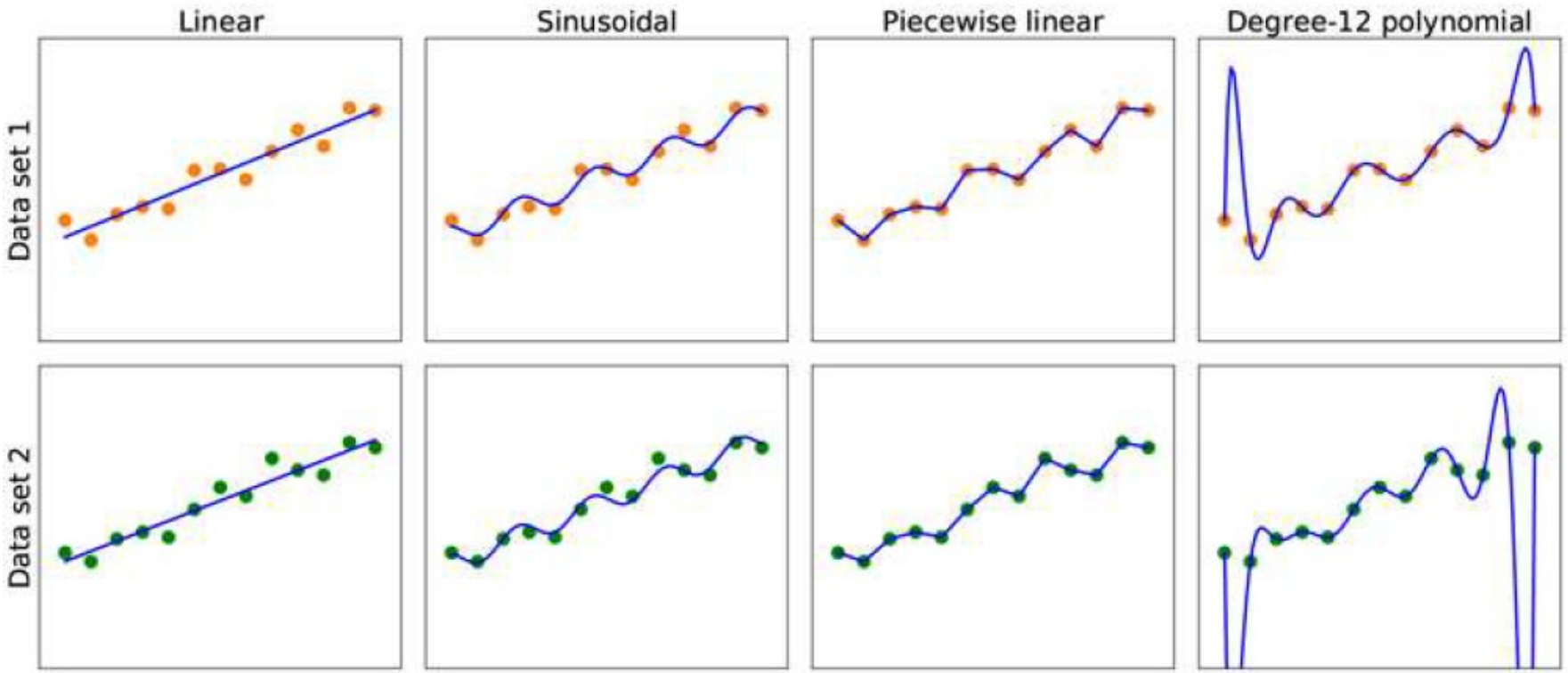
Datasets & Validations

- Datasets
  - Training set
  - Validation set
  - Test set
- Validations
  - Validation Test
  - K-fold cross validation Test
  - Independent Test

The hypothesis space (H)

- How do we choose a hypothesis space
  - Prior knowledge
  - Exploratory data analysis
    - Visualizations
      - histograms
      - scatter plots
      - Box plots
      - Etc.
- Consistent Hypothesis
  - Gives correct answer for the data point

The hypothesis space (H)



Finding hypotheses to fit data. **Top row**: four plots of best-fit functions from four different hypothesis spaces trained on data set 1. **Bottom row**: the same four functions, but trained on a slightly different data set (sampled from the same $f(x)$ function).

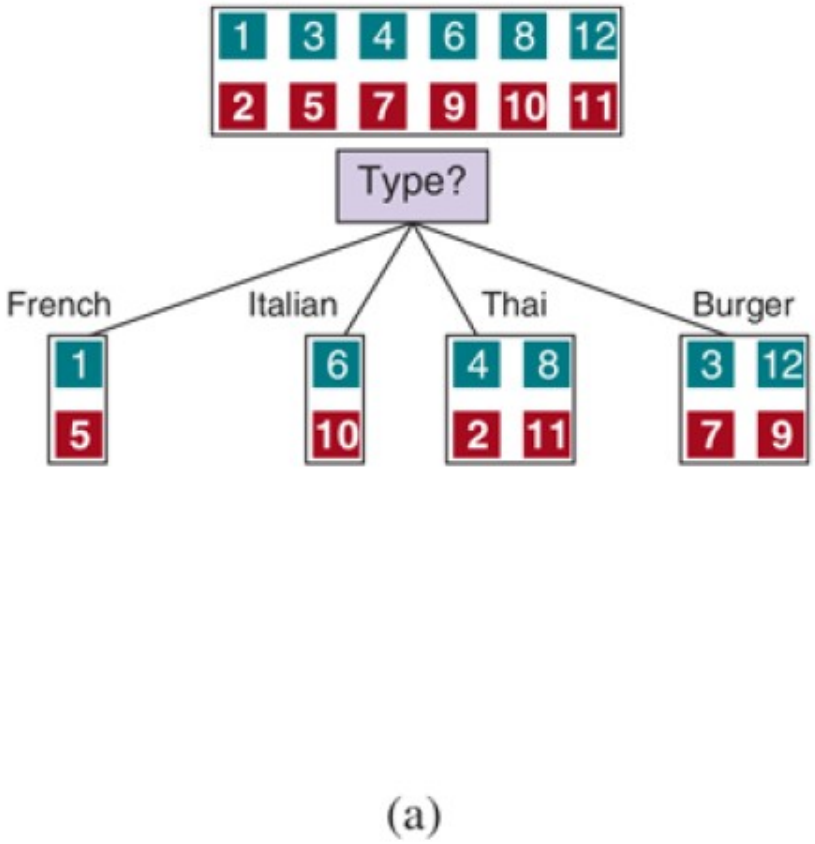
Example: Will we wait for a table?

1. **ALTERNATE:** whether there is a suitable alternative restaurant nearby.
2. **BAR:** whether the restaurant has a comfortable bar area to wait in.
3. **FRI/SAT:** true on Fridays and Saturdays.
4. **HUNGRY:** whether we are hungry right now.
5. **PATRONS:** how many people are in the restaurant (values are *None*, *Some*, and *Full*).
6. **PRICE:** the restaurant's price range (\$, \$\$, \$\$\$).
7. **RAINING:** whether it is raining outside.
8. **RESERVATION:** whether we made a reservation.
9. **TYPE:** the kind of restaurant (French, Italian, Thai, or burger).
10. **WAITESTIMATE:** host's wait estimate: 0 – 10, 10 – 30, 30 – 60, or >60 minutes.

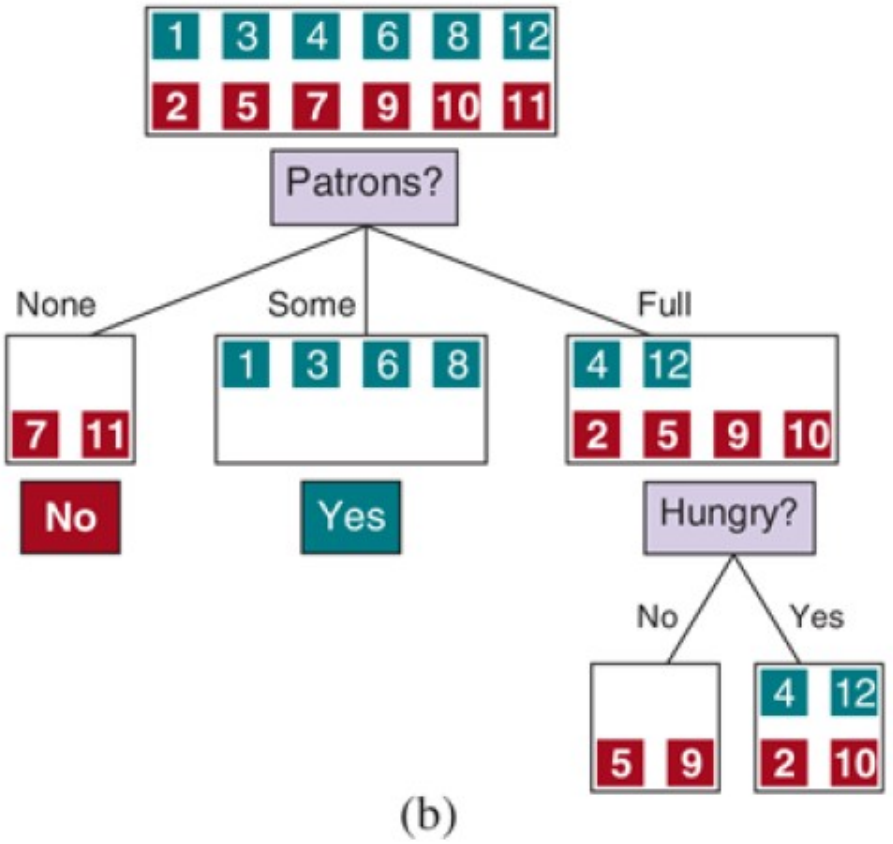
Will we wait for a table?

Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

How to select the best attribute?



(a)



(b)

Choosing attribute tests

Entropy: $$H(V) = \sum_k P(v_k) \log_2 \frac{1}{P(v_k)} = -\sum_k P(v_k) \log_2 P(v_k).$$

B(q) is the entropy of a Boolean random variable that is true with probability q

$$B(q) = -(q \log_2 q + (1-q) \log_2 (1-q)).$$

The entropy of the output variable on the whole set

$$H(Output) = B\left(\frac{p}{p+n}\right).$$

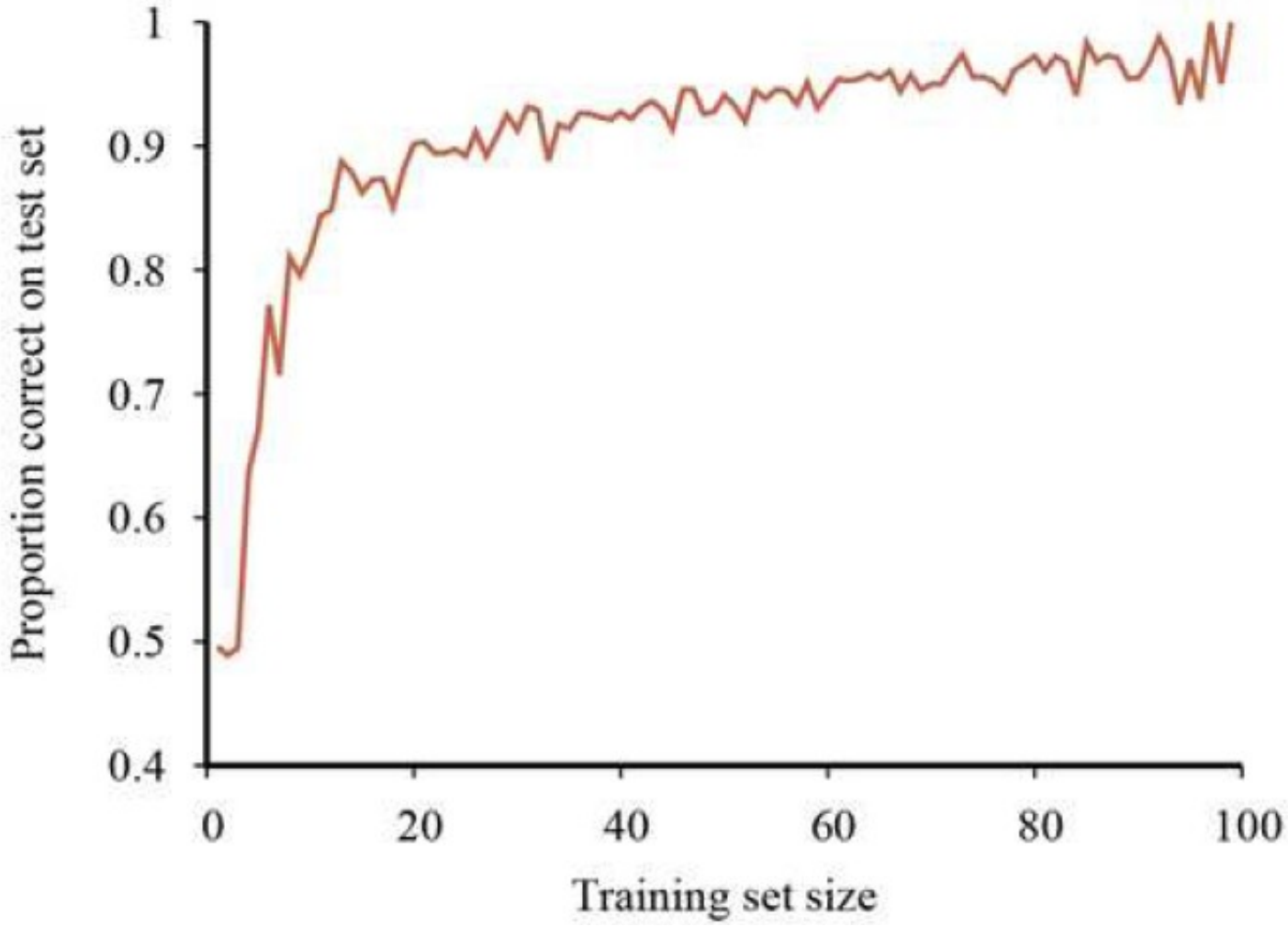
Choosing attribute tests

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p + n} B\left(\frac{p_k}{p_k + n_k}\right).$$

Information Gain

$$Gain(A) = B\left(\frac{p}{p+n}\right) - Remainder(A).$$

The learning curve



Confusion Matrix

Predicted Class

Actual Class

	WillWait = True	WillWait = False
WillWait = True	4 (TP)	2 (FN)
WillWait = False	3 (FP)	3 (TN)

https://en.wikipedia.org/wiki/Confusion_matrix

Performance Metrics

- Accuracy,
- Sensitivity (Recall)
- Specificity
- Precision
- F1-score
- MCC
- Etc.

Useful Resources

- Learning From Data (MOOC @ CALTECH)
  - Prof. Yaser Abu-Mostafa
- Machine Learning (Coursera Course)
  - Prof. Andrew Ng
- Essence of Linear Algebra
  - Grant Sanderson (3Blue1Brown)
- Practice and Compete at Kaggle
- Learn Python