Applied Analytics and Predictive Modeling

Spring 2020

Lecture-6

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Some of the slides adapted from Intro to Data Mining Tan et al. 2nd edition

Today’s agenda

- Building a decision tree
- Case Study presentations
- KNN algorithm
- Project description

Decision Tree

Computing Information Gain after Splitting

- Information Gain

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions; n_i is number of records in partition i

- Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3 and C4.5 decision tree algorithms

Class exercise

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Example from Han & Kamber Data Mining: Concepts and Techniques

Attribute Selection by Information Gain Computation

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"
- I(p, n) = I(9, 5) =0.940
- Compute the entropy for *age*:

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
30…40	4	0	0
>40	3	2	0.971

$$E(age) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = I(p,n) - E(age) = 0.246$$

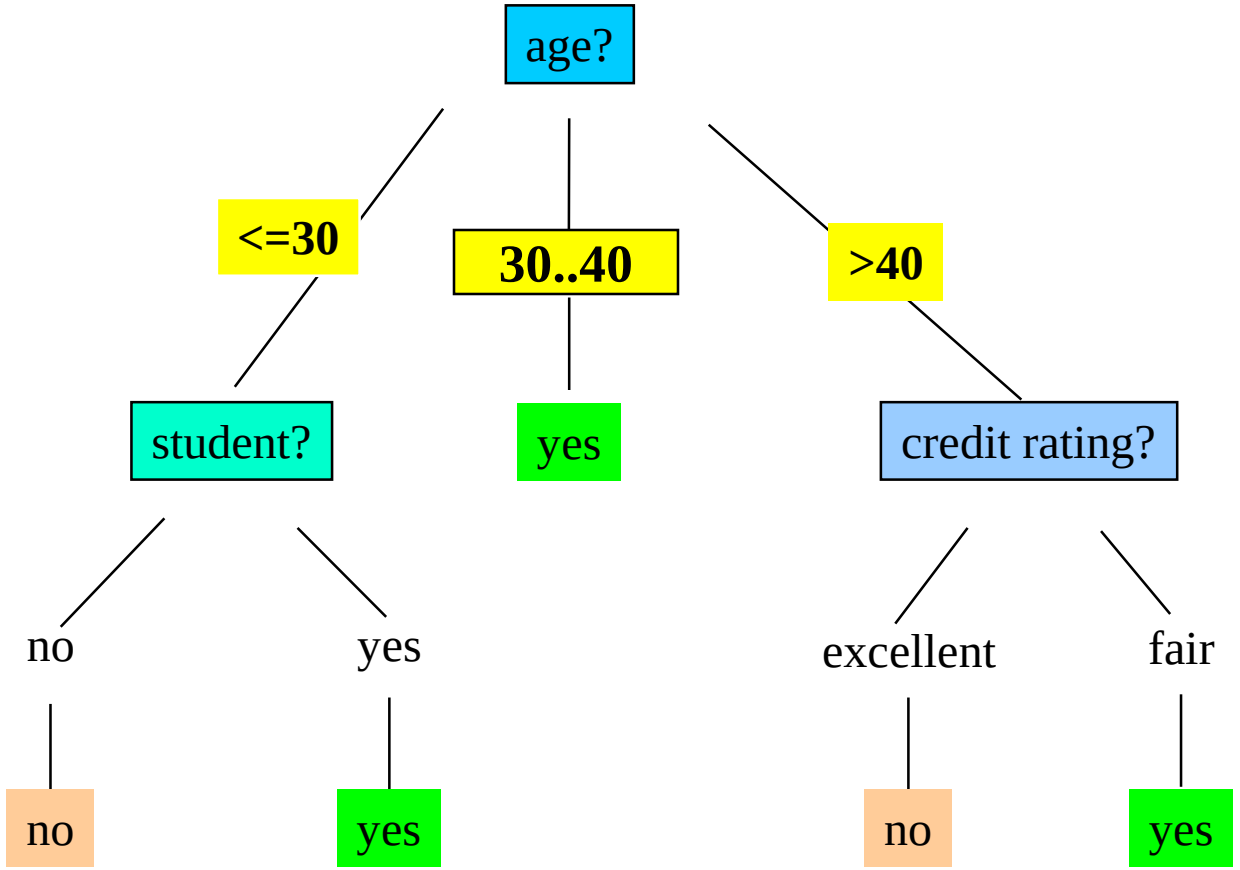
Similarly,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

Output: A Decision Tree for "*buys_computer*"



Case Study-2 Presentations

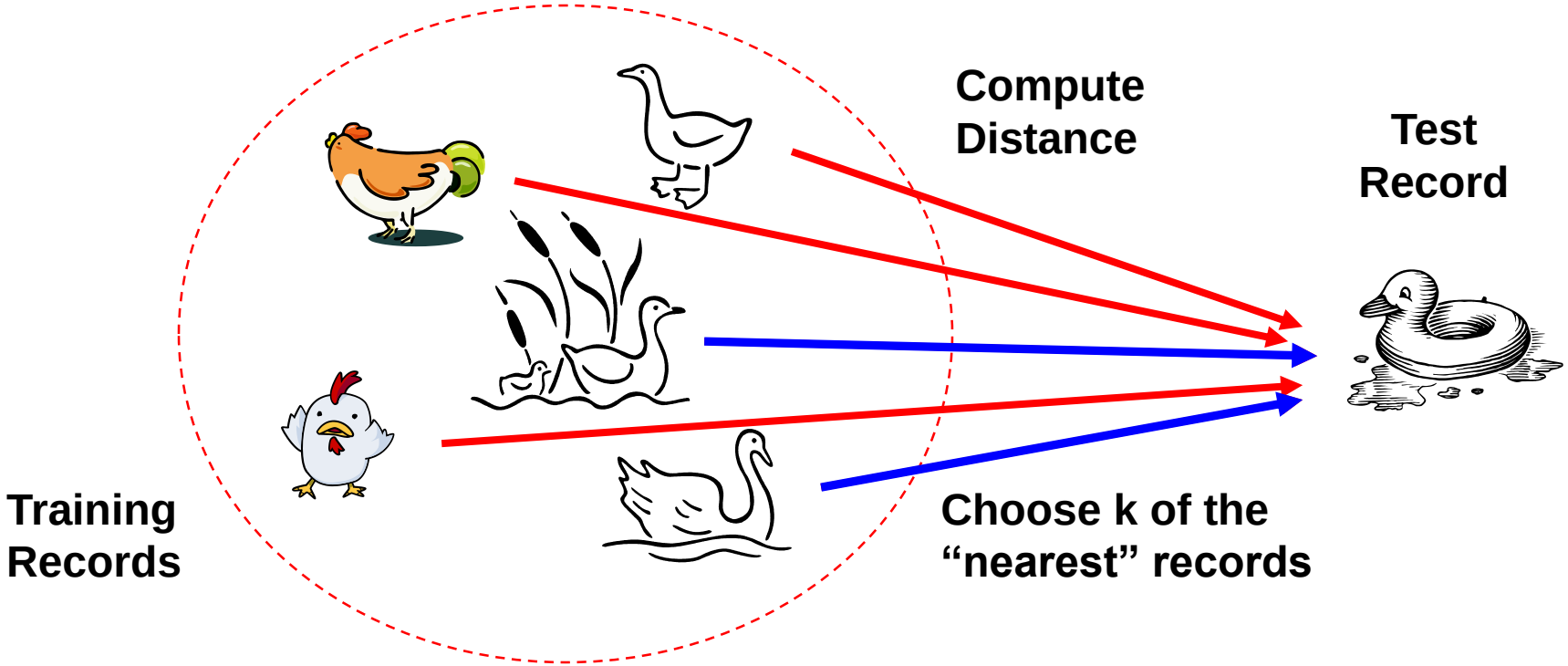
Groups 3,4,9

K-Nearest Neighbor Algorithm

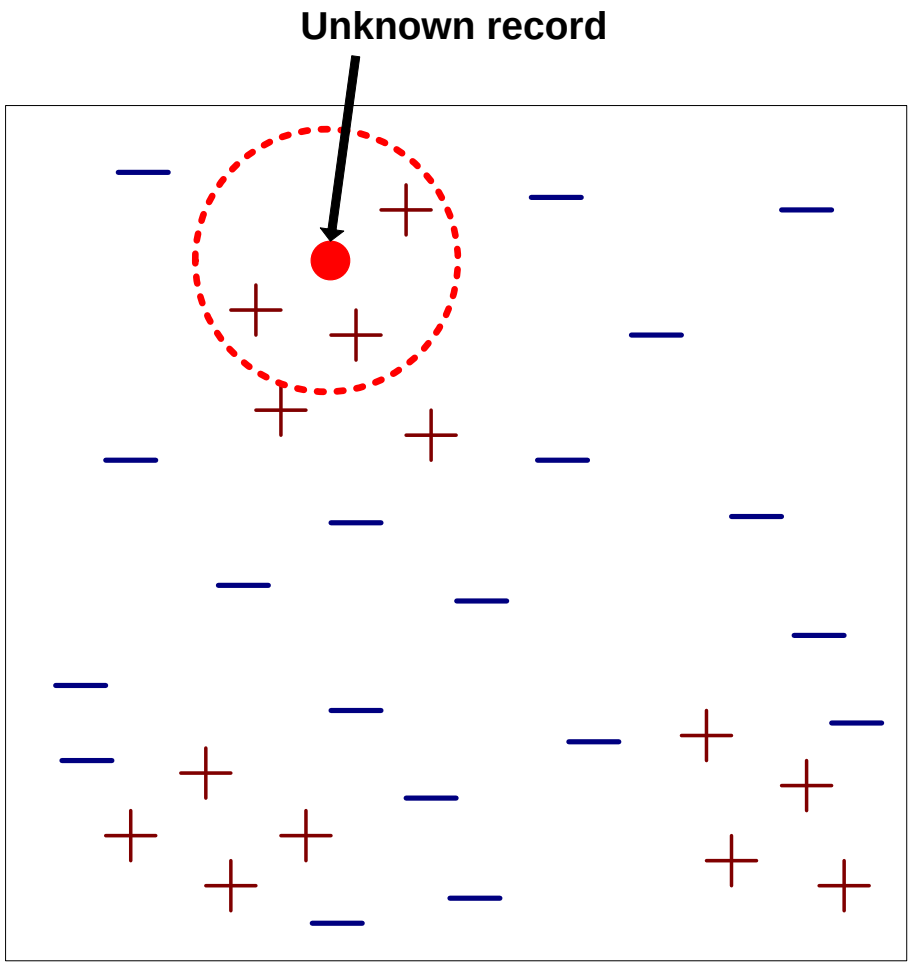
Adapted from Intro to Data Mining, Tan et al., 2nd edition.

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest-Neighbor Classifiers



- Requires three things
  - The set of labeled records
  - Distance metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Nearest Neighbor Classification

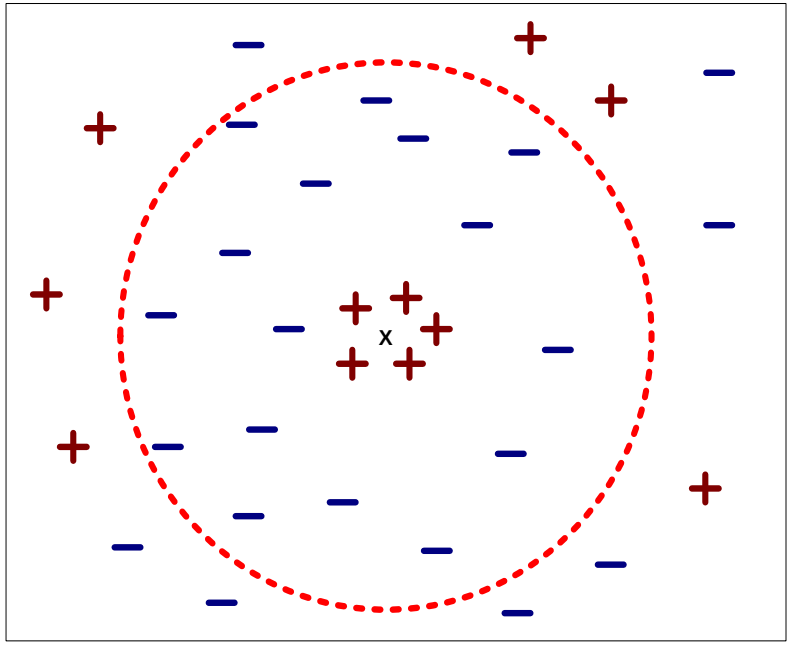
- Compute proximity between two points:
  - Example: Euclidean distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_i (\boldsymbol{x_i} - \boldsymbol{y_i})^2}$$

- Determine the class from nearest neighbor list
  - Take the majority vote of class labels among the k-nearest neighbors
  - Weight the vote according to distance
    - weight factor, $w = 1/d^2$

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



Nearest Neighbor Classification…

- **Choice of proximity measure matters**
  - For documents, cosine is better than correlation or Euclidean

1 1 1 1 1 1 1 1 1 1 1 0		0 0 0 0 0 0 0 0 0 0 0 1
	vs	
0 1 1 1 1 1 1 1 1 1 1 1		1 0 0 0 0 0 0 0 0 0 0 0

Euclidean distance = 1.4142 for both pairs

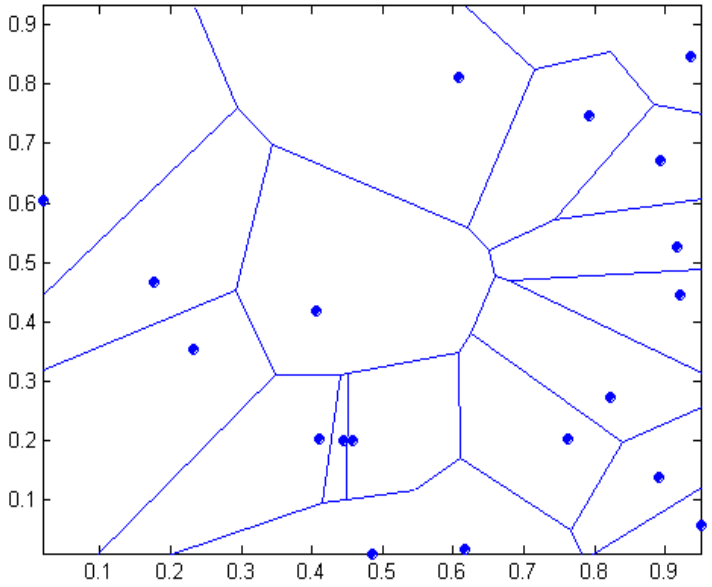
Nearest Neighbor Classification…

- **Data preprocessing is often required**
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
    - Example:
      - height of a person may vary from 1.5m to 1.8m
      - weight of a person may vary from 90lb to 300lb
      - income of a person may vary from $10K to $1M
  - Time series are often standardized to have 0 means a standard deviation of 1

Nearest-neighbor classifiers

- Nearest neighbor classifiers are local classifiers
- They can produce decision boundaries of arbitrary shapes.

1-nn decision boundary is a Voronoi Diagram



Nearest Neighbor Classification…

- **How to handle missing values in training and test sets?**
  - Proximity computations normally require the presence of all attributes
  - Some approaches use the subset of attributes present in two instances
    - This may not produce good results since it effectively uses different proximity measures for each pair of instances
    - Thus, proximities are not comparable

Nearest Neighbor Classification…

- **Handling irrelevant and redundant attributes**
  - Irrelevant attributes add noise to the proximity measure
  - Redundant attributes bias the proximity measure towards certain attributes
  - Can use variable selection or dimensionality reduction to address irrelevant and redundant attributes

Improving KNN Efficiency

- Avoid having to compute distance to all objects in the training set
  - Multi-dimensional access methods (k-d trees)
  - Fast approximate similarity search
  - Locality Sensitive Hashing (LSH)
- Condensing
  - Determine a smaller set of objects that give the same performance
- Editing
  - Remove objects to improve efficiency

Python Notebook

Exercises

Project Description

Dataset

- We are working with goal-oriented subreddits.
- Goal-oriented subreddits – Subreddits that are aimed towards helping users achieve their personal goals.
- In this case all of them are related to health.
  - /r/stopsmoking
  - /r/stopdrinking
  - /r/c25k
  - /r/loseit

Tasks to-do

- Phase-1 (due: 03/23/2020):
  1. Choose the data atleast from 2 subreddits – You can select all the data from 4 subreddits or choose 2 or 3 subreddits (depends on the task you are planning to do)
  2. Conduct preliminary analysis
     1. How are these subreddits different from each other in terms of posting behavior, post size, engagement level (if you can measure), types of content being posted
     2. How many unique users are posting and how are the users behaving in terms of average number of posts they are making on each subreddit, how long of posts they are making, how long did they join the group (just count from when they made a 1st post in the subreddit), etc.
     3. Present any visualizations related to this or any other aspects that you can think of
  3. Are there any overlapping set of users? If so, how is their behavior across these subreddits?
  4. What are the 1 task (or max 2 tasks if you are interested) that you want to do? Provide a brief outline and you don't have to finish solving this task for this deadline

Tasks to-do

- What are the tasks that you can do with this data?

1. Build a prediction algorithm to detect truthful vs fake posts/users
2. Classifying a user as an expert (who already achieved the goal but hanging around the subreddit to support other users) or a non-expert (someone who is working towards the goal)
3. Build a classifier to detect hate speech or trolls on these forums, if any.
4. Can you use these posts to build a rough financial portfolio of users?
5. Can you think of any unsupervised learning approach to provide aggregate analysis of various aspects latently present in the dataset?
6. Or can you use any pretrained classifiers to see if this data is biased in any way?
7. [Anything else that you can think of and share with me before you start working on it seriously..]

Tasks to-do

- Phase-2 (due: 04/13 and 04/20 during project presentations for feedback; Final report due on 04/27):
  1. Based on the classification model you are building, present all the details on why you chose this task; why is this an important aspect with regard to this dataset; how is your solution helping solve the task at hand?
  2. Clearly define your task and create a dummy example to show details.
  3. For classification purposes, since there is no ground truth data given here, you are expected to build the training data and testing data. Your training data doesn't have to be perfect but try your best.

If you don’t like this project, here is an option

- https://sites.google.com/view/icwsm2020datachallenge/home
- Working towards a publication