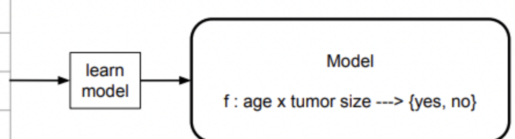
CAS CS 506
Lec 11

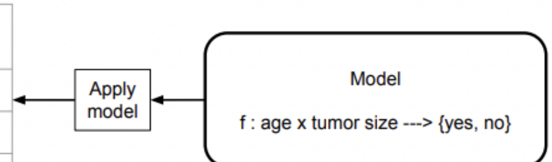
Classification

1. Classification Tasks
    a. Predicting tumor cells as benign or malignant
    b. Classifying images
    c. Classifying credit card transactions as being legitimate or fraudulent
    d. Many more
2. Classification Techniques
    a. Instance-Based classification
    b. Decision trees
    c. Naive bayes
    d. Support Vector Machines
    e. Neural networks
3. What is classification?
    a. Given a training set where data is labeled with a special attribute called a class (a discrete value)
    b. We want to find a model describing how the class attribute varies as a function of the values of the other attributes
    c. Goal: use this model on unlabeled data to assign a class as accurately as possible
4. Example
    a.

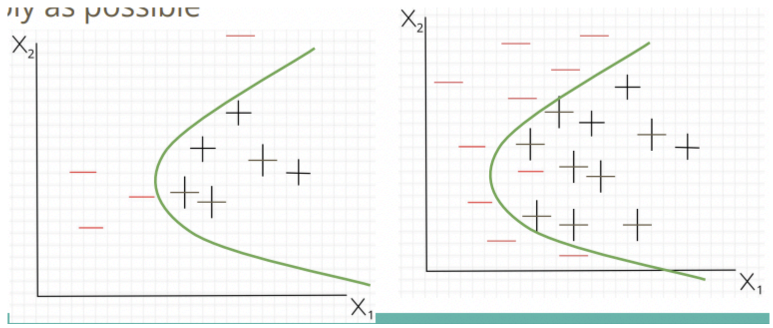
age	Tumor size	malignant?
20	10	no
30	15	yes
40	20	no
50	25	yes

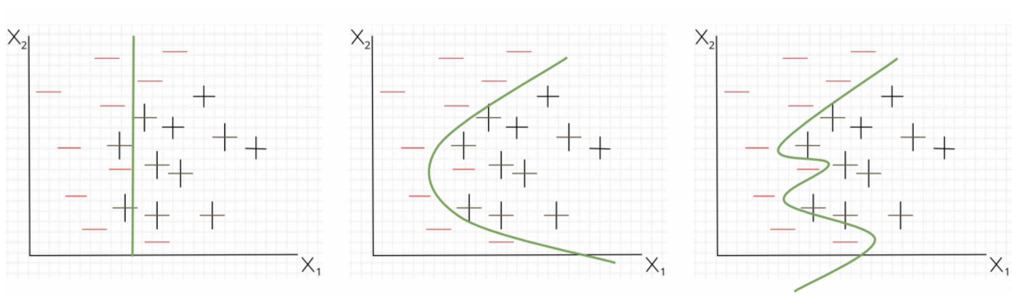
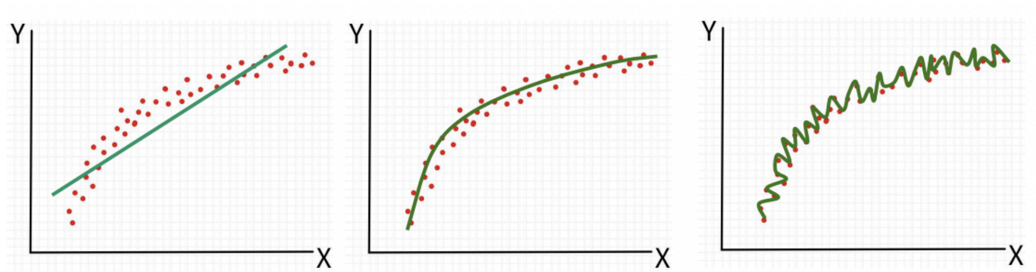


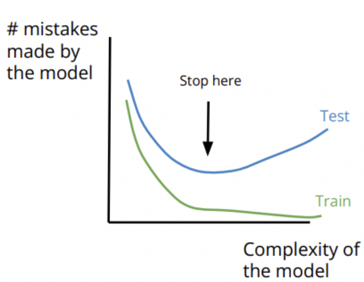
age	Tumor size	malignant?
25	5	?
35	10	?
45	25	?



5. Modeling Philosophy
    a. What constitutes a good feature?
    b. What constitutes a good set of features?
        i. Change in $F_1, \ldots, F_m$ means expect a change in Y
    c. Correlation vs causation

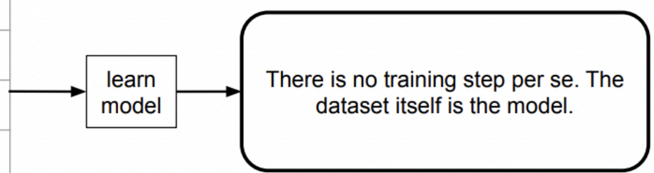
d. Primary goal is to capture the general trend / relationship between class and features as simply as possible
    i. Outliers
    ii. Noise
    iii. 
e. Model performance / evaluations
    i. Overfitting vs underfitting
f. All models are wrong but some are useful. What value does your model provide?

6. Underfitting vs Overfitting
    a. 
    b. 

7. Model Evaluation (simply)
    a. Evaluating a model on the data it was trained on is cheating - can just memorize
    b. Distinction between data used for training and data left out used for testing / evaluation
    c. 

8. Instance-Based Classifiers
    a. Use the stored training records to predict the class label of unseen cases
    b. Rote-learners
        i. Preform classification only if the attributes of the unseen exactly match a record in our training set
9. Instance-Based Classifiers: Training Step
    a.

age	Tumor size	malignant?
20	10	no
30	15	yes
40	20	no
50	25	yes

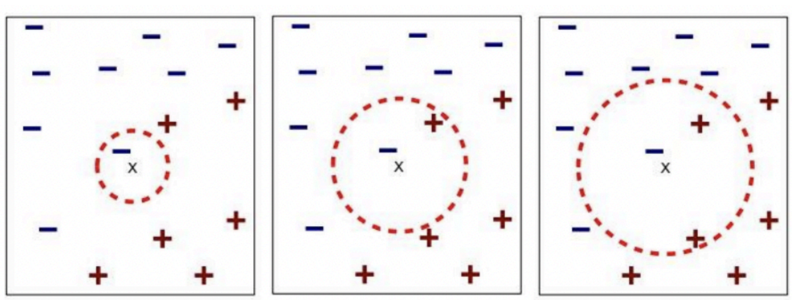


10. Instance-Based Classifiers: Applying the Model
    a.

age	Tumor size	malignant?
20	10	no
30	15	yes
40	20	no
50	25	yes

age	Tumor size	malignant?
20	10	no

11. K Nearest Neighbor Classifier
    a. Requires
        i. Training set
        ii. Distance function
        iii. Value for k
    b. How to classify an unseen record
        i. Compute distance of unseen record to all training records
        ii. Identify the k nearest neighbors
        iii. Aggregate the labels of these k neighbors to predict the unseen record class (ex: majority rule)
    c.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

d. Aggregation methods
    i. Majority rule
    ii. Weighted majority based on distance (w = 1/d^2)
e. Scaling issues
    i. Attributes should be scaled to prevent distance measures from being dominated by one attribute.
    ii. Example
        1. Age: 0 → 100
        2. Income: 10k → 1 million

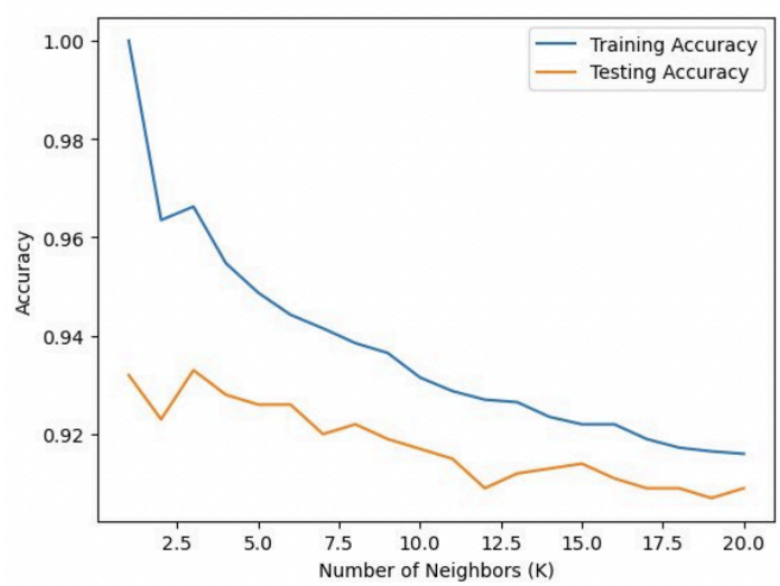
12. Scaling Attributes

a. 

13. K Nearest Neighbor Classifier

a. Choosing the value of k
    i. If k is too small
        1. Sensitive to noise points + overfitting (doesn't generalize well)
    ii. If k is too big
        1. Neighborhood may include points from other classes

14. How to Choose K

a. 

15. K Nearest Neighbor Classifier
    a. Pros
        i. Simple to understand why a given unseen record was given a particular class
    b. Cons
        i. Expensive to classify new points
        ii. KNN can be problematic in high dimensions (curse of dimensionality)