Chapter 5:

k-Nearest Neighbor Algorithm

Supervised vs. Unsupervised
k-Nearest Neighbor Algorithm
Main Questions of the k-NN
k-nearest Neighbor Algorithm for Estimation and Prediction

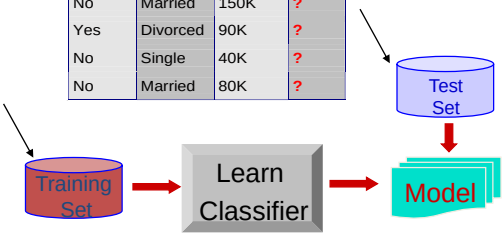
Supervised vs. Unsupervised

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Classification (Review): Supervised & Discrete

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

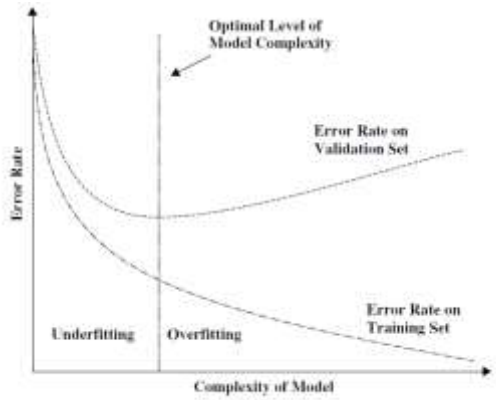


- k-nearest neighbor (Chapter 5)
- Decision tree (Chapter 6)
- Neural network (BP) (Chapter 7)

In supervised approaches, the algorithm needs to guard against "memorizing" the training set => prevent overfitting

Overfitting

- When the model become more complex than the ideal model to fit the training set



The optimal level of model complexity is at the minimum error rate on the validation set.

Division of the Data sets
to handle over-fitting

- **Training set**:
  - A set of examples used for learning, that is to fit the parameters [i.e., weights] of the classifier.
- **Validation set**:
  - A set of examples used to tune the parameters [or complexity] of a classifier, for example to choose the number of hidden units in a neural network or handle over-fitting.
- **Test set (completely unseen data)**:
  - A set of examples used only to assess the performance [generalization] of a fully specified classifier.
- Usually all the data set is divided by 60%-20%-20% or 70%-20%-10%

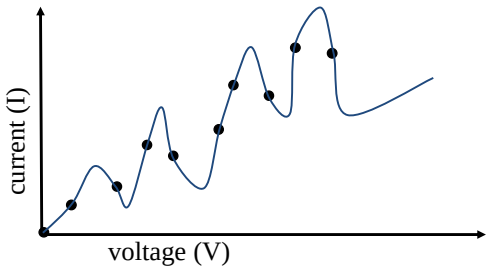
5

Overfitting Example

Testing Ohms Law: V = IR

Experimentally measure 10 points

Fit a curve to the Resulting data.



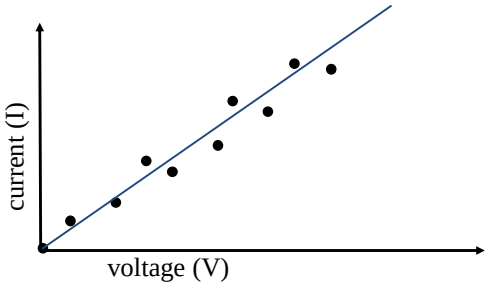
Perfect fit to training data with an 9th degree polynomial
(can fit n points exactly with an n-1 degree polynomial)

Ohm was wrong, we have found a more accurate function!

6

Overfitting Example

Testing Ohms Law: V = IR



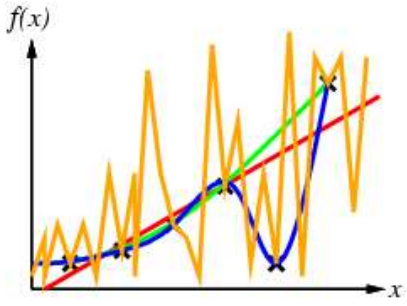
Better generalization with a linear function that fits training data less accurately.

➔ Bias-Variance trade-off

7

Bias-Variance Trade-off

- Learning a function from examples

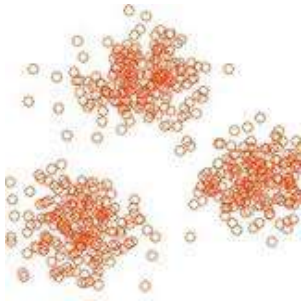


Occam's Razor: **prefer the simplest hypothesis consistent with data**

Clustering (Review): Unsupervised & Discrete

- Similar to the classification, except unsupervistion
- example: Grouping similar instances

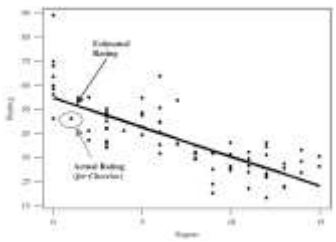
- *k*-means clustering (Chapter 8)
- Kohonen networks (Chapter 9)



9

Regression (Review): Supervised & Continuous

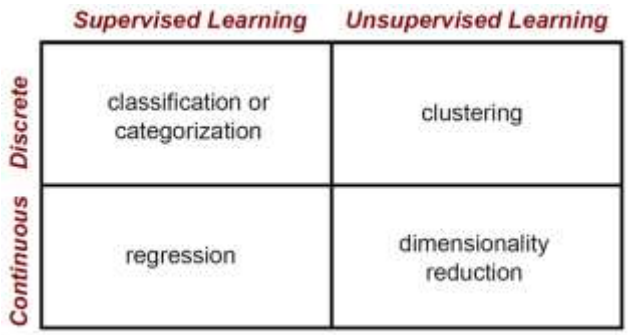
Cereal Name	Sugars	Rating
100% Bran	6	68.4030
100% Natural Bran	8	33.9837
All-Bran	5	59.4255
All-Bran Extra Fiber	0	93.7049
Almond Delight	8	34.3848
Apple Cinnamon Cheerios	10	29.5095
Apple Jacks	14	33.1741
Basic 4	8	37.0386
Bran Chex	6	49.1203
Bran Flakes	5	53.3138
Cap'n'Crunch	12	18.0429
Cheerios	1	50.7650
Cinnamon Toast Crunch	9	19.8236
Clusters	7	40.4002
Cocoa Puffs	13	22.7364



The regression equation is
Rating = 59.4 - 2.42 Sugars

10

Review: Supervised vs. Unsupervised



Dimension Reduction: Image Compression
Algorithms: PCA, ICA, …

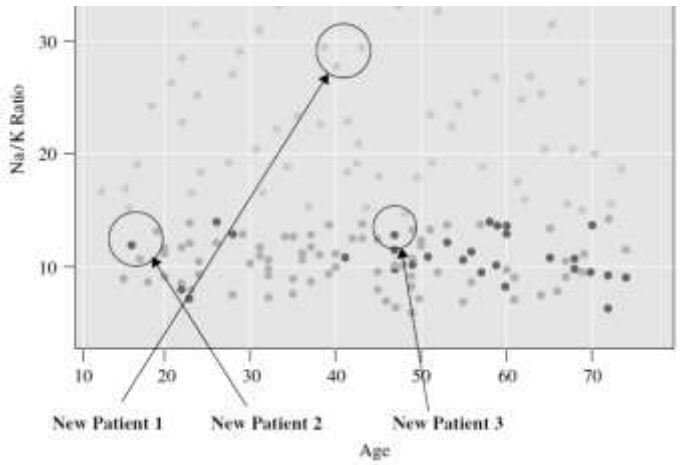
11

k-Nearest Neighbor Algorithm

- An example of *instance-based learning*
- The training data set is stored. A classification for a new unclassified record is simply by comparing to the most similar records in the training set

12

Example: classification of patients for the type of the drug



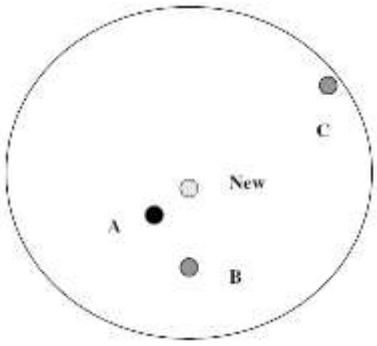
Light gray: drug Y;

Medium gray: drug A or X;

Dark gray: drug B or C.

Suppose: k = 1, new Patient 1: Drug Y

New Patient 2, close up view



- K = 1 ➔ dark gray
- K = 2
  - One dark, one medium
  - No classification!
- K = 3
  - One dark, two medium
  - Voting ➔ medium

Main questions of the k-NN algorithm

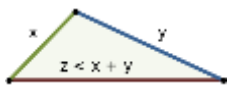
- How do we measure distance?
- How do we combine the information from more than one observation?
  - Should all points be weighted equally, or should some points have more influence than others?
- How many neighbors should we consider? what is *k*?

15

Distance Metric

- Three principles:

1. $d(x,y) \geq 0$, and $d(x,y) = 0$ if and only if $x = y$
2. $d(x,y) = d(y,x)$
3. $d(x,z) \leq d(x,y) + d(y,z)$ *triangle inequality*



16

Distance Metric

- Numeric variable:
  - Euclidean distance

$$d_{\text{Euclidean}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Categorical variable:
  - "different from" distance

$$\text{different}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases}$$

17

Distance metric: Example

- Which patient is more similar to a 50-year-old male:
  - a 20-year-old male, or
  - a 50-year-old female?
- Without normalization
- Combining Euclidean with the "different from" measures

Patient	Age	Gender
A	50	Male
B	20	Male
C	50	Female

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

$$d(A,C) = \sqrt{(20-20)^2 + 1^2} = 1$$

18

Distance metric: Example

MMN: min-max normalization Zscore: Z-Score standardization

Patient	Age	Age_{MMN}	Age_{Zscore}	Gender
A	50	$\frac{50-10}{50}=0.8$	$\frac{50-45}{15}=0.33$	Male
B	20	$\frac{20-10}{50}=0.2$	$\frac{20-45}{15}=-1.67$	Male
C	50	$\frac{50-10}{50}=0.8$	$\frac{50-45}{15}=0.33$	Female

$$d_{MMN}(A,B)=\sqrt{(0.8-0.2)^2+0^2}=0.6$$

$$d_{MMN}(A,C)=\sqrt{(0.8-0.8)^2+1^2}=1.0$$

$$d_{Zscore}(A,B)=\sqrt{[0.33-(-1.67)]^2+0^2}=2.0$$

$$d_{Zscore}(A,C)=\sqrt{(0.33-0.33)^2+1^2}=1.0$$

when mixing categorical and continuous variables, the min-max normalization may be preferred

19

TOC: Main questions of the k-NN algorithm

- How do we measure distance?
- How do we combine the information from more than one observation?
  - Should all points be weighted equally, or should some points have more influence than others?
- How many neighbors should we consider? what is *k*?

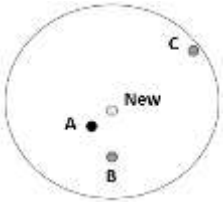
20

Combination Function

- Simple Unweighted Voting
- Weighted Voting
  - the influence of a particular record is inversely proportional to the distance of the record from the new record to be classified.

21

Weighted Voting: Example (3-NN)



Record	Age	Na/K	Age_{MMN}	Na/K_{MMN}
New	17	12.5	0.05	0.25
A (dark gray)	16.8	12.4	0.0467	0.2471
B (medium gray)	17.2	10.5	0.0533	0.1912
C (medium gray)	19.5	13.5	0.0917	0.2794

$$d(new,A) = \sqrt{(0.05 - 0.0467)^2 + (0.25 - 0.2471)^2} = 0.004393$$

$$d(new,B) = \sqrt{(0.05 - 0.0533)^2 + (0.25 - 0.1912)^2} = 0.58893$$

$$d(new,C) = \sqrt{(0.05 - 0.0917)^2 + (0.25 - 0.2794)^2} = 0.051022$$

$$\text{votes (dark gray)} = \frac{1}{d(new,A)^2} = \frac{1}{0.004393^2} \simeq 51,818$$

$$\text{votes (medium gray)} = \frac{1}{d(new,B)^2} + \frac{1}{d(new,C)^2} = \frac{1}{0.058893^2} + \frac{1}{0.051022^2} \simeq 672$$

22

Quantifying Attribute Relevance

- The importance of all the attributes are not equal
- Quantifying the relevance:
  - An expert person
  - Cross-validation approach

23

Quantifying Attribute Relevance: example

- Suppose: The Na/K ratio was determined to be three times as important as age for drug classification.

age Na/K

$$d(\text{new},A) = \sqrt{(0.05-0.0467)^2 + [3(0.25-0.2471)]^2} = 0.009305$$

$$d(\text{new},B) = \sqrt{(0.05-0.0533)^2 + [3(0.25-0.1912)]^2} = 0.17643$$

$$d(\text{new},C) = \sqrt{(0.05-0.0917)^2 + [3(0.25-0.2794)]^2} = 0.09756$$

24

TOC: Main questions of the k-NN algorithm

- How do we measure distance?
- How do we combine the information from more than one observation?
  - Should all points be weighted equally, or should some points have more influence than others?
- How many neighbors should we consider? what is *k*?

25

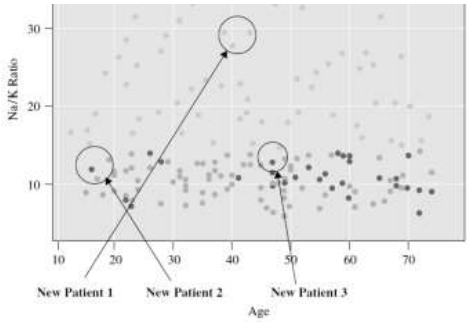
Choosing K

- Too large k:
  - Higher performance (+)
  - Slow computational time (-)
  - Locally interesting behaviors will be neglected (-)
- Too small k:
  - Fast response (+)
  - Considering locally interesting behaviors (+)
  - Classification would be affected by outliers and noise (-)
  - Overfitting (-)
- Ideal k can be determined by trial and error, cross validation

26

Database consideration

- For instance-based learnings (like *k*-NN), a rich balance database is required
- Maintaining a rich database => restrictions on memory space.
  - Solution: retaining the points that are near a classification "boundary."
  - For example, all records with Na/K > 19 could be omitted



27

K-NN for estimation and prediction

- 3-NN for blood pressure estimation

Record	Age	Na/K	BP	Age_{MMN}	Na/K_{MMN}	Distance
New	17	12.5	?	0.05	0.25	—
A	16.8	12.4	120	0.0467	0.2471	0.009305
B	17.2	10.5	122	0.0533	0.1912	0.16783
C	19.5	13.5	130	0.0917	0.2794	0.26737

- Locally weighted averaging of the BP of 3 nearest neighbors can estimate the BP of the new record.
- How?

28

K-NN for estimation and prediction

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{\frac{120}{0.009305^2} + \frac{122}{0.17643^2} + \frac{130}{0.09756^2}}{\frac{1}{0.009305^2} + \frac{1}{0.17643^2} + \frac{1}{0.09756^2}} = 120.0954$$

29