GEOMETRIC VIEW OF DATA

David Kauchak
CS 158 – Spring 2022

1

Admin

Assignment 2

Assignment 1 solution posted under the "Resources" tab on sakai (use them to debug!)

Assignment 1 back soon

Keep reading

Videos?

2

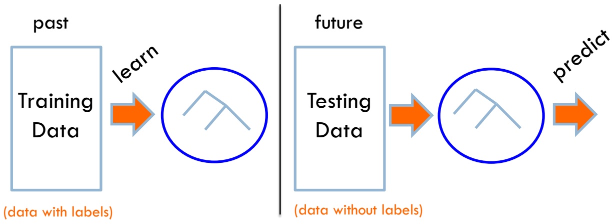
Proper Experimentation

u13007351 fotosearch.com

3

Experimental setup

REAL WORLD USE OF ML ALGORITHMS

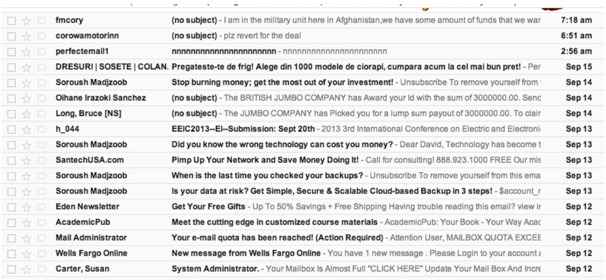


How do we tell how well we're doing?

4

Real-world classification

Google has labeled training data, for example from people clicking the "spam" button, but when new messages come in, they're not labeled



☐	☆	fmcory	**(no subject)** - I am in the military unit here in Afghanistan,we have some amount of funds that we war	7:18 am
☐	☆	corowamotorinn	**(no subject)** - plz revert for the deal	6:51 am
☐	☆	perfectemail1	**nnnnnnnnnnnnnnnnnnnnnn** - nnnnnnnnnnnnnnnnnnnnnn	2:56 am
☐	☆	DRESURI | SOSETE | COLAN.	**Pregateste-te de frig! Alege din 1000 modele de ciorapi, cumpara acum la cel mai bun pret!** - Per	Sep 15
☐	☆	Soroush Madjzoob	**Stop burning money; get the most out of your investment!** - Unsubscribe To remove yourself from	Sep 14
☐	☆	Oihane Irazoki Sanchez	**(no subject)** - The BRITISH JUMBO COMPANY has Award your Id with the sum of 3000000.00. Send	Sep 14
☐	☆	Long, Bruce [NS]	**(no subject)** - The JUMBO COMPANY has Picked you for a lump sum payout of 3000000.00. To clair	Sep 14
☐	☆	h_044	**EEIC2013--EI--Submission: Sept 20th** - 2013 3rd International Conference on Electric and Electroni	Sep 13
☐	☆	Soroush Madjzoob	**Did you know the wrong technology can cost you money?** - Dear David, Technology has become t	Sep 13
☐	☆	SantechUSA.com	**Pimp Up Your Network and Save Money Doing It!** - Call for consulting! 888.923.1000 FREE Our mis	Sep 13
☐	☆	Soroush Madjzoob	**When is the last time you checked your backups?** - Unsubscribe To remove yourself from this ema	Sep 13
☐	☆	Soroush Madjzoob	**Is your data at risk? Get Simple, Secure & Scalable Cloud-based Backup in 3 steps!** - $account_r	Sep 13
☐	☆	Eden Newsletter	**Get Your Free Gifts** - Up To 50% Savings + Free Shipping Having trouble reading this email? view ir	Sep 12
☐	☆	AcademicPub	**Meet the cutting edge in customized course materials** - AcademicPub: Your Book - Your Way Acac	Sep 12
☐	☆	Mail Administrator	**Your e-mail quota has been reached! (Action Required)** - Attention User, MAILBOX QUOTA EXCEE	Sep 12
☐	☆	Wells Fargo Online	**New message from Wells Fargo Online** - You have 1 new message . Please Login to your account	Sep 12
☐	☆	Carter, Susan	**System Administrator.** - Your Mailbox Is Almost Full "CLICK HERE" Update Your Mail Box And Incre	Sep 12

5

Classification evaluation

Labeled data

Data	Label
🟨	0
🟨	0
🟨	1
🟨	1
🟨	0
🟨	1
🟨	0

Use the labeled data we have already to create a test set with known labels!

Why can we do this?

We assume there's an underlying distribution that generates both the training and test examples

6

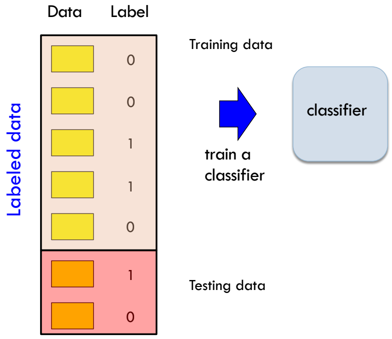
Classification evaluation

Labeled data

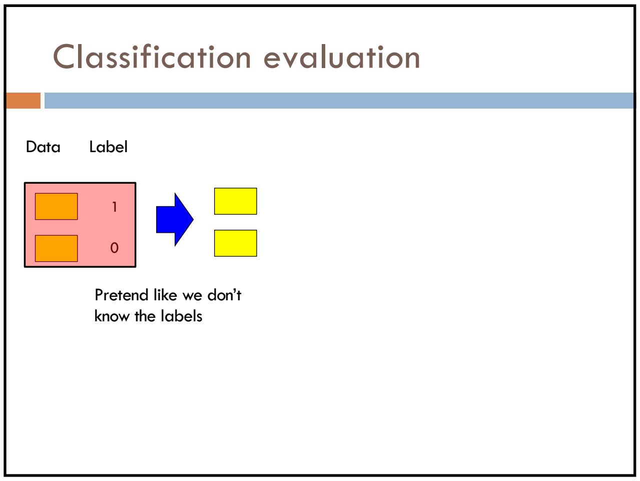
Data	Label	
🟨	0	Training data
🟨	0	
🟨	1	
🟨	1	
🟨	0	
🟧	1	Testing data
🟧	0	

7

Classification evaluation



8



Classification evaluation

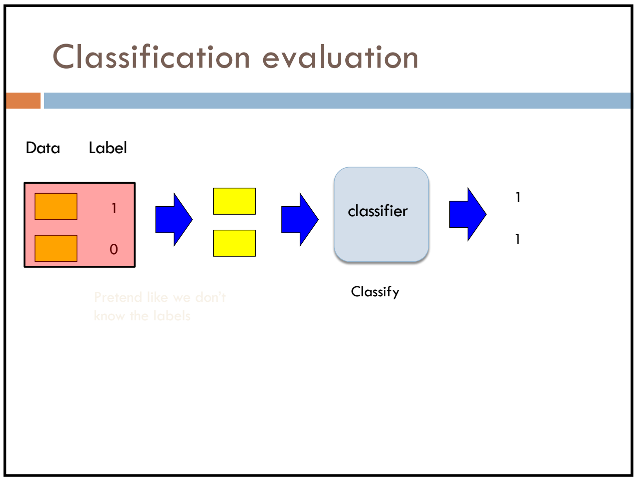
Data Label

1

0

Pretend like we don't know the labels

9



Classification evaluation

Data Label

1

0

classifier

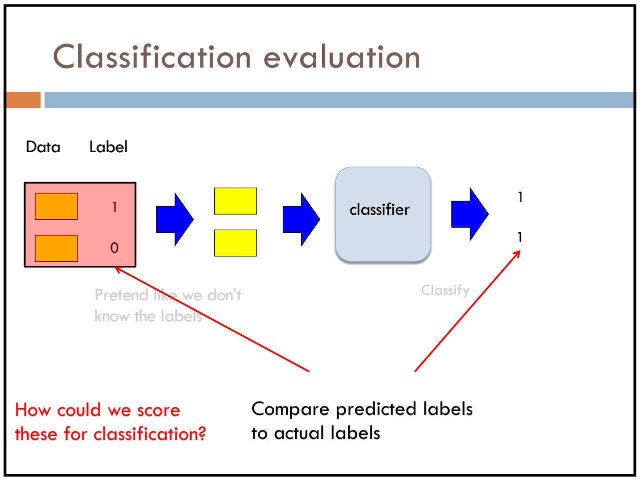
1

1

Pretend like we don't know the labels

Classify

10



Classification evaluation

Data Label

1

0

classifier

1

1

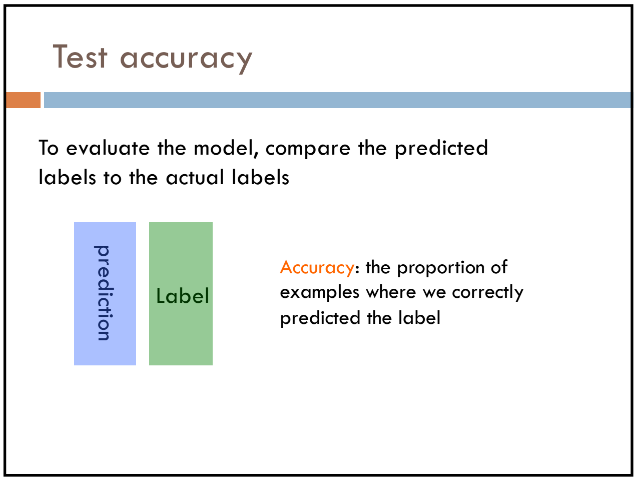
Pretend like we don't know the labels

Classify

How could we score these for classification?

Compare predicted labels to actual labels

11



Test accuracy

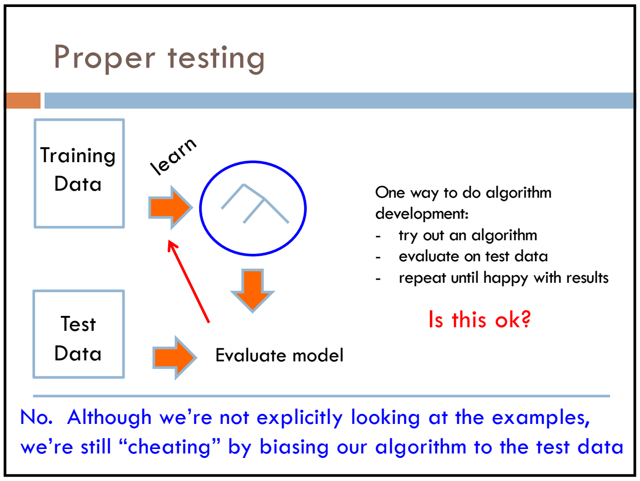
To evaluate the model, compare the predicted labels to the actual labels

prediction

Label

Accuracy: the proportion of examples where we correctly predicted the label

12



Proper testing

Training Data

learn

Test Data

Evaluate model

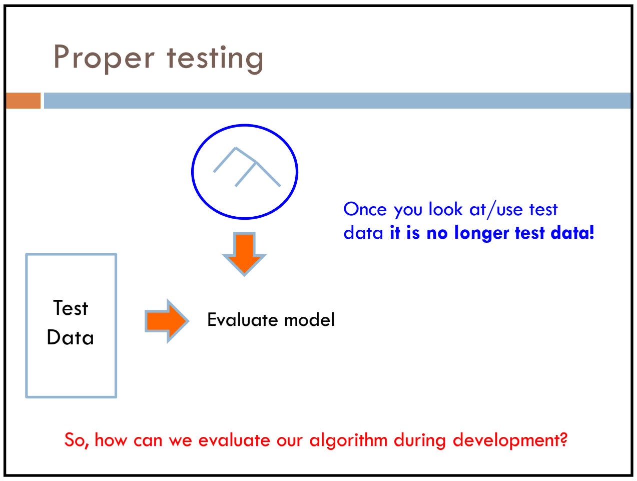
One way to do algorithm development:

- try out an algorithm
- evaluate on test data
- repeat until happy with results

Is this ok?

No. Although we're not explicitly looking at the examples, we're still "cheating" by biasing our algorithm to the test data

13



Proper testing

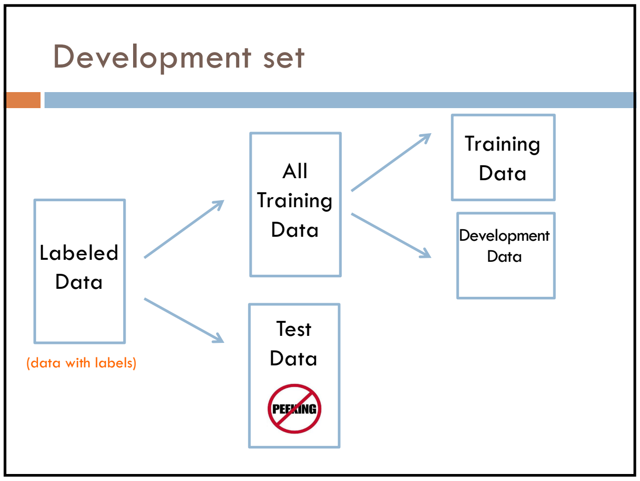
Once you look at/use test data **it is no longer test data!**

Test Data

Evaluate model

So, how can we evaluate our algorithm during development?

14



Development set

Labeled Data

(data with labels)

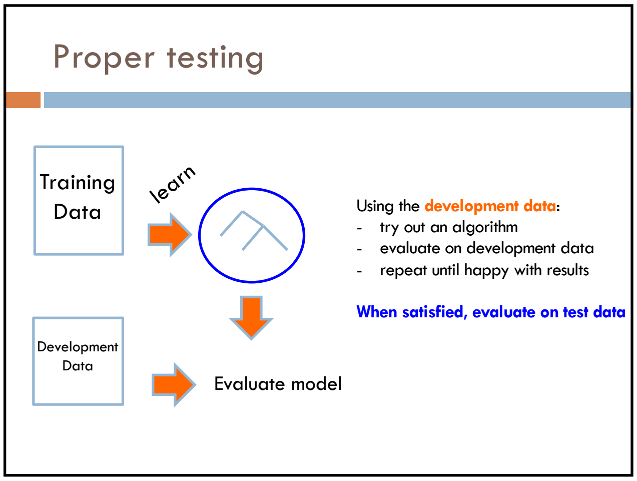
All Training Data

Training Data

Development Data

Test Data

15



Proper testing

Training Data

learn

Development Data

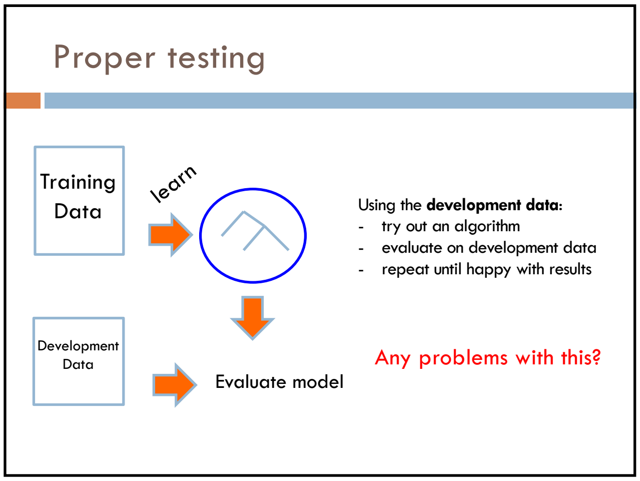
Evaluate model

Using the **development data**:

- try out an algorithm
- evaluate on development data
- repeat until happy with results

When satisfied, evaluate on test data

16



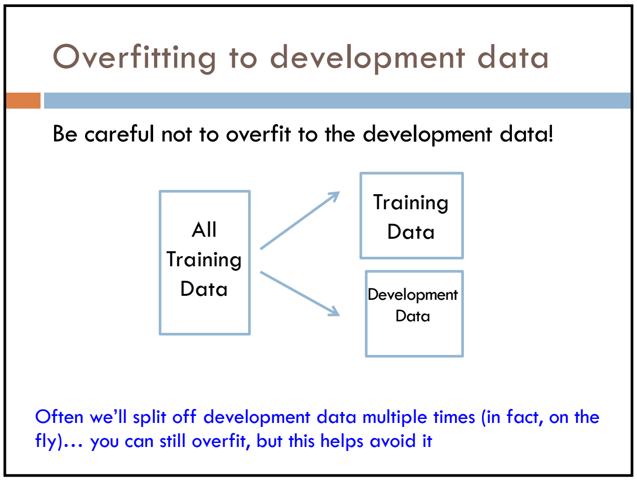
Proper testing

Using the **development data**:

- try out an algorithm
- evaluate on development data
- repeat until happy with results

Any problems with this?

17

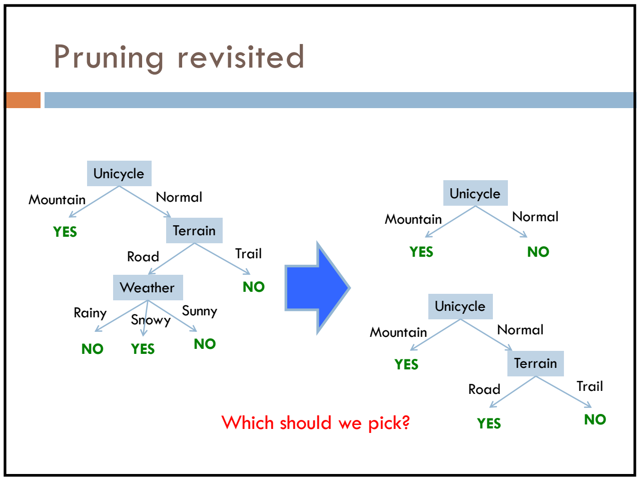


Overfitting to development data

Be careful not to overfit to the development data!

Often we'll split off development data multiple times (in fact, on the fly)… you can still overfit, but this helps avoid it

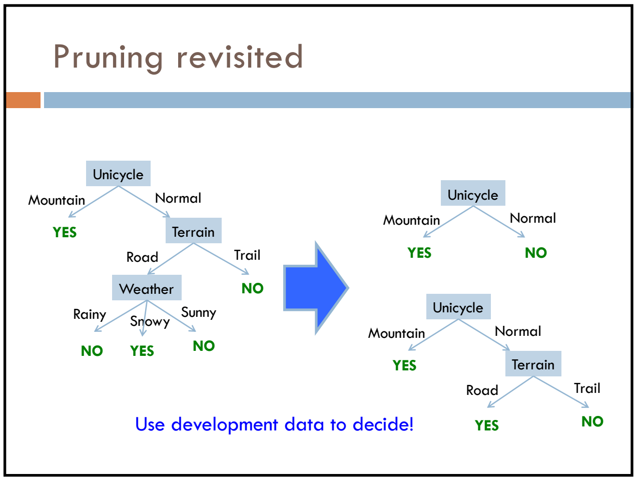
18



Pruning revisited

Which should we pick?

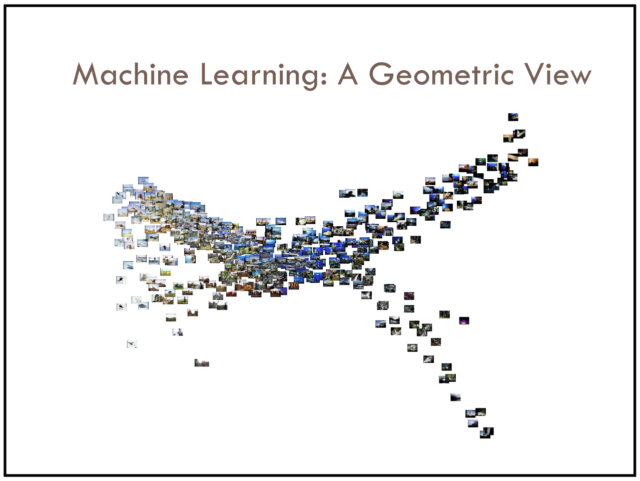
19



Pruning revisited

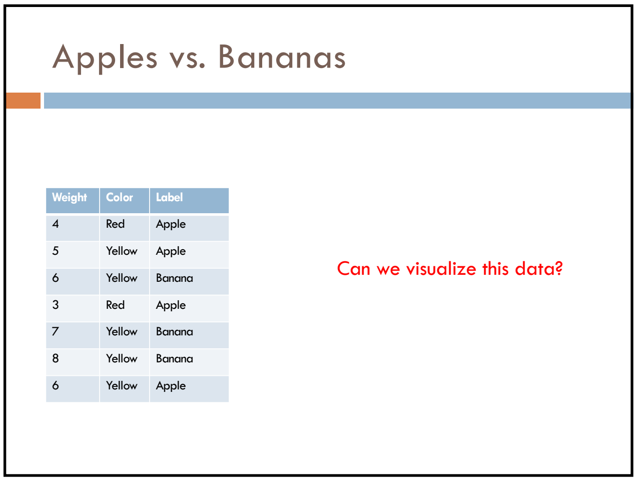
Use development data to decide!

20



Machine Learning: A Geometric View

21

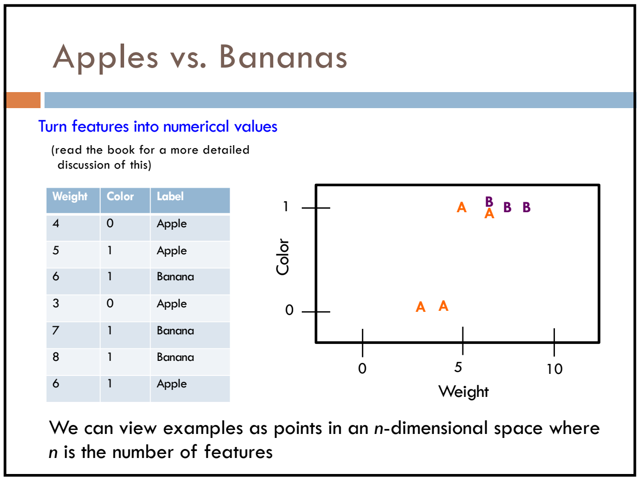


Apples vs. Bananas

Weight	Color	Label
4	Red	Apple
5	Yellow	Apple
6	Yellow	Banana
3	Red	Apple
7	Yellow	Banana
8	Yellow	Banana
6	Yellow	Apple

Can we visualize this data?

22



Apples vs. Bananas

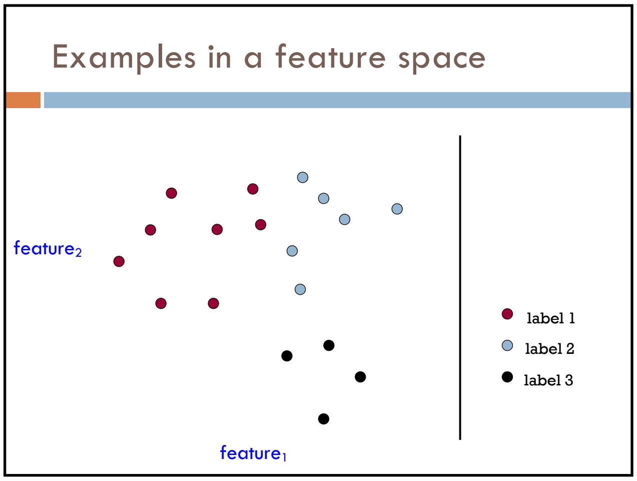
Turn features into numerical values

(read the book for a more detailed discussion of this)

Weight	Color	Label
4	0	Apple
5	1	Apple
6	1	Banana
3	0	Apple
7	1	Banana
8	1	Banana
6	1	Apple

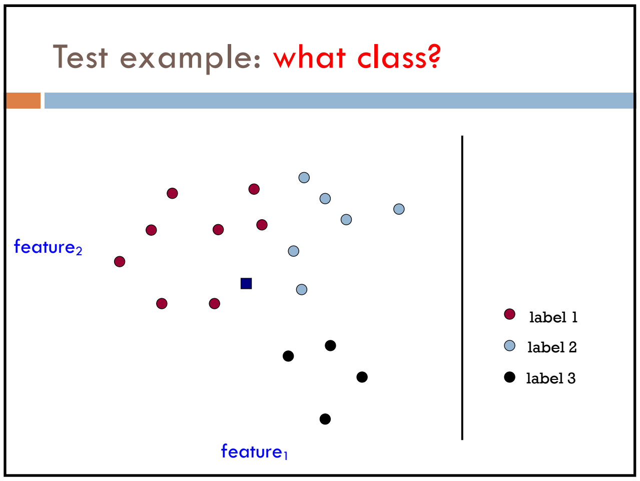
We can view examples as points in an *n*-dimensional space where *n* is the number of features

23



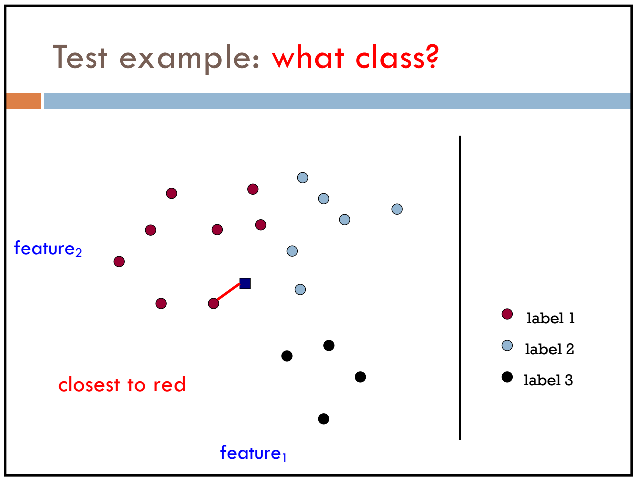
Examples in a feature space

24



Test example: what class?

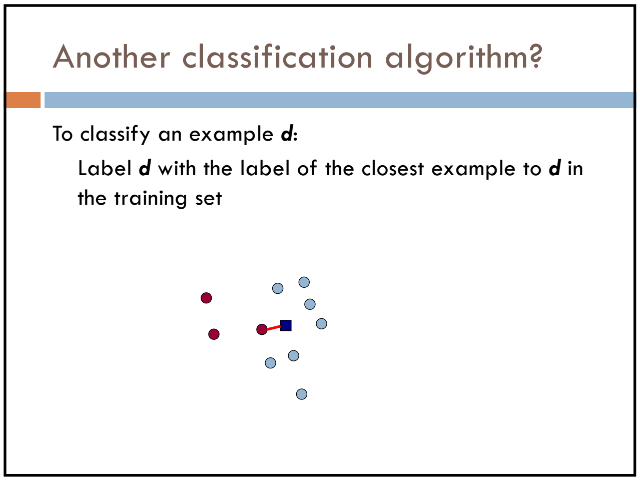
25



Test example: what class?

closest to red

26

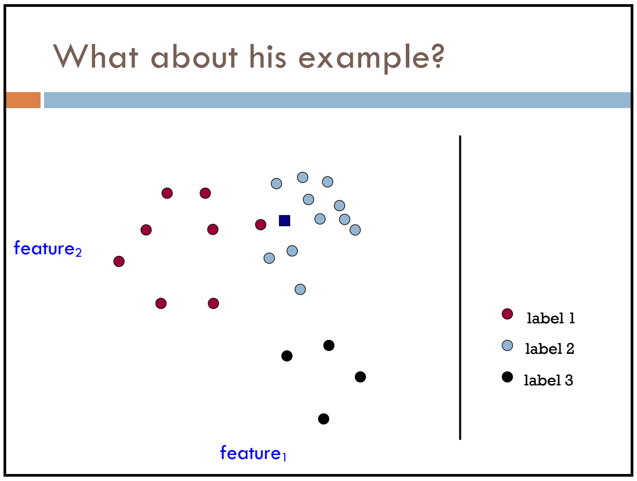


Another classification algorithm?

To classify an example ***d***:

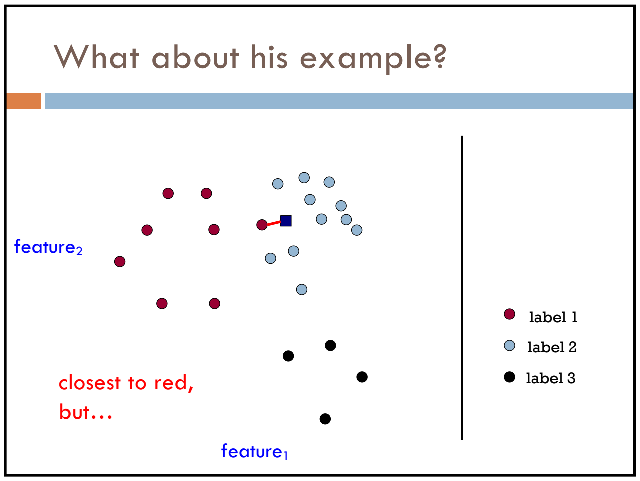
Label ***d*** with the label of the closest example to ***d*** in the training set

27



What about his example?

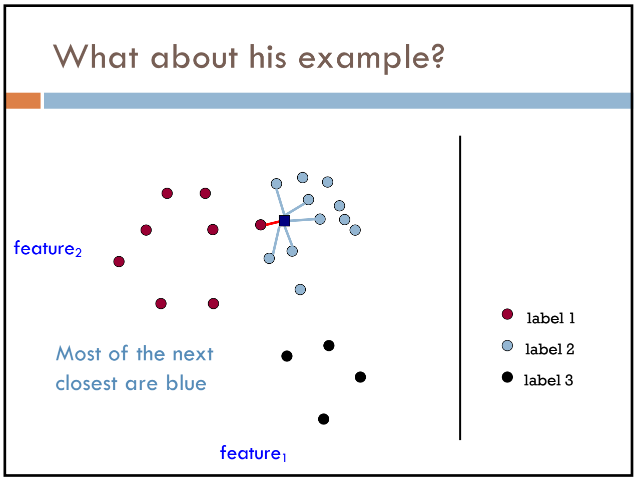
28



What about his example?

closest to red, but…

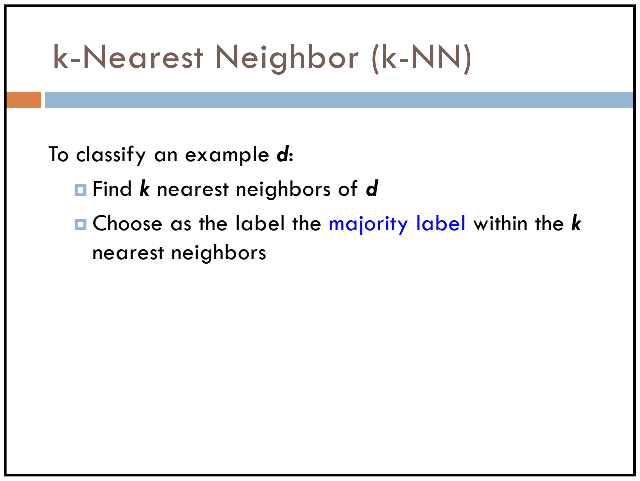
29



What about his example?

Most of the next closest are blue

30

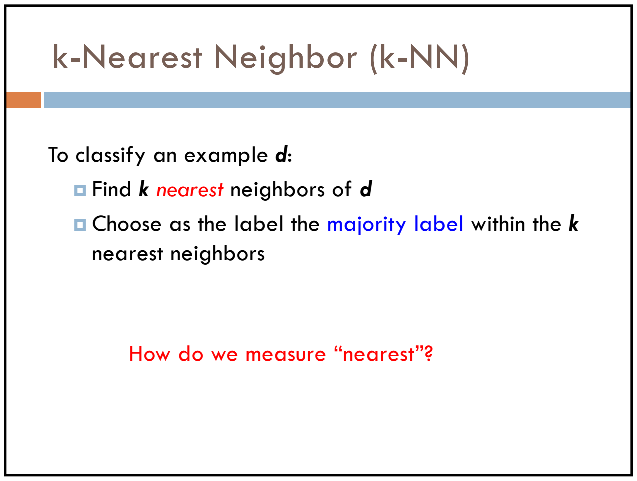


k-Nearest Neighbor (k-NN)

To classify an example ***d***:

- Find ***k*** nearest neighbors of ***d***
- Choose as the label the majority label within the ***k*** nearest neighbors

31



k-Nearest Neighbor (k-NN)

To classify an example ***d***:

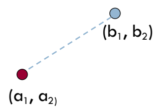
- Find ***k*** *nearest* neighbors of ***d***
- Choose as the label the majority label within the ***k*** nearest neighbors

How do we measure "nearest"?

32

Euclidean distance

In two dimensions, how do we compute the distance?

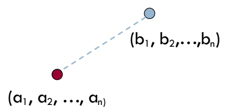


$$D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

33

Euclidean distance

In n-dimensions, how do we compute the distance?

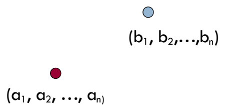


$$D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

34

Euclidean distance

In n-dimensions, how do we compute the distance?

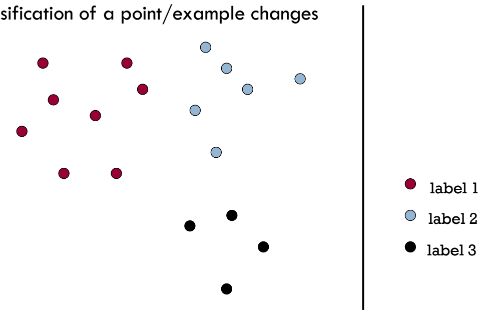


Measuring distance/similarity is a domain-specific problem and there are many, many different variations!

35

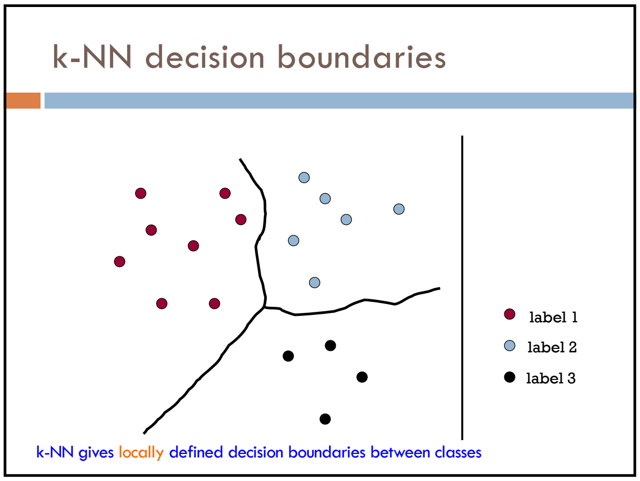
Decision boundaries

The **decision boundaries** are places in the features space where the classification of a point/example changes



Where are the decision boundaries for k-NN?

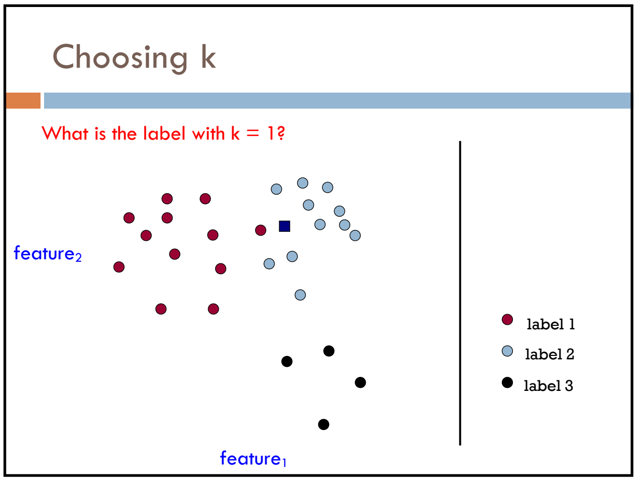
36



k-NN decision boundaries

k-NN gives locally defined decision boundaries between classes

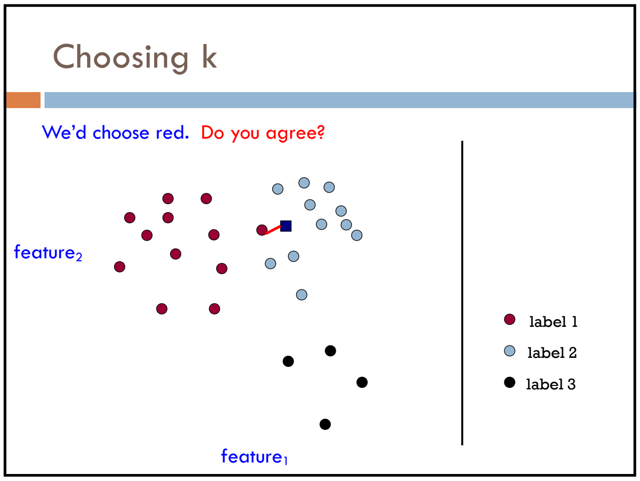
37



Choosing k

What is the label with k = 1?

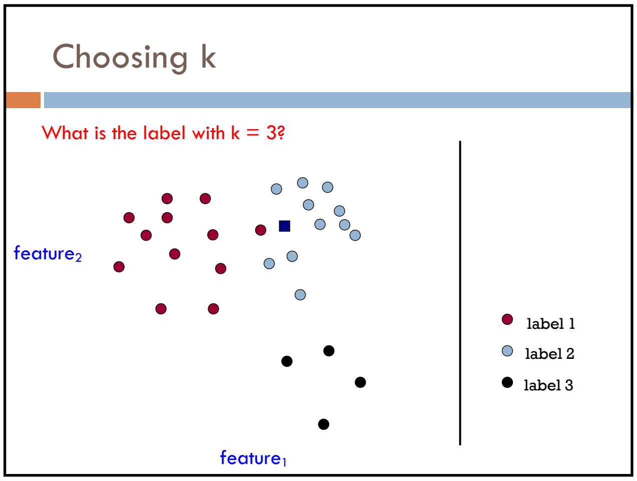
38



Choosing k

We'd choose red. Do you agree?

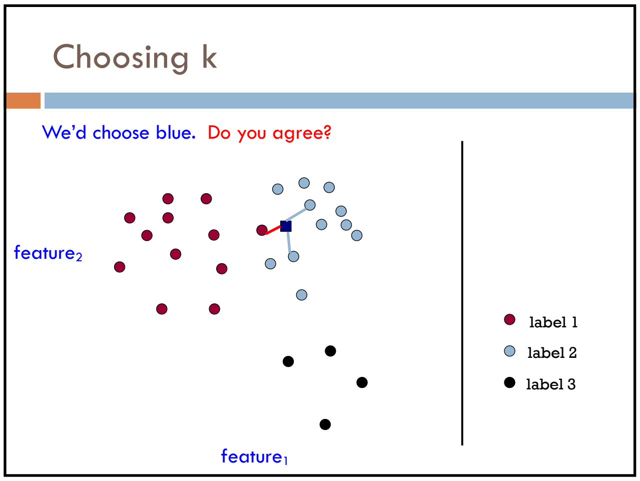
39



Choosing k

What is the label with k = 3?

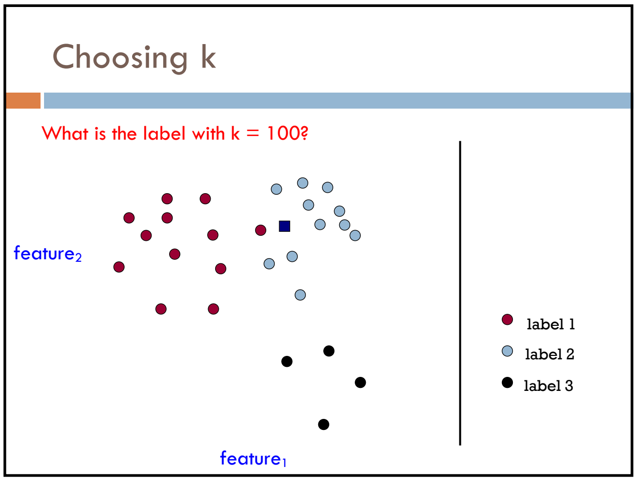
40



Choosing k

We'd choose blue. Do you agree?

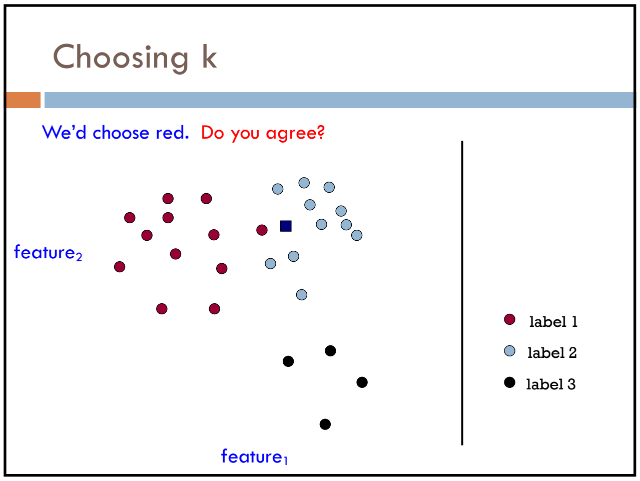
41



Choosing k

What is the label with k = 100?

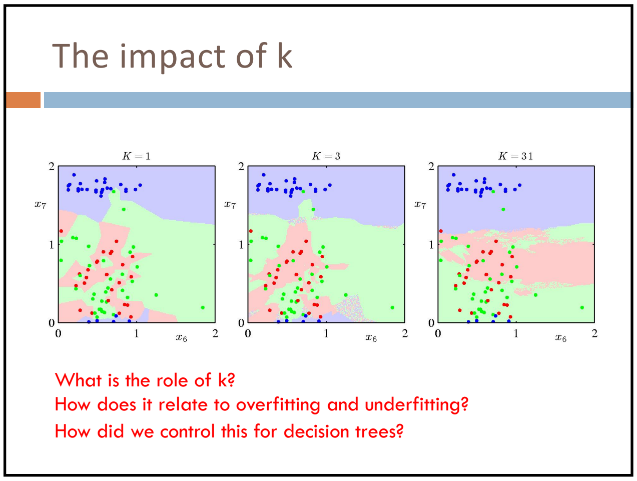
42



Choosing k

We'd choose red. Do you agree?

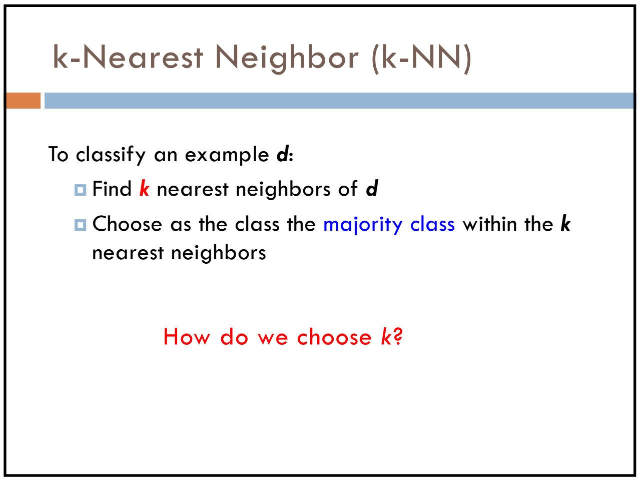
43



The impact of k

What is the role of k?
How does it relate to overfitting and underfitting?
How did we control this for decision trees?

44



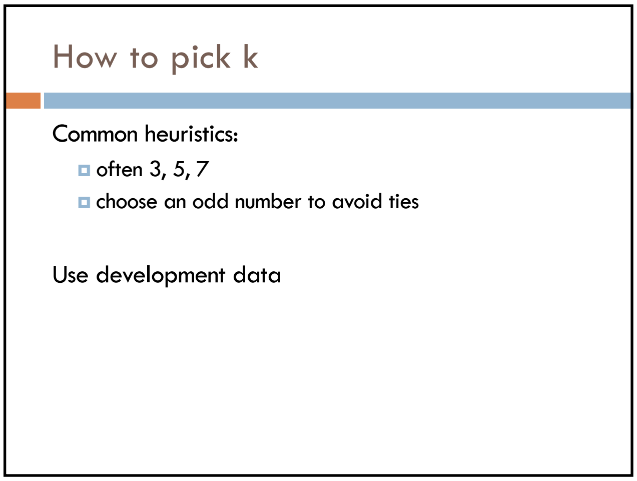
k-Nearest Neighbor (k-NN)

To classify an example ***d***:

- Find ***k*** nearest neighbors of ***d***
- Choose as the class the majority class within the ***k*** nearest neighbors

How do we choose *k*?

45



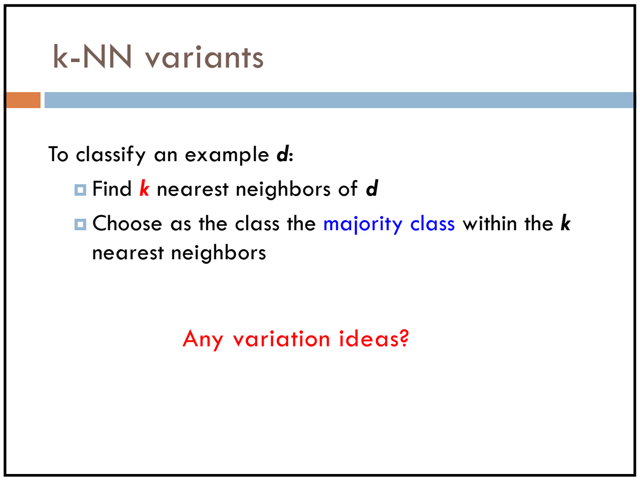
How to pick k

Common heuristics:

- often 3, 5, 7
- choose an odd number to avoid ties

Use development data

46



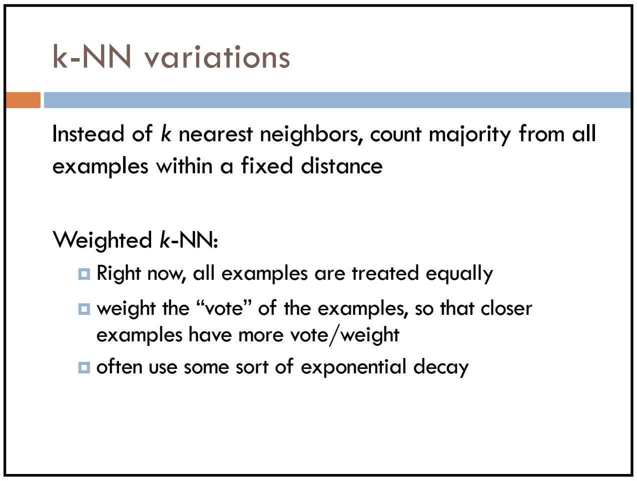
k-NN variants

To classify an example ***d***:

- Find ***k*** nearest neighbors of ***d***
- Choose as the class the majority class within the ***k*** nearest neighbors

Any variation ideas?

47



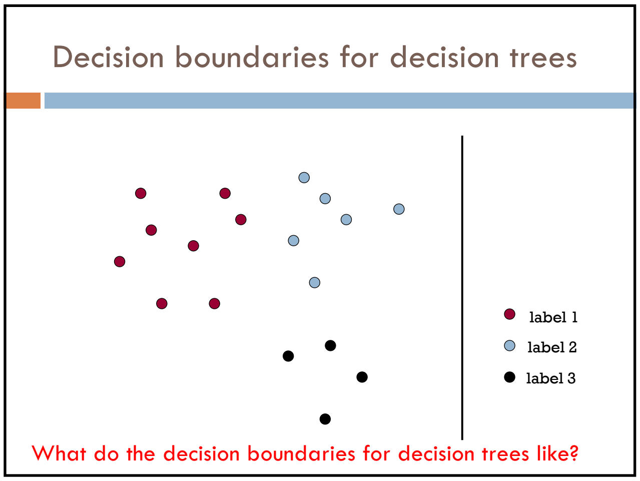
k-NN variations

Instead of *k* nearest neighbors, count majority from all examples within a fixed distance

Weighted *k*-NN:

- Right now, all examples are treated equally
- weight the "vote" of the examples, so that closer examples have more vote/weight
- often use some sort of exponential decay

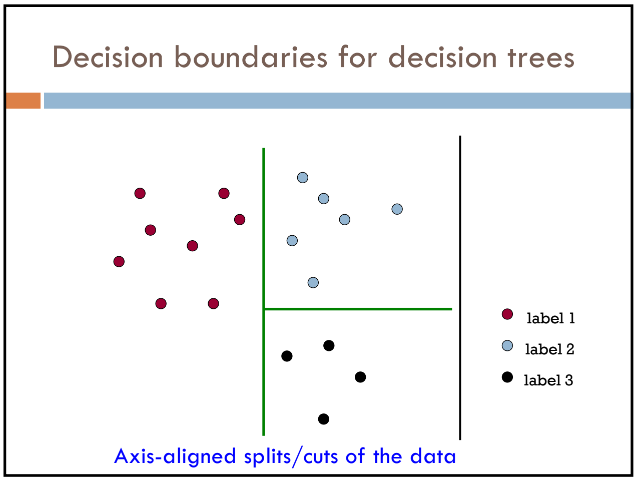
48



Decision boundaries for decision trees

What do the decision boundaries for decision trees like?

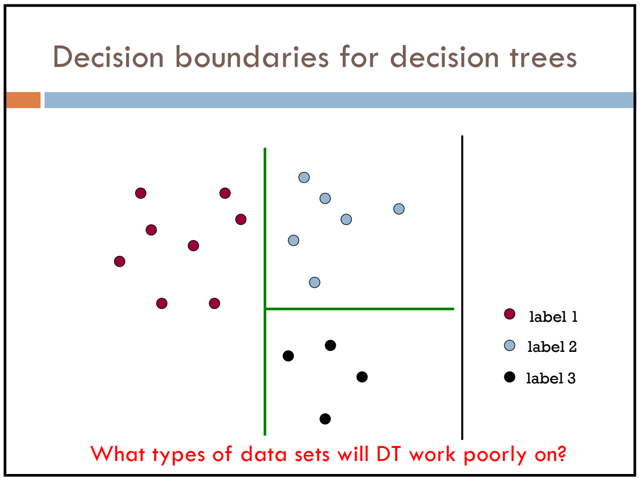
49



Decision boundaries for decision trees

Axis-aligned splits/cuts of the data

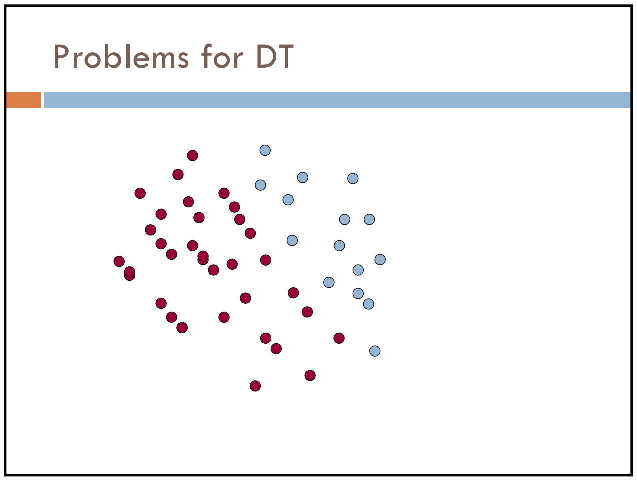
50



Decision boundaries for decision trees

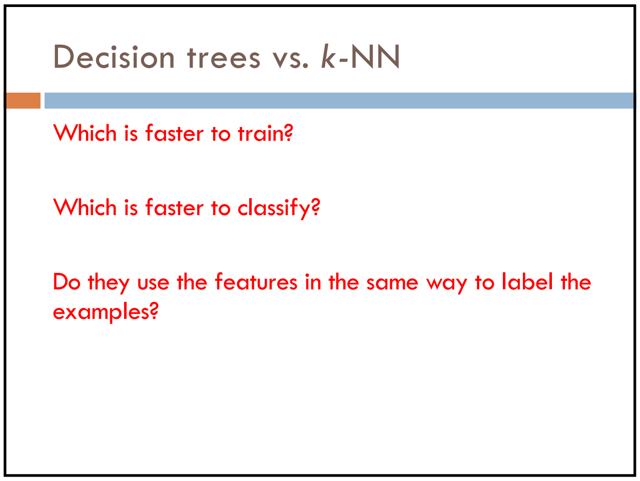
What types of data sets will DT work poorly on?

51



Problems for DT

52



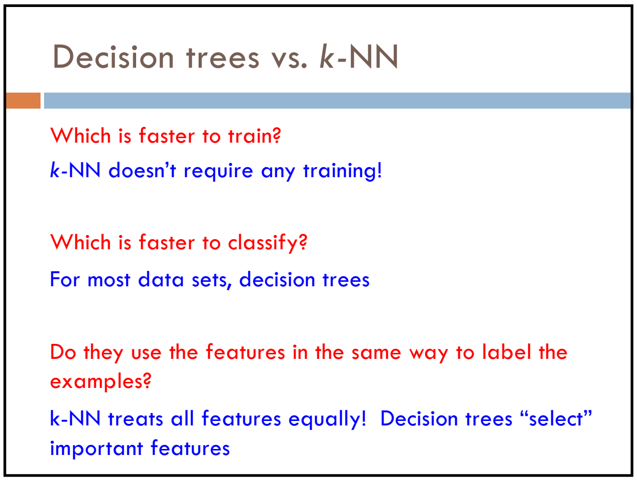
Decision trees vs. *k*-NN

Which is faster to train?

Which is faster to classify?

Do they use the features in the same way to label the examples?

53



Decision trees vs. *k*-NN

Which is faster to train?

k-NN doesn't require any training!

Which is faster to classify?

For most data sets, decision trees

Do they use the features in the same way to label the examples?

k-NN treats all features equally! Decision trees "select" important features

54