Classification

Boston University CS 506 - Lance Galletti

What is Classification?

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

relation that make it

malignant

$\Downarrow$

age VS Tumor size

What is Classification?

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

CLASS

PREDICTORS / FEATURES / ATTRIBUTES

What is Classification?

twice tumor size greater than age.

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

learn model

Model

f : age x tumor size ---> {yes, no}

What is Classification?

age	Tumor size	malignant?
20	9	no
30	16	yes
40	18	no
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
40	18	no
30	16	yes
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
40	18	no
30	16	yes
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
40	18	no
30	16	yes
50	28	yes

What is Classification?

age	Tumor size	malignant?
20	9	no
40	18	no
30	16	yes
50	28	yes

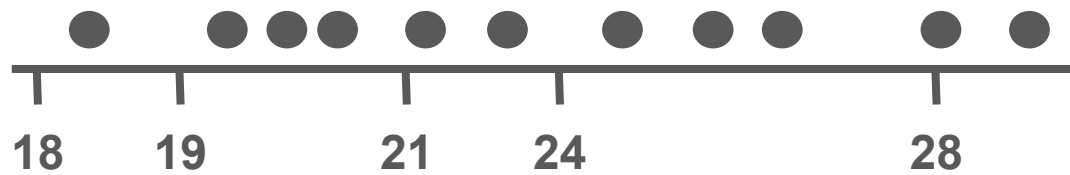**What property / combination of age and tumor size is unique to malignant tumors?**

Is twice tumor size > age
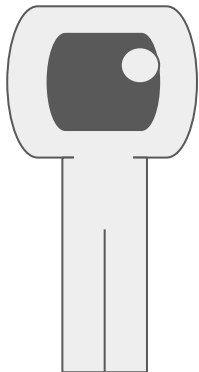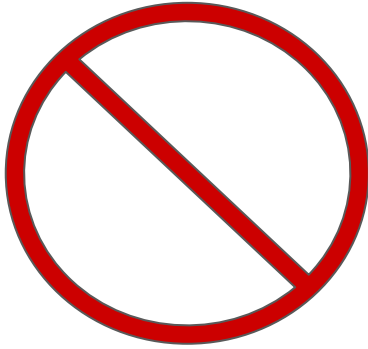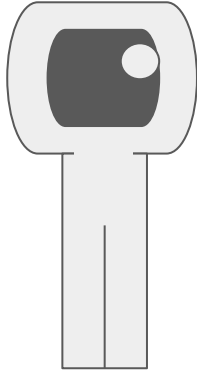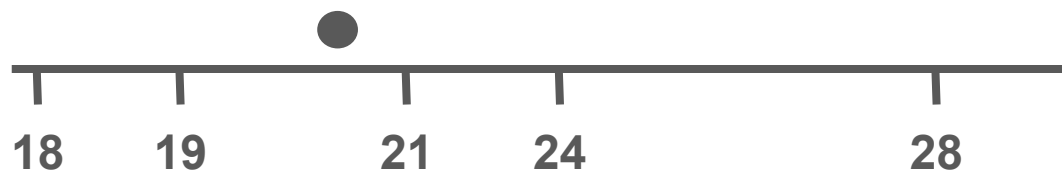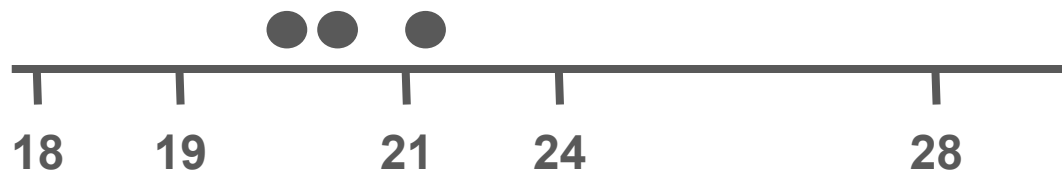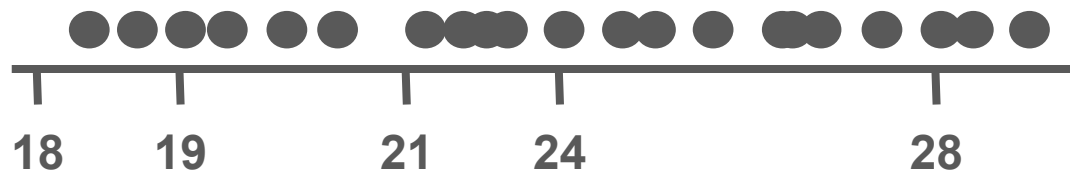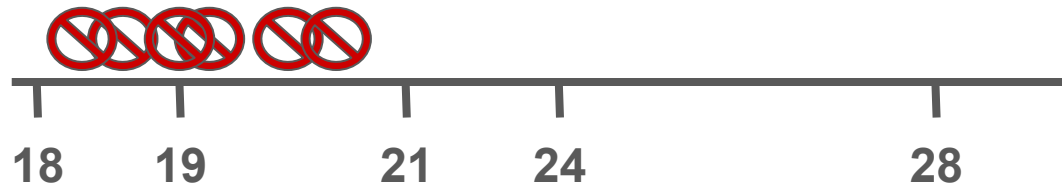
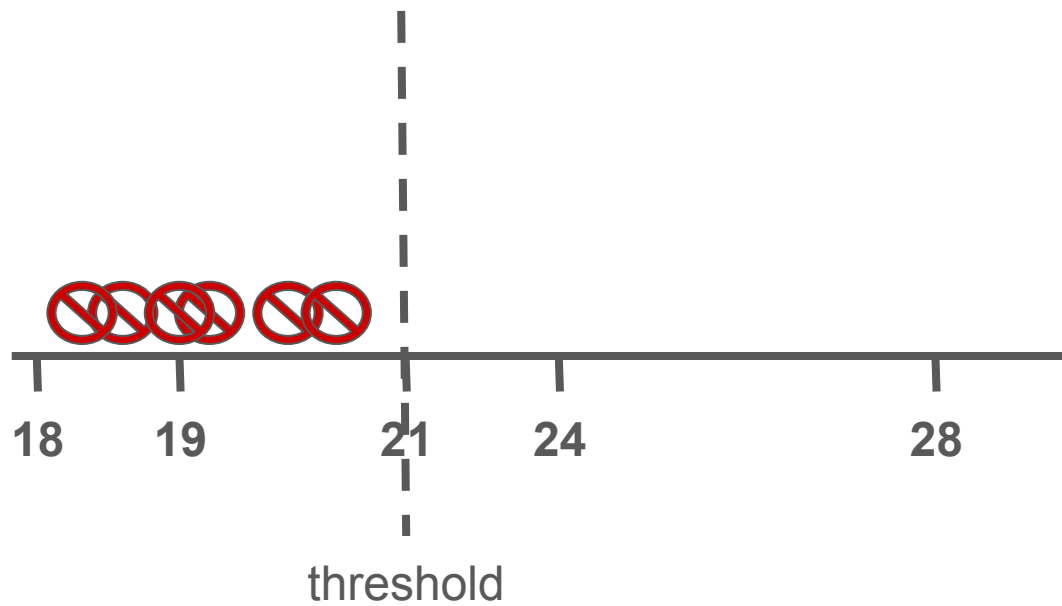18   19   21   24   28

threshold

Sometimes there are **many** correct answers

18   19   21   24   28

Sometimes there are **many** correct answers

18   19   21   24   28

Sometimes there are **many** correct answers

18   19   21   24   28
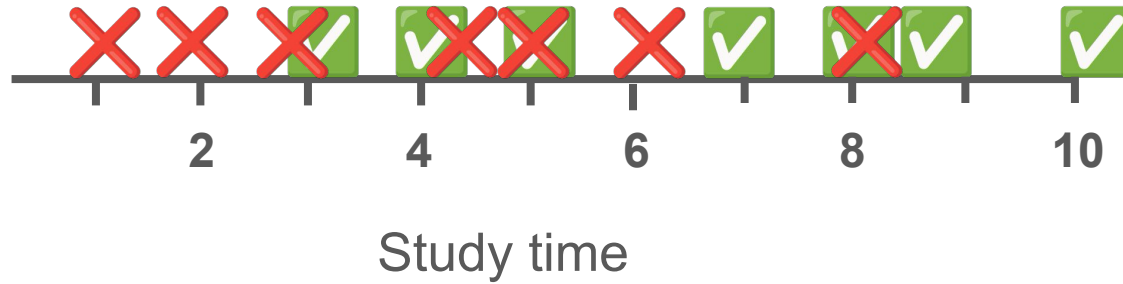
Sometimes there are **many** correct answers
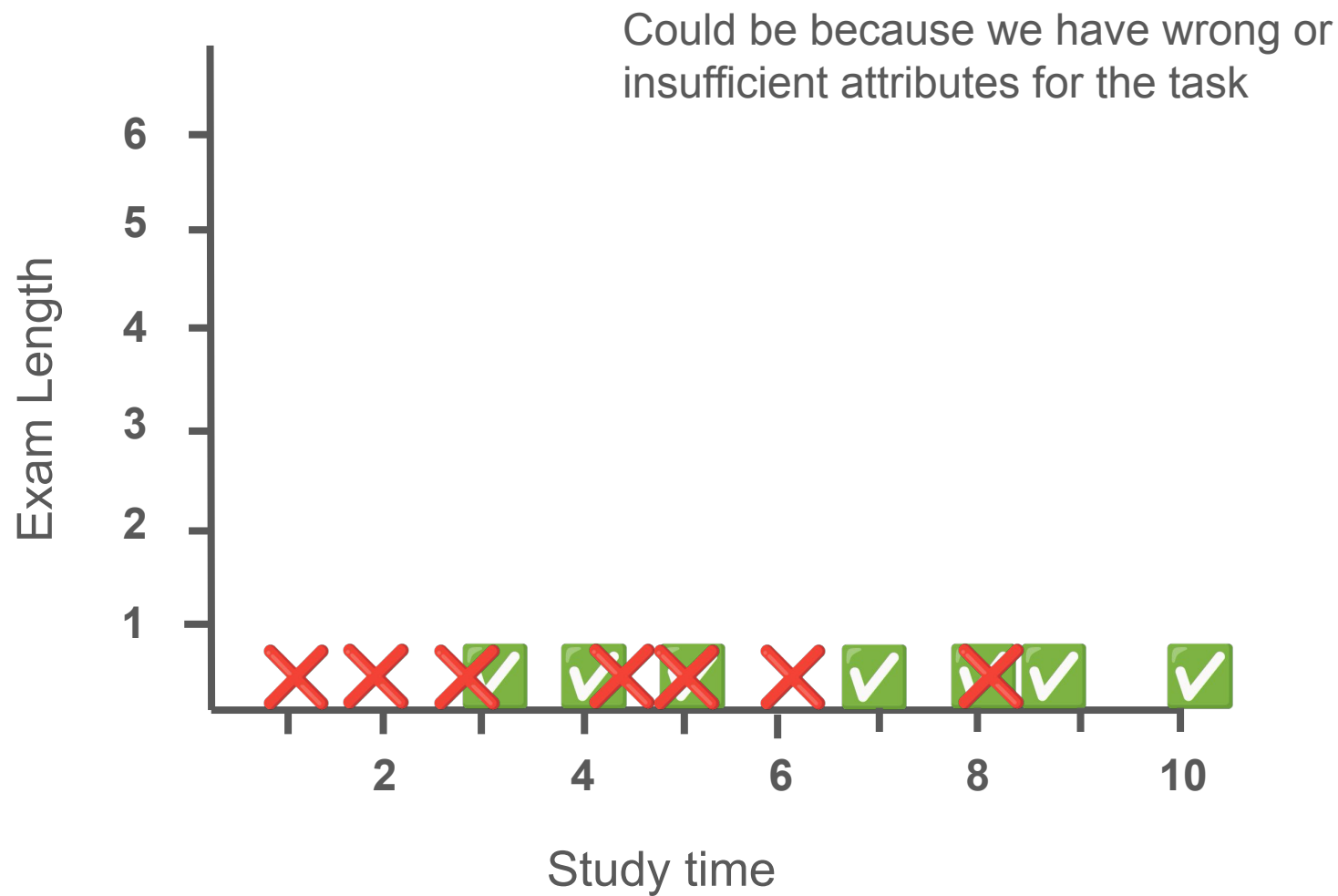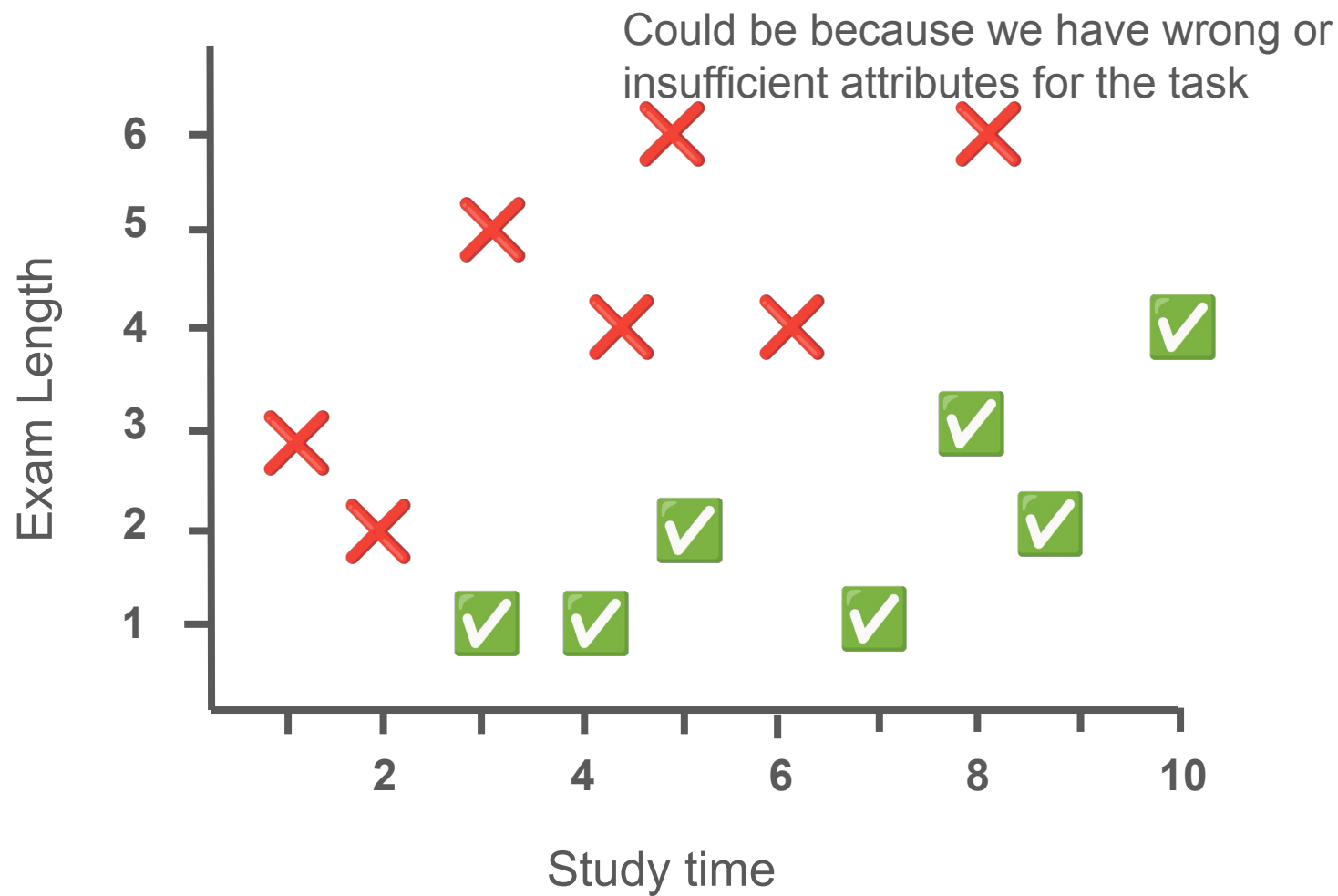
18   19   21   24   28
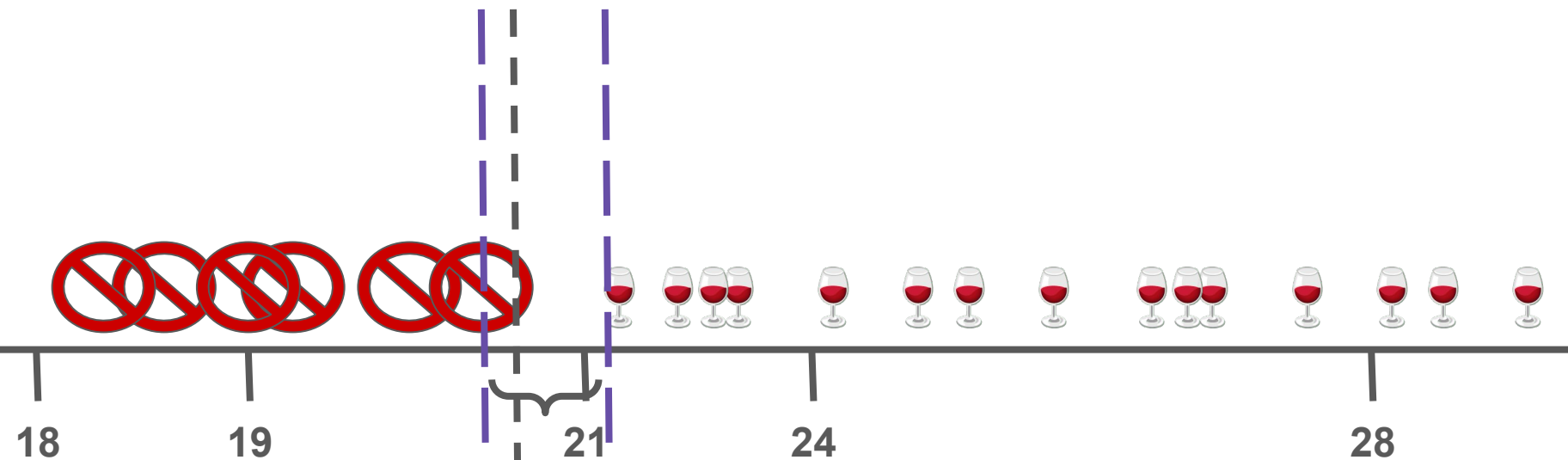
Sometimes there are **no** correct answers
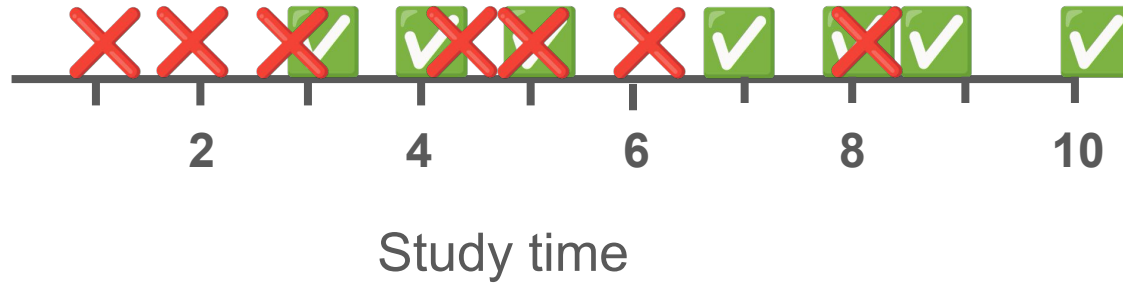
Study time

Could be because we have wrong or insufficient attributes for the task

Study time

Could be because we have wrong or insufficient attributes for the task
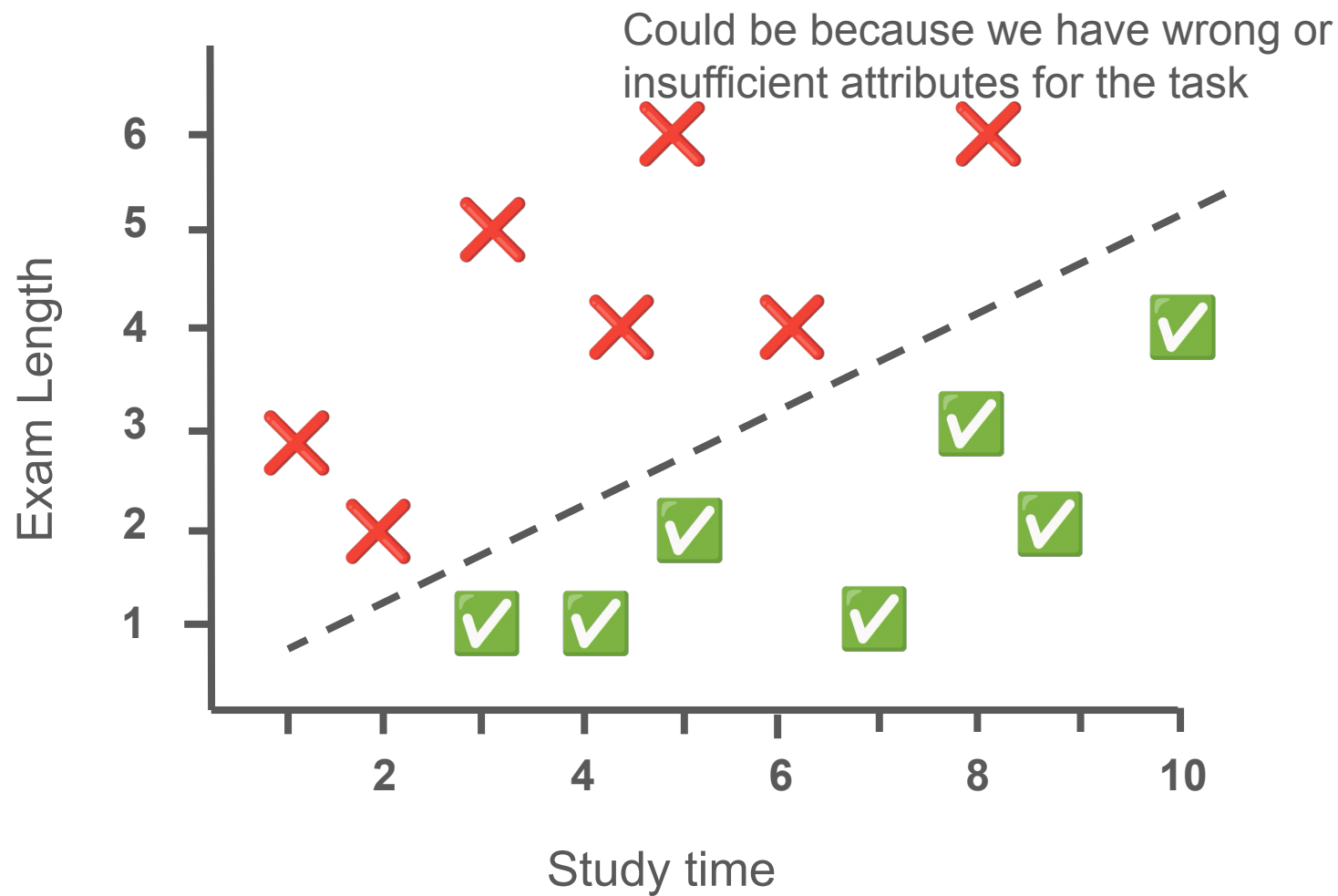
Could be because we have wrong or insufficient attributes for the task

Could be because the problem just doesn't have an exact solution

Rhinos   overlp ∿   Elephants

Weight (in Tons)

0   2   4   6   8   10

Could be because the problem just doesn't have an exact solution

Definitely a Rhino

Not Sure

Definitely an Elephant

Could be because the problem just doesn't have an exact solution

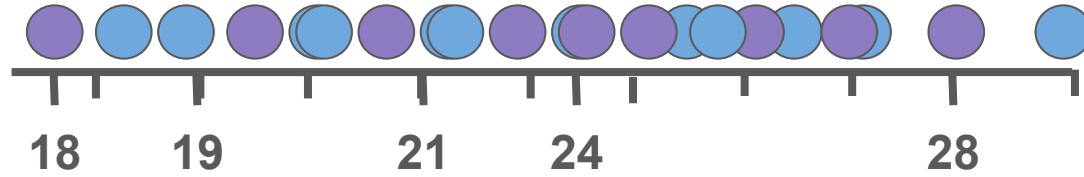All models are wrong but some are useful
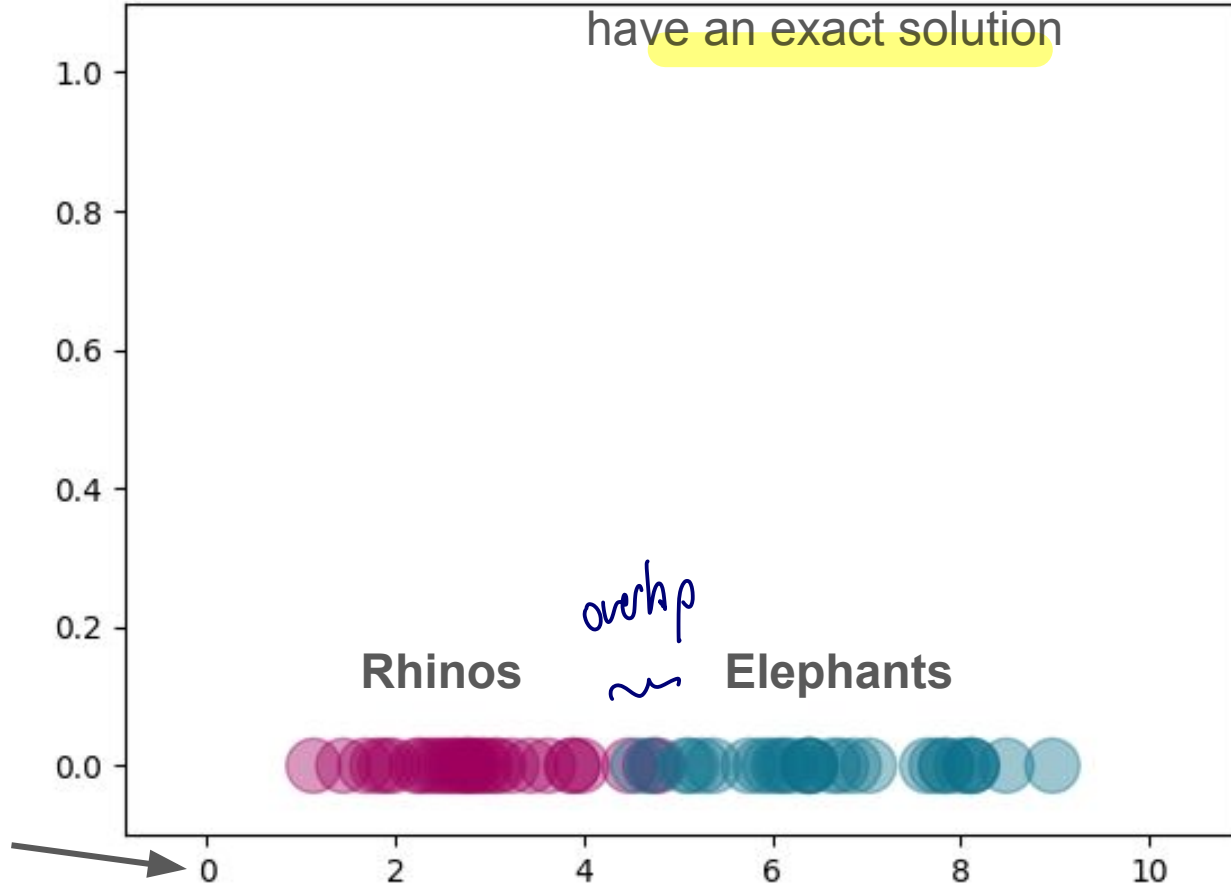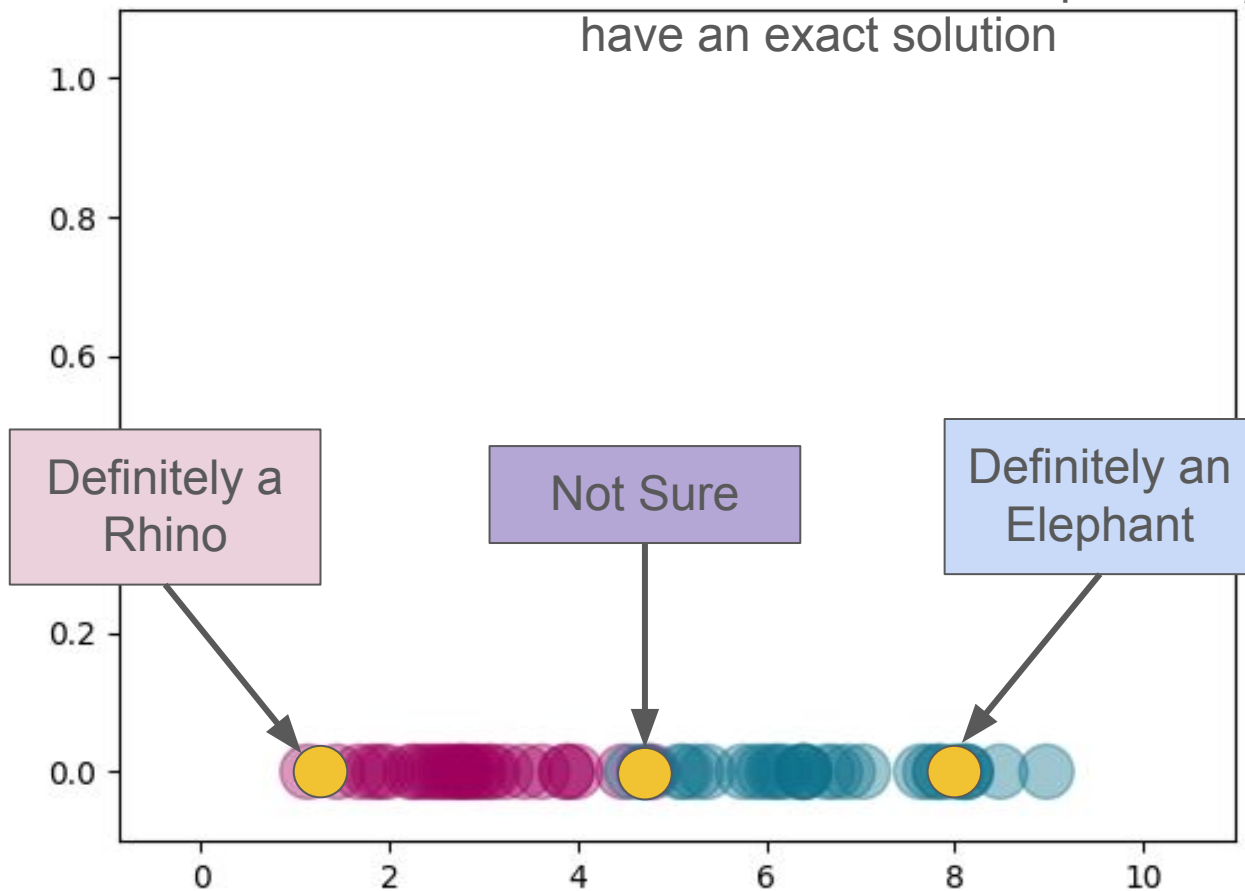
The feasibility of a classification task completely depends on the relationship between the attributes (or predictors) and the class.

For example if we used age instead of weight for elephants and rhinos

18   19   21   24   28

age
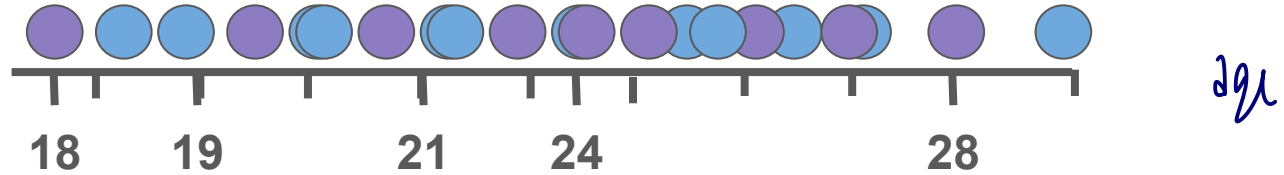
hard time for age
not good predictor

Age cannot distinguish rhinos and elephants

Takeaways

- There could be **many correct answers**
- There could be **no correct answers**
  - And maybe that's ok - no relationship is interesting information too
  - But the model could **still be useful** if it's more or less correct most of the time
- Whether a task is feasible depends on:
  - The relationship between the predictors and the class

Lots of Questions

- How do we know if we have good predictors for a task?
- How do we know we have done a good job at classification?

Lots of Questions

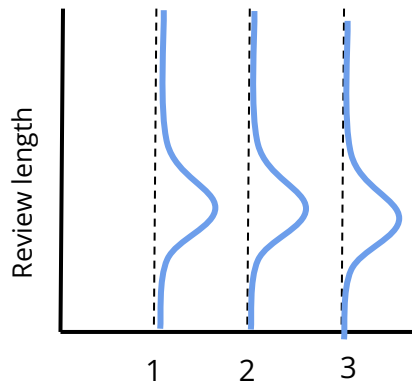- **How do we know if we have good predictors for a task?**
- How do we know we have done a good job at classification?

How do we know we have good predictors?

- What constitutes a good feature/predictor?

amazon
review
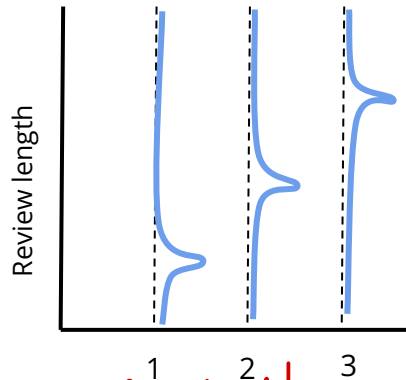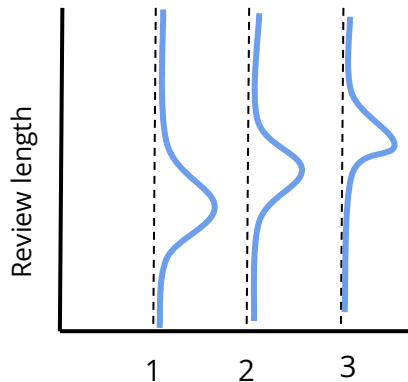
good preds

How do we know we have good predictors?

- What constitutes a good feature/predictor?
- What constitutes a good set of features/predictors?
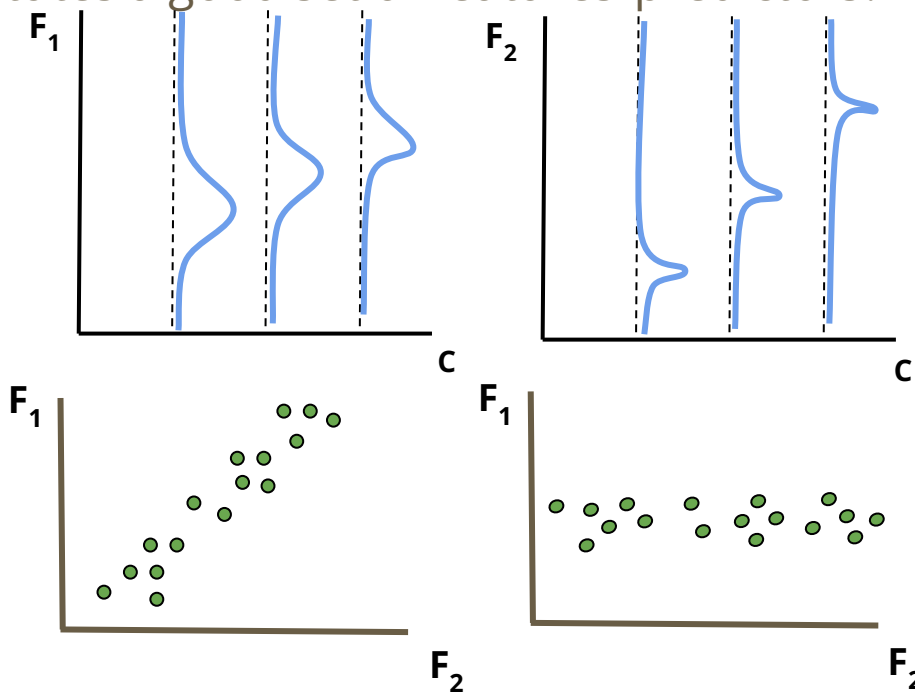
want to independent
respond to

How do we know we have good predictors?

- What constitutes a good feature/predictor?
- What constitutes a good set of features/predictors?

We want features that are related to the target but not to each other. How do we know if features are related?

Correlation

How do we know we have good predictors?

Correlation Δ in X matter for Δy (positively or negatively)

Correlation between X and Y (continuous variables) can be measured as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

* quadratic will not work

How do we know we have good predictors?

Correlation

Correlation between X and Y (continuous variables) can be measured as:
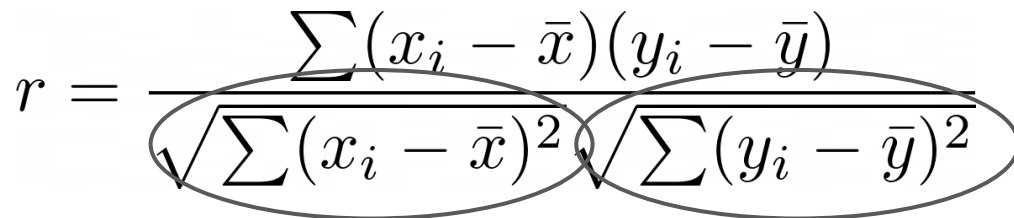
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Covariance

How do we know we have good predictors?

Correlation

Correlation between X and Y (continuous variables) can be measured as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

Standard Deviation of X   **Standard Deviation of Y**

How do we know we have good predictors?

Correlation

positive. correlation
~ 1

Negative relation
~ -1.

Y   X

Y   X

Y   X
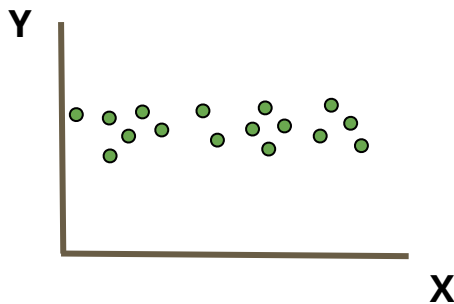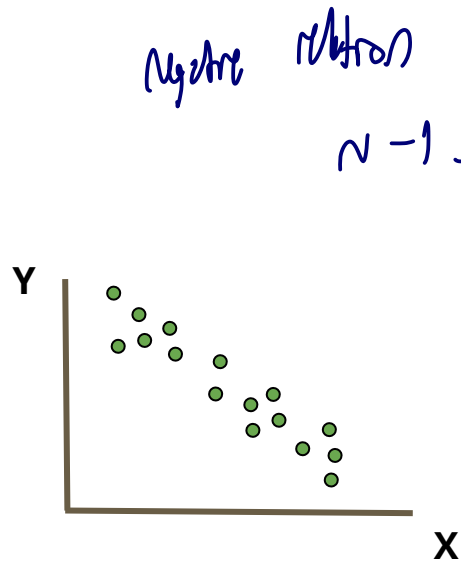
r ~ 1   r = 0   r ~ -1

How do we know we have good predictors?

No correlation doesn't mean there isn't a relationship

def some relation but

he can't capture it

r = 0   r = 0   r = 0
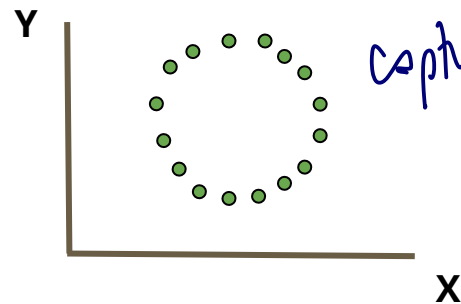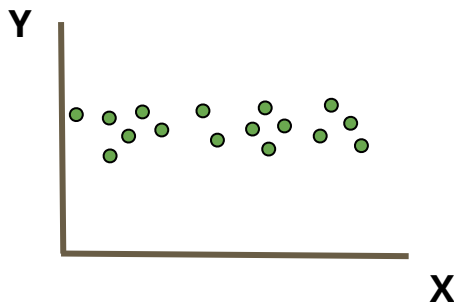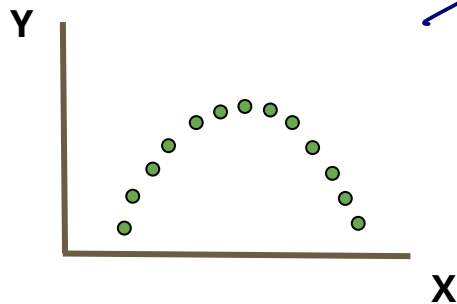
How do we know we have good predictors?

Correlation

What if X is continuous and Y is a class?

How do we know we have good predictors?

Correlation

What if X is continuous and Y is a class?

- Y is {Yes, No}, Colors, Cities etc   Normal
- Y is {Terrible, Bad, Ok, Good, Great}   Ordinal

How do we know we have good predictors?

Correlation

What if X is continuous and Y is a class?

- Y is {Yes, No}, Colors, Cities etc **Nominal**
- Y is {Terrible, Bad, Ok, Good, Great} **Ordinal**

How do we know we have good predictors?

Correlation

What if X is continuous and Y is a class?

- Y is {Yes, No}, Colors, Cities etc   **Nominal**
- Y is {Terrible, Bad, Ok, Good, Great}   **Ordinal**

No order, need to look at the means of X and how they differ across each Nominal Value of Y

There is a natural Order and we can assign numbers to each category like {0, 1, 2, 3, 4}. BUT it's not clear if the difference between Terrible and Bad is the same as between OK and Good…

How do we know we have good predictors?

Correlation

What if X is continuous and Y is a class?

- Y is {Yes, No}, Colors, Cities etc   **Nominal**
- Y is {Terrible, Bad, Ok, Good, Great}   **Ordinal**

No order, need to look at the means of X and how they differ across each Nominal Value of Y

Instead of using the exact numbers assigned we can compare their rank / position in the data. So it doesn't matter how far Bad is from Ok it just matters that Ok comes after Bad.

Spearman Coefficient Example

X	Y
10	1
20	0
30	2
40	3
50	4

Spearman Coefficient Example

X	Y
10	1
20	0
30	2
40	3
50	4

R(X)	R(Y)
1	2
2	1
3	3
4	4
5	5

Spearman Coefficient Example

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

X	Y
10	1
20	0
30	2
40	3
50	4

R(X)	R(Y)	d
1	2	-1
2	1	1
3	3	0
4	4	0
5	5	0

Spearman Coefficient Example

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

-1 one to 1

sum squared 2.

X	Y
10	1
20	0
30	2
40	3
50	4

R(X)	R(Y)	d
1	2	-1
2	1	1
3	3	0
4	4	0
5	5	0

p = .9

How do we know we have good predictors?

- What constitutes a good feature/predictor?
- What constitutes a good set of features/predictors?

We want features that are related to the target but not to each other. How do we know if features are related? **Correlation (Pearson or Spearman)**

BUT…

Correlation is not causation.

How do we know we have good predictors?

Correlation VS Causation

1. Temperature and ice cream sales are positively correlated

How do we know we have good predictors?

Correlation VS Causation

1. Temperature and ice cream sales are positively correlated
   a. Temperature increases seem to cause ice cream sales to spike
      i. BUT in the desert where there is no ice cream, there is no spike in sales.
   b. Ice cream sale increases do not cause the temperature to rise

counterexample

How do we know we have good predictors?

Correlation VS Causation

1. Temperature and ice cream sales are positively correlated
   a. Temperature increases seem to cause ice cream sales to spike
      i. BUT in the desert where there is no ice cream, there is no spike in sales.
   b. Ice cream sale increases do not cause the temperature to rise
2. Sleeping with shoes on is strongly correlated with waking up with a headache.

How do we know we have good predictors?

Correlation VS Causation

1. Temperature and ice cream sales are positively correlated
   a. Temperature increases seem to cause ice cream sales to spike
      i. BUT in the desert where there is no ice cream, there is no spike in sales.
   b. Ice cream sale increases do not cause the temperature to rise
2. Sleeping with shoes on is strongly correlated with waking up with a headache.
   a. But neither causes the other…
   b. There's a third common factor causing this correlation: going to bed drunk.

Causation

Testing for causality requires specific testing / experimentation with a control group

But it's very hard to show that things are causally linked through observational data... Especially if the relationship isn't deterministic.

Causation

Testing for causality requires specific testing / experimentation with a control group

But it's very hard to show that things are causally linked through observational data... Especially if the relationship isn't deterministic.

- Not everyone who smokes will get lung cancer

Causation

Testing for causality requires specific testing / experimentation with a control group

But it's very hard to show that things are causally linked through observational data... Especially if the relationship isn't deterministic.

- Not everyone who smokes will get lung cancer

Hard to say what would have happened if you hadn't smoked cause you can't rewrite the past.

Causation

Testing for causality requires specific testing / experimentation with a control group

Ethical considerations when testing effects of new drugs

Example 1

women

men

	applied	accepted	rate	applied	accepted	rate
TOTAL	1184	274	**23%**	2470	584	**24%**

Example 1

aggregate or flipped.

wm

m

	applied	accepted	rate	applied	accepted	rate
Computer science	26	7	27%	228	58	25%
Economics	240	63	26%	512	112	22%
Engineering	164	52	32%	972	252	26%
Medicine	416	99	24%	578	140	24%
Veterinary Med	338	53	16%	180	22	12%
TOTAL	1184	274	**23%**	2470	584	**24%**

Example 1

Simpson's Paradox

When the relationship is reversed when data is aggregated.

Example 2

Bob is diagnosed with lung cancer. He thinks it's because of exposure to asbestos while working and decides to sue his previous employer.

It cannot be proven that the cancer would not have occurred without exposure to asbestos. So how does Bob go about this?

you cant prov?.

Example 2

Bob is diagnosed with lung cancer. He thinks it's because of exposure to asbestos while working and decides to sue his previous employer.

- Out of 1000 people like Bob, 10 are expected to develop lung cancer
- Exposure to asbestos more than doubles the expected development of lung cancer. So out of 1000 people exposed to asbestos like Bob, let's say about 25 would get lung cancer.
- So of those exposed to asbestos that go on to develop lung cancer, less than half developed lung cancer if they had no been exposed.
- So it's more likely than not that the asbestos caused the cancer.

Bob wins

Example 2

Forensic epidemiology

* closr to 1 or -1 is good.

Lots of Questions

- How do we know if we have good predictors for a task?
- **How do we know we have done a good job at classification?**

How do we know we've done well at classification?
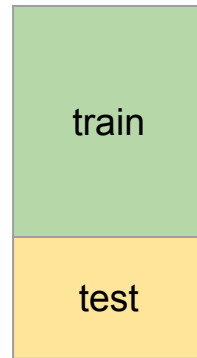
How do we know we've done well at classification?

- Testing without cheating. Learning not memorizing.

— evaluate on data it has never seen before

*if deliver trained model, just ship it as it is

How do we know we've done well at classification?

- Testing without cheating. Learning not memorizing.
  - Split up our data into a training set and a separate testing set
  - Use the training set to find patterns and create a model
  - Use the testing set to evaluate the model on data it has not seen before

train

test

How do we know we've done well at classification?

- Testing without cheating. Learning not memorizing.
  - Split up our data into a training set and a separate testing set
  - Use the training set to find patterns and create a model
  - Use the testing set to evaluate the model on data it has not seen before
- Also allows us to check that we have not learned a model TOO SPECIFIC to the dataset ~~avoid~~ (not generalized enough)
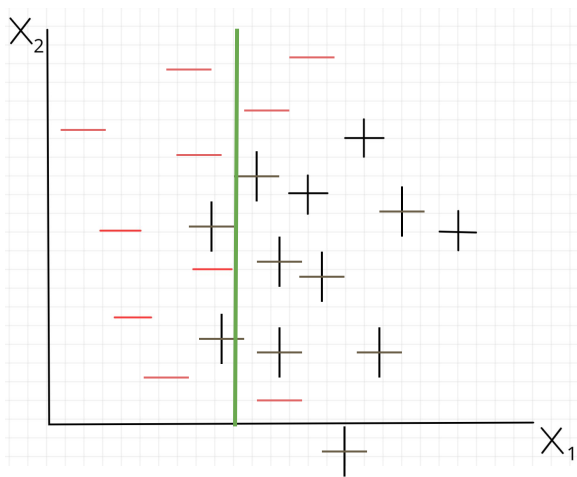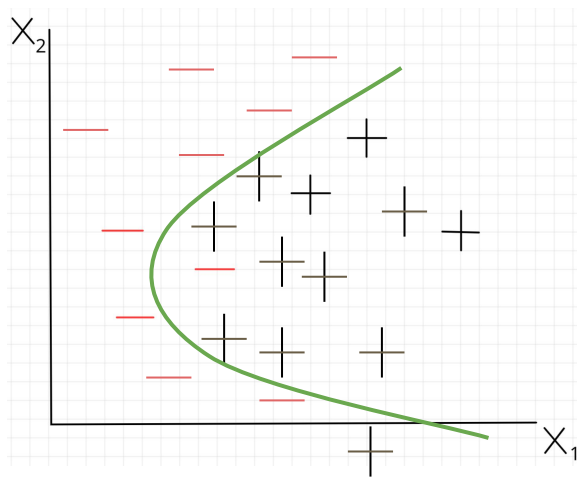 - Overfitting vs underfitting

complex enough!!

train
test

How do we know we've done well at classification?
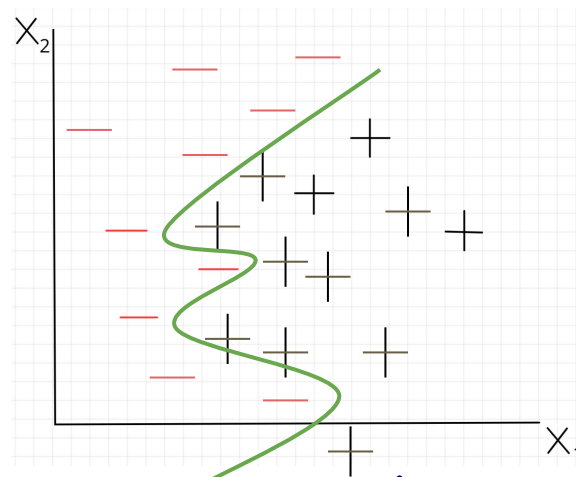
Underfitting VS Overfitting

to simple   good   too specific.

How do we know we've done well at classification?

Underfitting VS Overfitting

mistakes made by the model

underfitting

sweet spot! memorize enough, im good!

Stop here

Test

overfitting
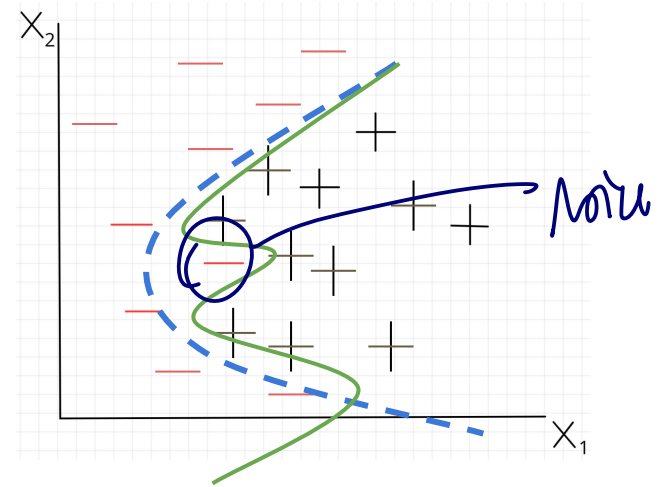
Train
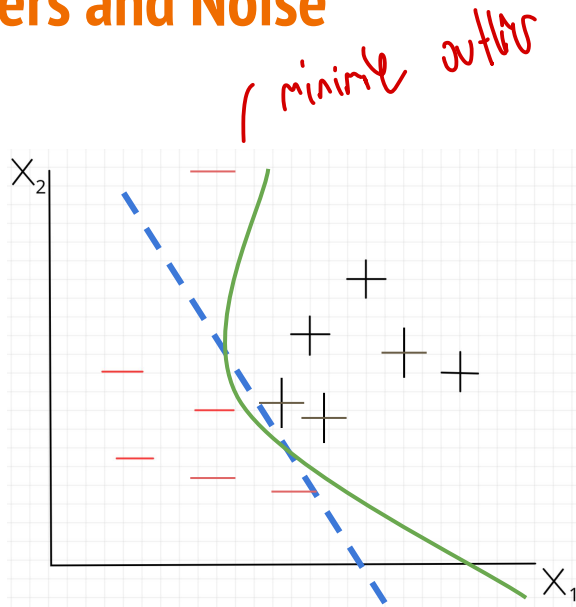
Complexity of the model

How do we know we've done well at classification?

- Testing without cheating:
  - Split up our data into a training set and a separate testing set
  - Use the training set to find patterns and create a model
  - Use the testing set to evaluate the model on data it has not seen before
- Also allows us to check that we have not learned a model TOO SPECIFIC to the dataset
  - Overfitting vs underfitting
  - Goal is to capture general trends
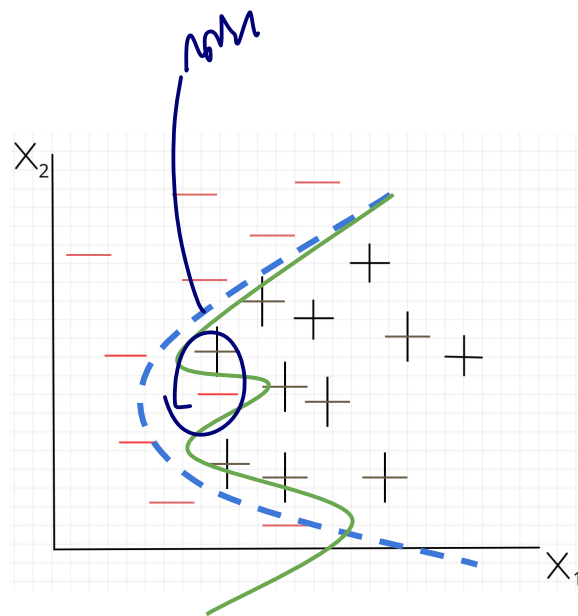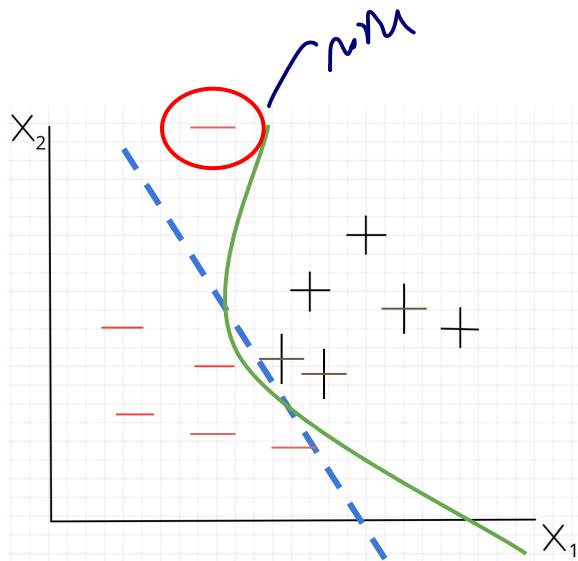   - Watch out for outliers and noise

How do we know we've done well at classification?
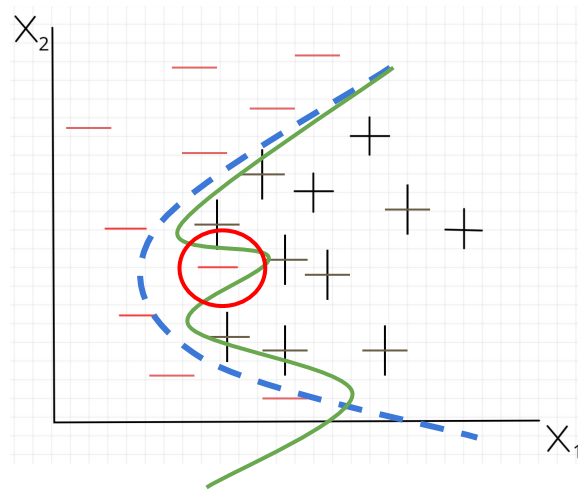
Outliers and Noise

minime outlier

noise

How do we know we've done well at classification?

Outliers and Noise

How do we know we've done well at classification?

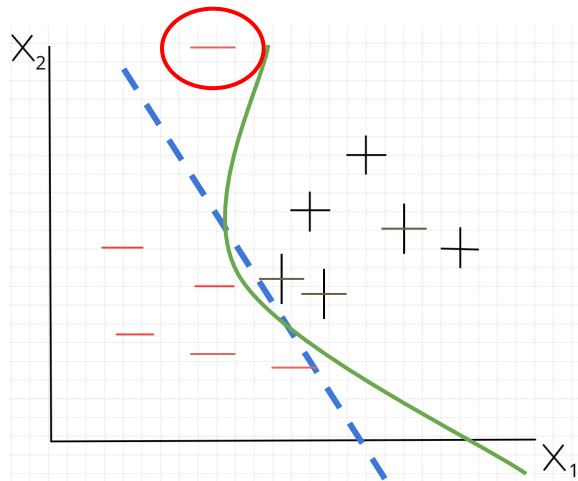## Outliers and Noise

How do we know we've done well at classification?

- Testing without cheating:
  - Split up our data into a training set and a separate testing set
  - Use the training set to find patterns and create a model
  - Use the testing set to evaluate the model on data it has not seen before
- Also allows us to check that we have not learned a model TOO SPECIFIC to the dataset
  - Overfitting vs underfitting
  - Goal is to capture general trends
    - Watch out for outliers and noise
- The types of mistakes made matters

How do we know we've done well at classification?

Types of mistakes

- Testing for a rare disease
  - Out of 1000 data points, only 10 have this rare disease. A model that simply tells folks they don't have the disease will have an accuracy of 99%.

Classification

- Training Step
  - Create the model based on the examples / data points in the training set
- Testing Step
  - Use the model to fill in the blanks of the testing set
  - Compare the result of the model to the true values

Instance-Based Classifiers

(1) classifier
lazy! just data itself -

- Use the stored training records to predict the class label of unseen cases
- Rote-learners:
  - Perform classification only if the attributes of the unseen record exactly match a record in our training set

* find existing datapoint
that v found.

Instance-Based Classifiers: Training Step

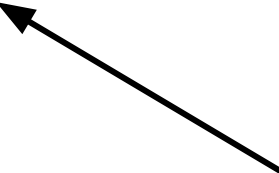| age	Tumor size	malignant?
20	10	no
30	15	yes
40	20	no
50	25	yes

learn model

There is no training step per se. The dataset itself is the model.

Instance-Based Classifiers: Applying the model

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

age	Tumor size	malignant?
20	10	?

Instance-Based Classifiers: Applying the model

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

age	Tumor size	malignant?
20	10	**no**
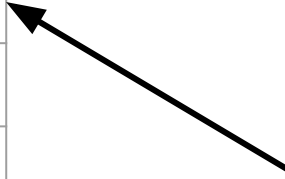
work well for large dataset

Instance-Based Classifiers

- Use the stored training records to predict the class label of unseen cases
- Rote-learners:
  - Perform classification only if the attributes of the unseen record exactly match a record in our training set

Instance-Based Classifiers

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

not much esnthy -

age	Tumor size	malignant?
25	5	?

closet one ?.

aggregate than 1

Nearest Neighbor Classifier

Use **SIMILAR** records to perform classification

K Nearest Neighbor Classifier

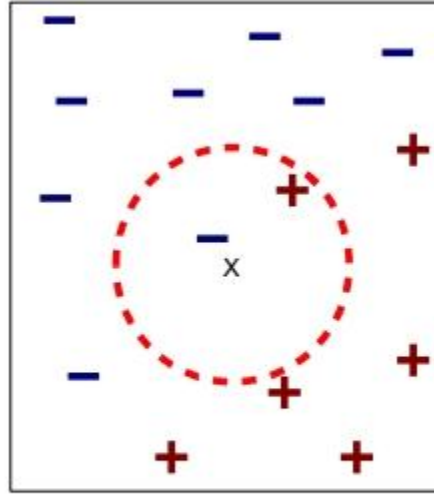Requires:
- Training set
- Distance function
- Value for k

How to classify an unseen record:
1. Compute distance of unseen record to all training records
2. Identify the k nearest neighbors
3. Aggregate the labels of these k neighbors to predict the unseen record class (ex: majority rule)
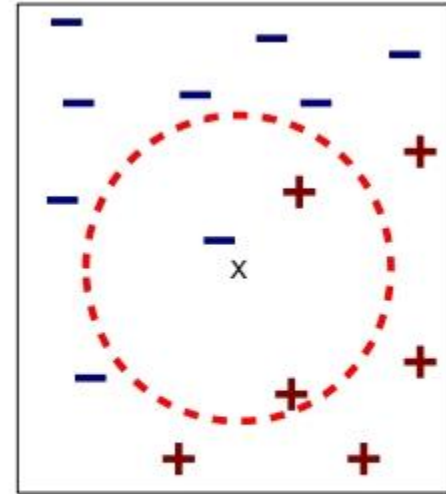
K Nearest Neighbor Classifier
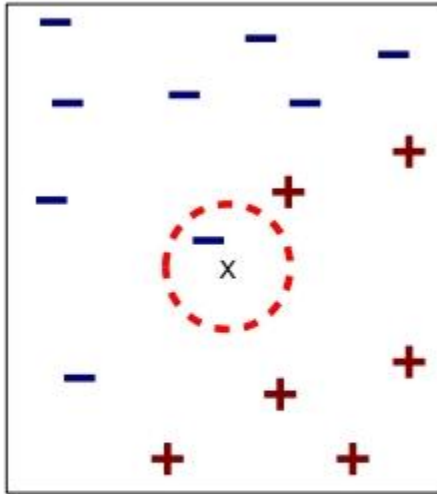
(a) 1-nearest neighbor

(b) 2-nearest neighbor

(c) 3-nearest neighbor

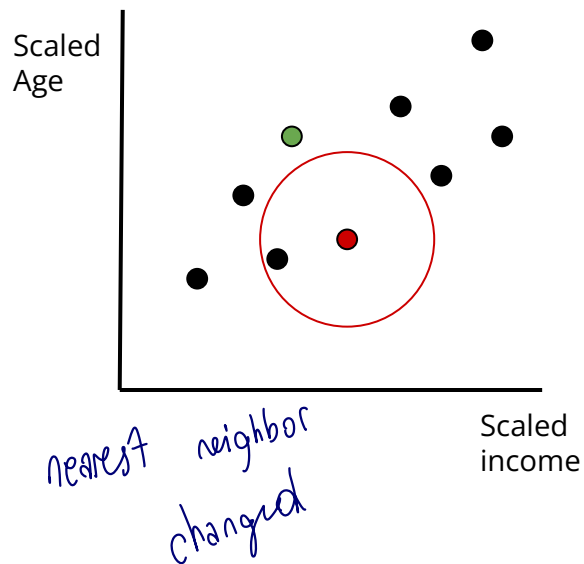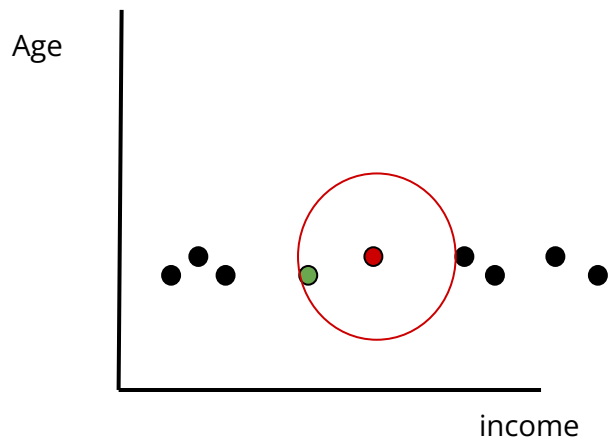# K Nearest Neighbor Classifier

scale note , weight note .

Aggregation methods:
- Majority rule
- Weighted majority based on distance ($w = 1/d^2$)

Scaling issues:
- Attributes should be scaled to prevent distance measures from being dominated by one attribute. Example:
  - Age: 0 -> 100
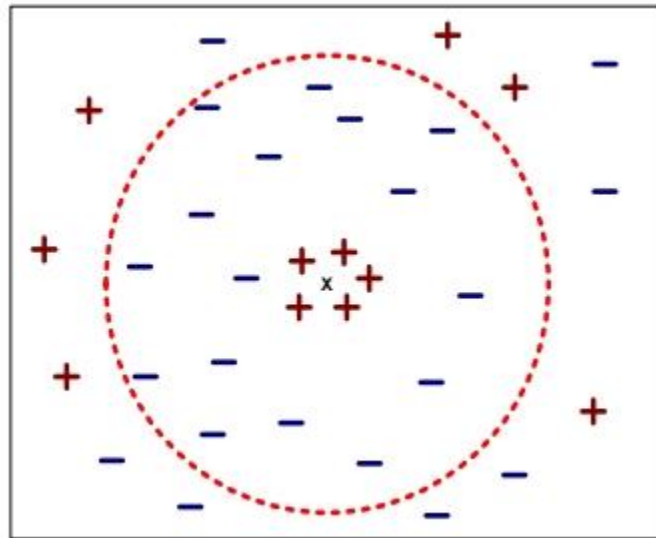 - Income: 10k -> 1million

Scaling Attributes
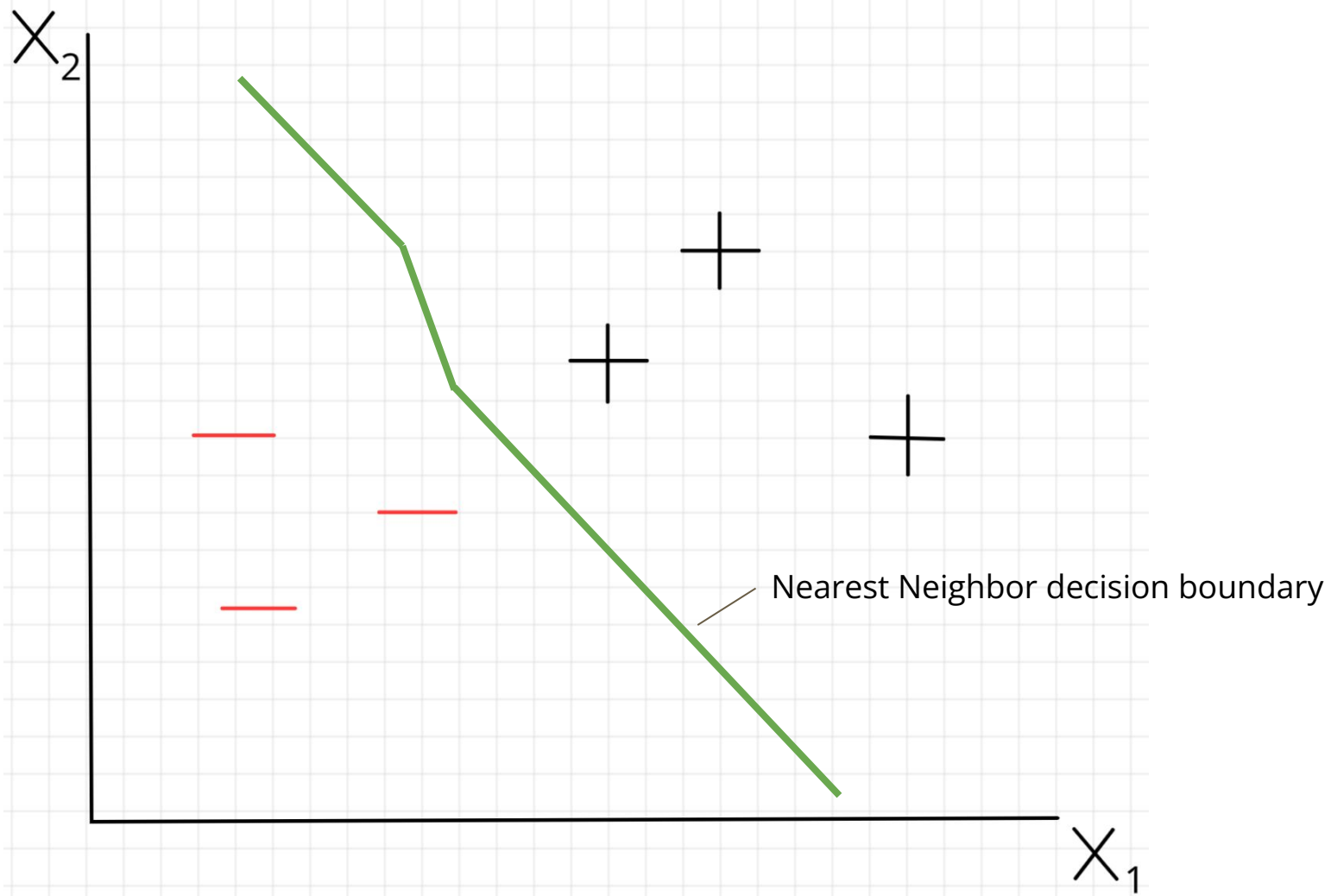
K Nearest Neighbor Classifier

Choosing the value of k:

- If k is too small ->
  - sensitive to noise points + doesn't generalize well ~overfitting 2. if k=1

- If k is too big ->
  - neighborhood may include points from other classes , suffer

Nearest Neighbor decision boundary

K Nearest Neighbor Classifier

Pros:
- Simple to understand why a given unseen record was given a particular class

Cons:
- Expensive to classify new points
- KNN can be problematic in high dimensions (curse of dimensionality)

A B C

Where would the KNN decision boundary be for K=1?

18   19   21   24   28