Intro to Data Science:

KNN Classification

Ed Podojil Data Engineer/Scientist, Animoto

Warm Up

How can we could predict the winner of an election?

What kind of data do we need?

What kind of answer are we trying to solve?

Agenda

I. Classification Problems

II. Building Effective Classifiers

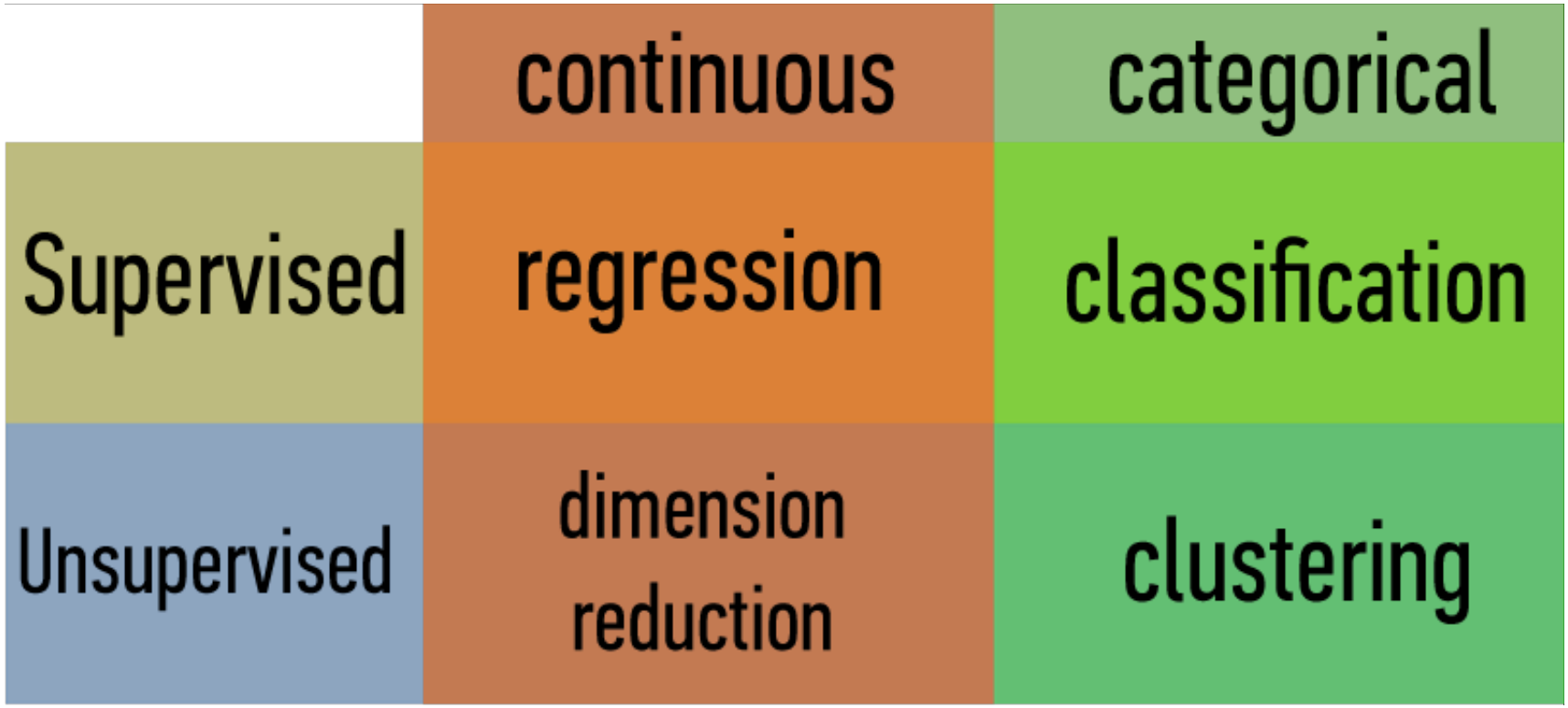
III. The KNN Classification Model

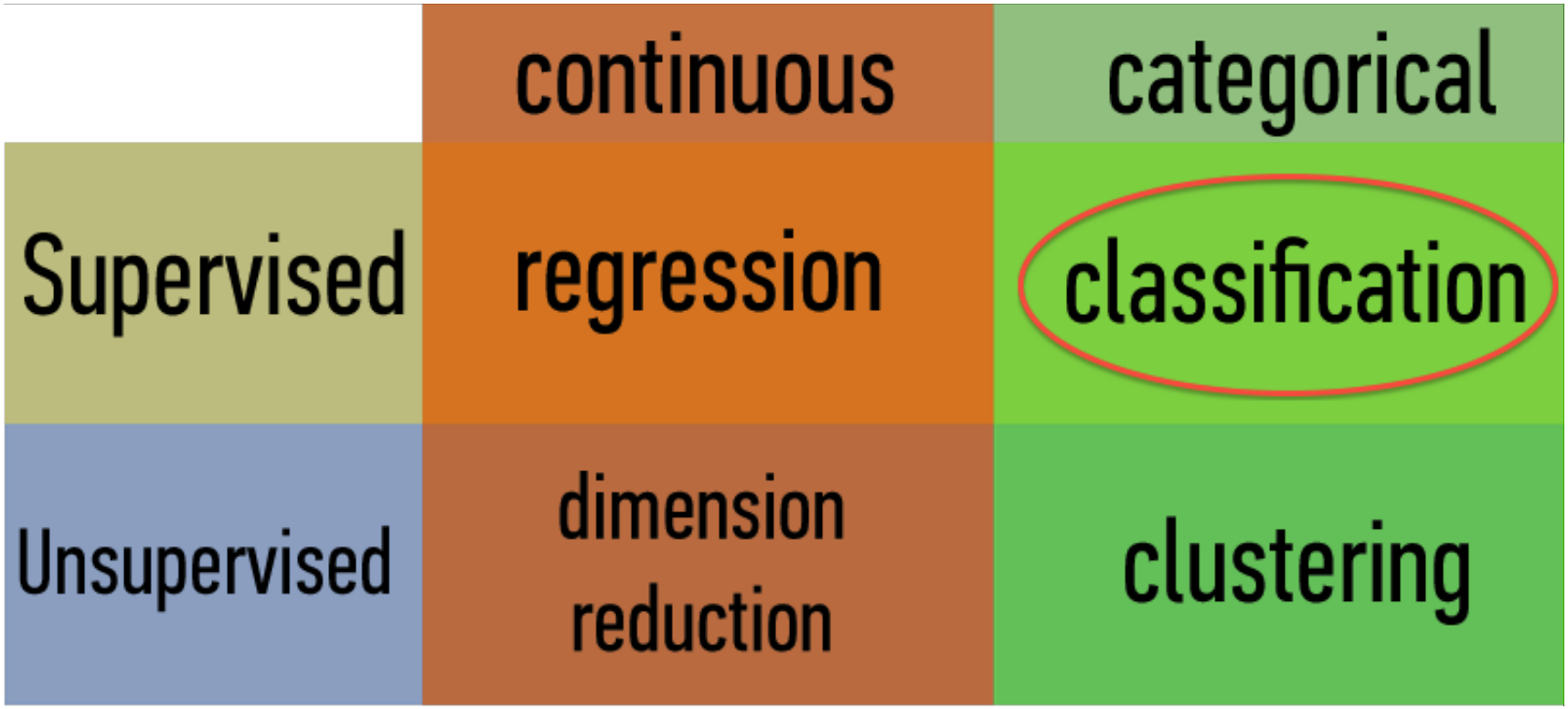
I. Classification Problems

Goals

- Clarify what a **classification problem** is
- Understand what data looks like for a classification problem
- Review **supervised learning**
- Determine steps required for classification


continuous
categorical
Supervised
Unsupervised


continuous
categorical
Supervised
regression
classification
Unsupervised
dimension reduction
clustering


continuous
categorical
Supervised
regression
classification
Unsupervised
dimension reduction
clustering

Example Data

```
> head(iris, 10)
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
3           4.7         3.2          1.3         0.2  setosa
4           4.6         3.1          1.5         0.2  setosa
5           5.0         3.6          1.4         0.2  setosa
6           5.4         3.9          1.7         0.4  setosa
7           4.6         3.4          1.4         0.3  setosa
8           5.0         3.4          1.5         0.2  setosa
9           4.4         2.9          1.4         0.2  setosa
10          4.9         3.1          1.5         0.1  setosa
```

What does "supervised" mean?

We already know the labels we are attempting to classify.

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

How does a classification problem work?

Input data, output predicted labels.

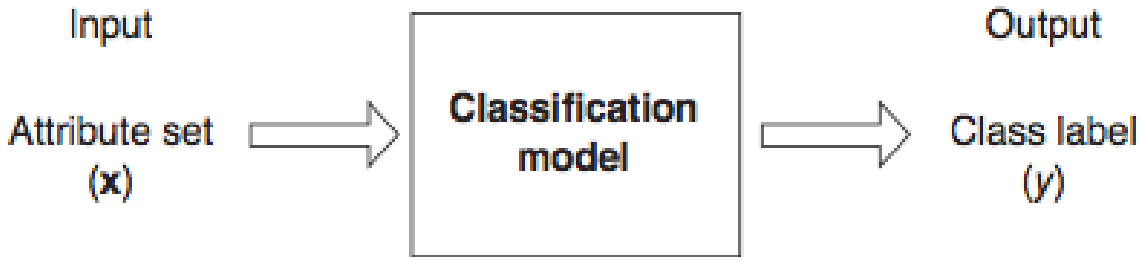
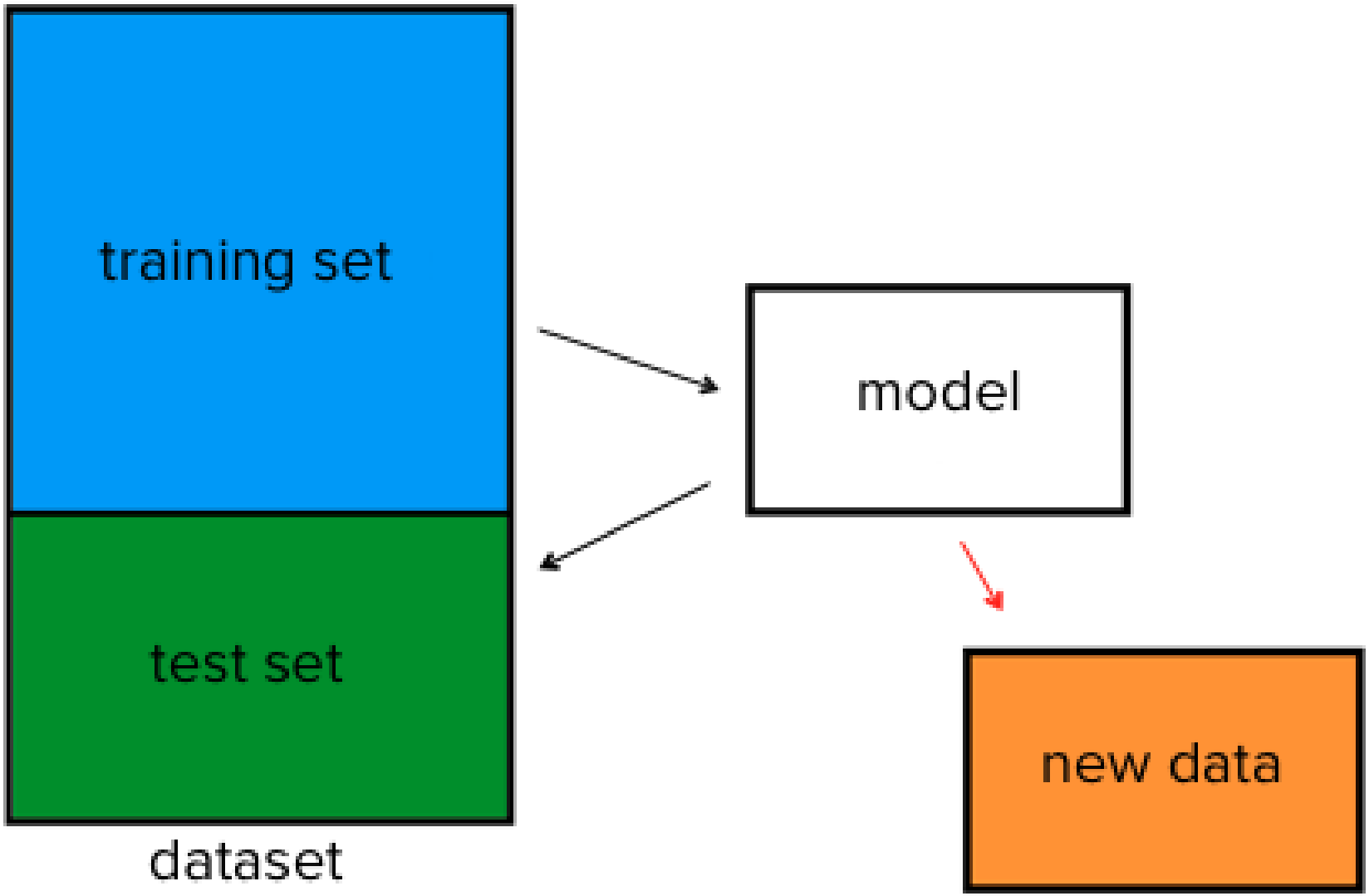


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

What steps does a classification problem require?

1. Split the dataset
2. Train the model
3. Test the model
4. Make predictions with new data

Model of breaking apart data for train, test, and output



II. Building Effective Classifiers

Goals:

- Determine types of **prediction errors**
- Explain why we use **training and test sets**
- Explain what **"overfitting"** and **"underfitting"** mean
- Describe the basic process of **n-fold cross validation**

Life with errors

aka "why predictive analytics is hard"

Types of errors we'll run into

Training error

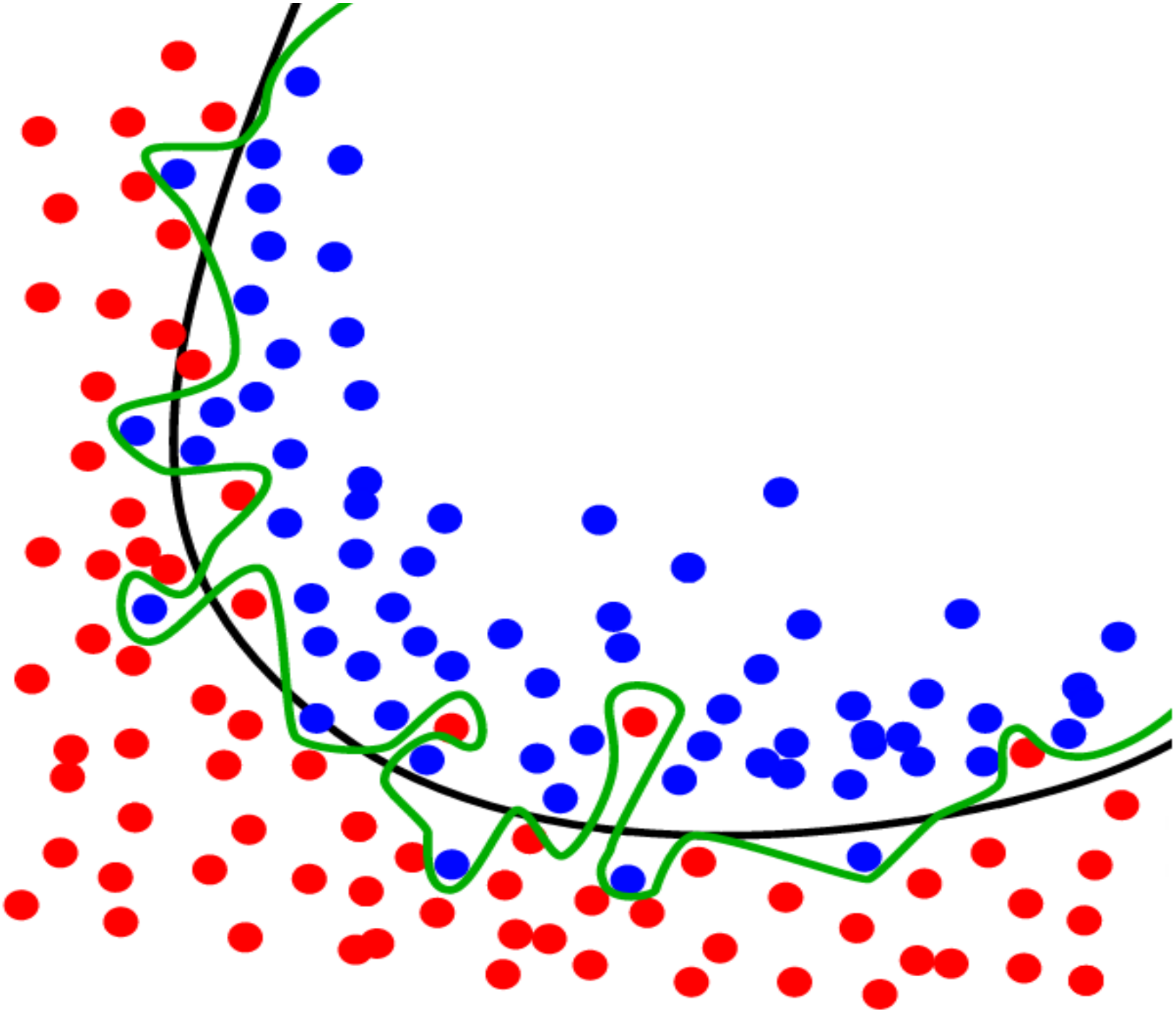
Generalization error

Out of Sample (new data) error

Training error

What happens if we had no test data and only used a training set?

Overfitting

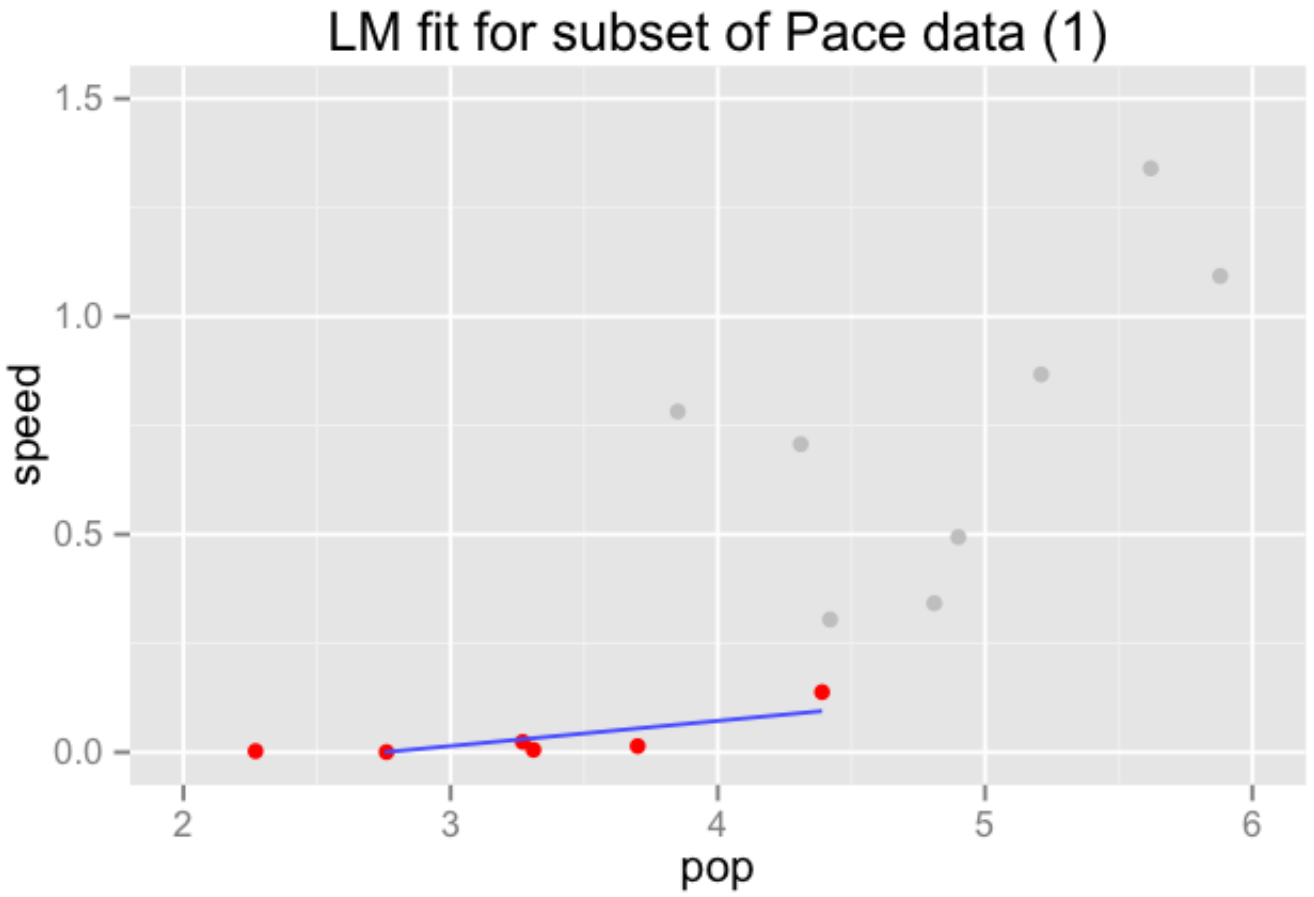


Generalization Error

How well does the model generalize to **unseen** data?

Generalization Error

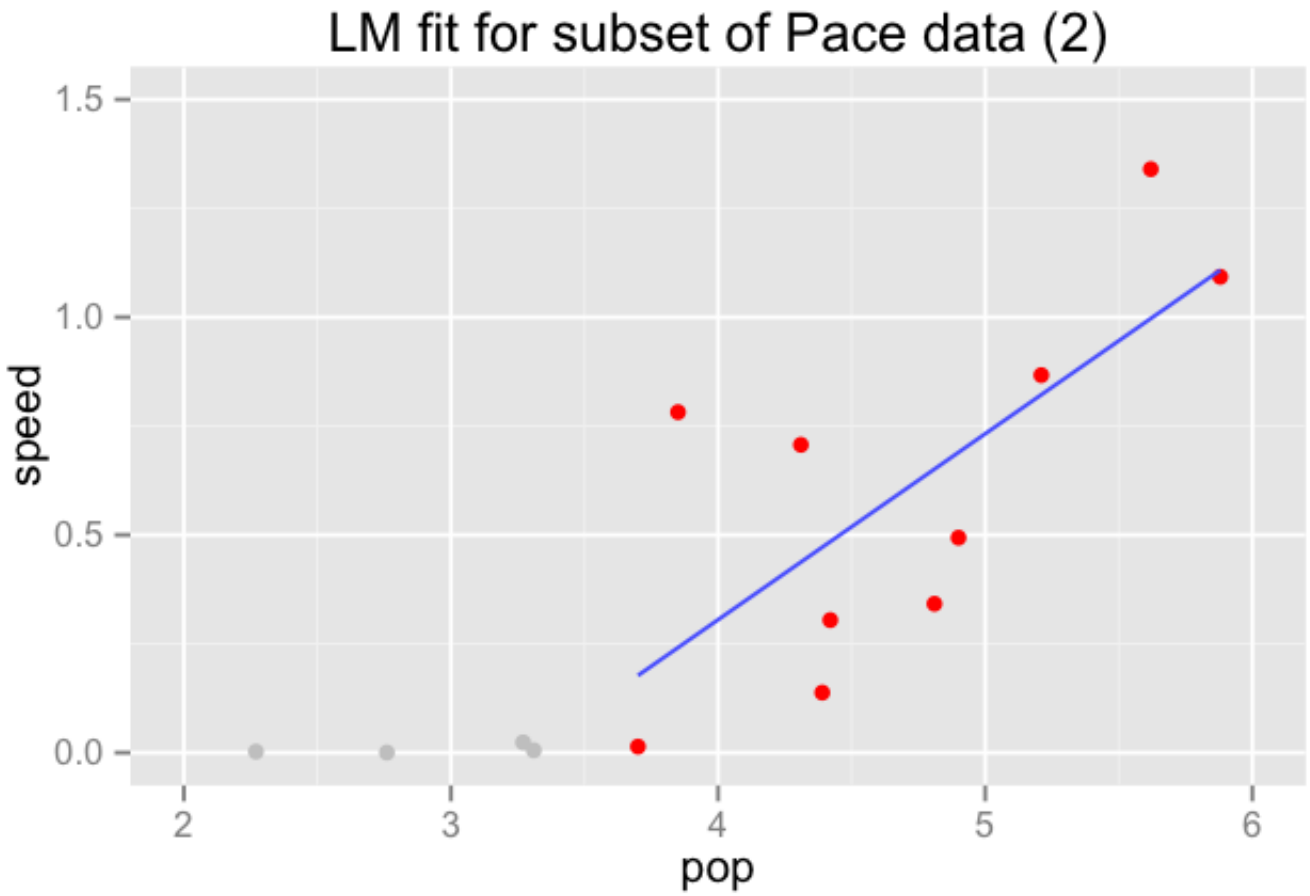
Try training the model on a subset of the data:



Does generalization error remain the same with different training data?

Generalization Error

Try training the model on a subset of the data:



Does generalization error remain the same with different training data? No!

Generalization Error

- Generalization error gives a high variance estimate of OOS error
- Insufficient training can lead to incorrect model selection

So how do we fix this?

Think about it: if we know that different test sets provide different results...

CROSS VALIDATE

ALL THE THINGS

Cross Validation

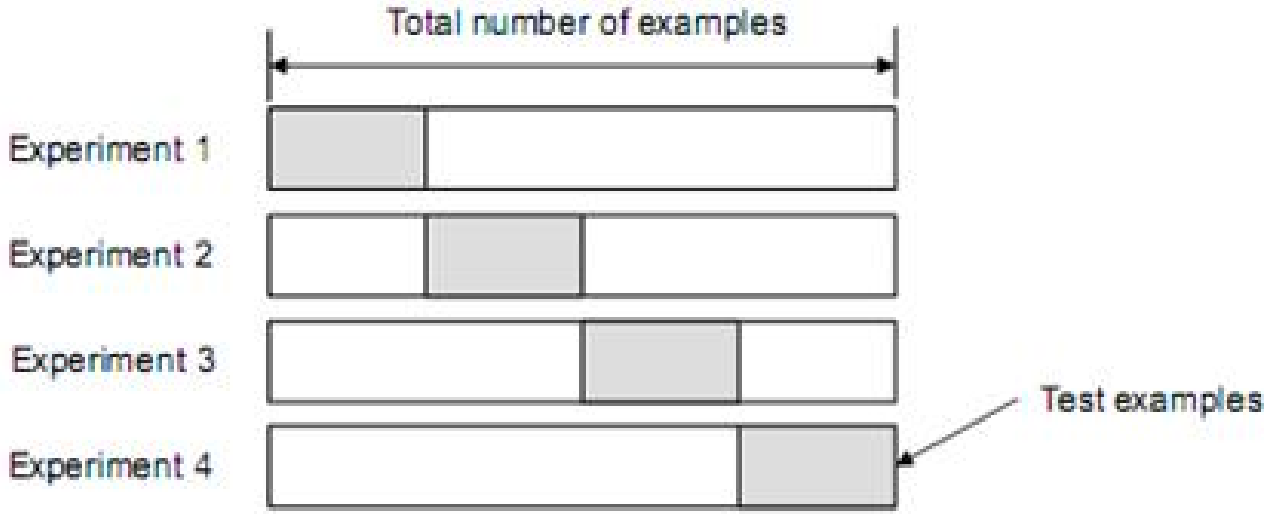
Assessing a model using different subsets of the data for training and testing

Examples

- N-fold cross validation
- Random sub-sampling validation
- Leave-one-out cross validation

N-fold cross validation

1. Randomly split the dataset into n equal groups
2. Use partition 1 as test set & union of other groups as training
3. Find generalization error
4. Repeat steps 2-3 using different group as test set at each iteration
5. Take average generalization error as estimate of OOS accuracy



Class break

III. KNN Classification

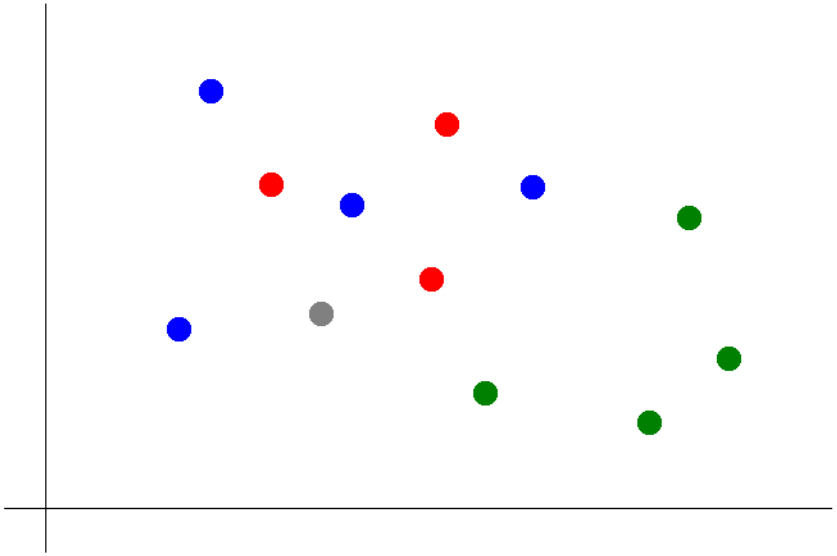
Goals

Define KNN Classification

Practice KNN Classification using Training and Test Set data

Suppose we want to predict the color of the gray dot.

1. Pick value for k
2. Find colors of k nearest neighbors
3. Assign most common color to gray dot



In Class Assignment

git push to your data science repo (save file as knn_classwork.R)

Extend the script we used in class to implement knn classification on the iris dataset using n-fold cross-validation.

CHALLENGE: Split the code into logical functions

```
knn.nfold <- function(n, … ) {
   # create n-fold partition of dataset
   # perform knn classification n times
   # n-fold generalization error = average over all iterations
}
```

Final Discussion