COMP 364: Computer Tools for Life Sciences

Notions of machine learning

Christopher J.F. Cameron and Carlos G. Oliver

Key course information

Assignment #3

- due tonight, November 15th at 11:59:59 pm

Assignment #4

- available now
- due Monday, November 27th at 11:59:59 pm
- first two parts can be completed now
  - remaining concepts will be taught over the next three lectures

Course evaluations

- available now at the following link:
  - `https://horizon.mcgill.ca/pban1/twbkwbis.P_WWWLogin?ret_code=f`

Problem: cat vs. bird

How would you write a computer program to identify a cat or bird in a photo?

Cats

Birds

Distinguishing features between cats and birds

There are some obvious features to distinguish cats and birds:

- **Cats**: fur, ears, a tail
- **Birds**: beaks, feathers, no teeth

How would you tell a computer to recognize a beak?

- fur? a tail?

Let's now say that we have:

- many example pictures of birds and cats (10s of thousands)
- each picture is labeled as either a bird or cat

How can we now learn to distinguish between cats and birds?

Machine learning

What is **machine learning (ML)**?

- the application of ML is currently a 'hot' trend in many fields
- everyone wants to perform it, but relatively few understand it

Google's DeepMind AI just taught itself to walk
`https://www.youtube.com/watch?v=gn4nRCC9TwQ`

What is ML?

ML is the study of computer algorithms that allow for learning

- a core subarea of **artificial intelligence (AI)**

Examples of learning problems:

- to complete a task
- make accurate predictions
- behave intelligently

Learning is always based on some sort of observation or data

- such as: examples, direct experience, or instruction
- **cat vs. bird**: many labeled pictures of cats and birds

What is ML? #2

In general, ML is about learning to do better in the future

- based on past experiences (**training data**)

The emphasis of ML is on **automatic methods**

- devise learning algorithms that do the learning automatically
- without human intervention or assistance
- can be viewed as 'programming by example'

Often we have a specific task in mind

- for example: identifying a cat or bird in a photo
- instead of creating a program to solve the task directly
- we implement methods where the computer learns to identify a cat or bird based on the provided examples

Why study ML?

It is unlikely that humans will develop a truly intelligent AI

- capable of facilities that we associate with intelligence
  - such as language or vision
- without using ML to get there
- these tasks are just too difficult to solve

We also would not consider a system to be truly intelligent

- if it were incapable of learning
- since learning is at the core of intelligence

Examples of ML

character recognition

- categorize images of handwritten characters by the letters represented

face detection

- find faces in images (or indicate if a face is present)

medical diagnosis

- diagnose a patient as a sufferer or non-sufferer of some disease
- predict the required dosage for successful treatment

fraud detection

- identify credit card transactions (for instance) which may be fraudulent in nature

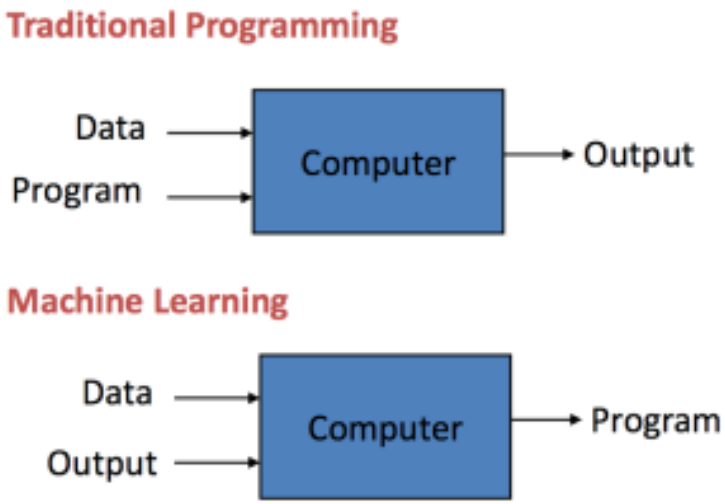
'Traditional' programming vs. ML

Traditional programming

- data and program are run on a computer to produce the output

ML

- data and output are run on a computer to create a program.
- program can then be used in traditional programming

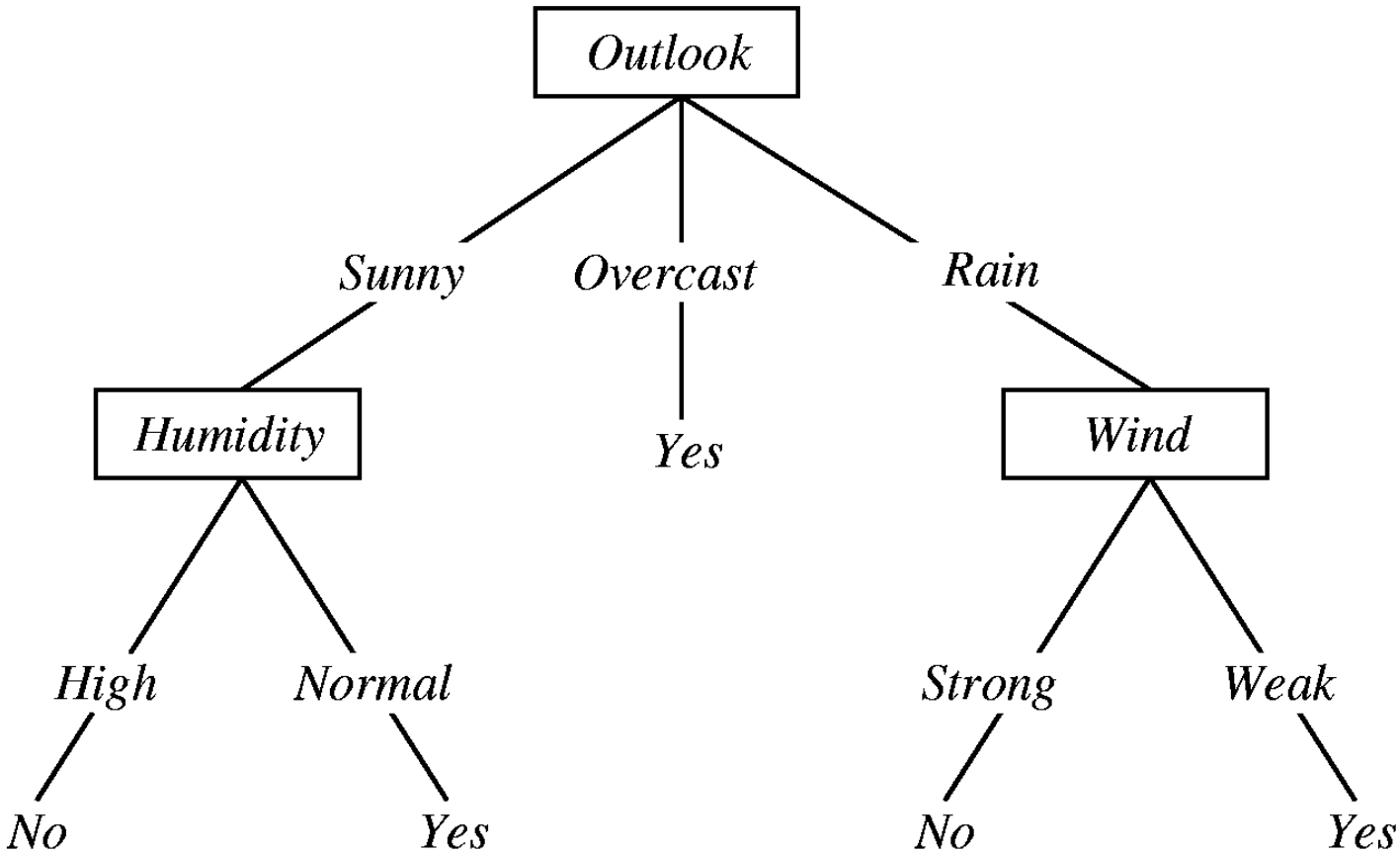


Types of ML algorithms

There are many types of ML algorithms:

- **logistic regression**: https://en.wikipedia.org/wiki/Logistic_regression
- **polynomial regression**: https://en.wikipedia.org/wiki/Polynomial_regression
- **decision tree**: https://en.wikipedia.org/wiki/Decision_tree
- **random forest**: https://en.wikipedia.org/wiki/Random_forest
- **artificial neural network**: https://en.wikipedia.org/wiki/Artificial_neural_network
- **support vector machine**: https://en.wikipedia.org/wiki/Support_vector_machine
- and many more...

Decision tree: to go outside or not



ML algorithm: decision tree

ML model: structure of decision tree

Key elements of ML

Every ML algorithm has three components:

1. **representation**: how to represent knowledge
   - what model should be chosen?
   - how to best structure data as input to the model

2. **evaluation**: the way to evaluate candidate programs
   - accuracy, prediction and recall, squared error, likelihood, etc.

3. **optimization**: the way in which candidate programs are generated is known as the 'search process'
   - there are infinite possible models to be chosen
   - how do we best select the ideal model?

Types of ML

Two common types of ML:

1. **supervised learning**: training data includes desired outputs
   - model input: a representation of a cat or bird photo
   - input label: either "cat" (0) or "bird" (1)

2. **unsupervised learning**: training data does not include desired outputs
   - e.g., clustering - it's hard to know what the correct grouping is

There are other types of ML

- in this course, we'll focus on the above two

Types of learning #2

Supervised learning is the most mature

- the most studied
- the type of learning used by most ML algorithms
- supervised learning is much easier than non-supervised
- also known as **inductive learning**

For inductive learning:

- the computer is provided examples from data (x)
- and the expected output of some function ($func(x)$)
- the goal of ML algorithm is to learn $func()$ for new data

Types of learned functions

There are three general types of functions for ML:

1. **classification**: the function being learned is discrete
   - identifying a cat or bird in a photo
2. **regression**: the function being learned is continuous
   - predicting stock market prices
3. **probability estimation**: the output of the function is a probability
   - will it rain tomorrow?

In practice, ML models perform regression functionality

- output is then transformed
  - discretized for classification
  - probabilities for probability estimation

Evaluating ML algorithms

How can we get an unbiased estimate of the accuracy for a learned model?

Using supervised learning

- split available data into **training** and **testing** datasets
- create a learned model from the training data
- apply learned model to testing data
- measure accuracy of model predictions

How would this work with our 'cat vs. bird' example?

Cat vs. bird ML example

total data: labeled pictures of cats and birds (50K each)

training data: labeled pictures of cats and birds (45K each)

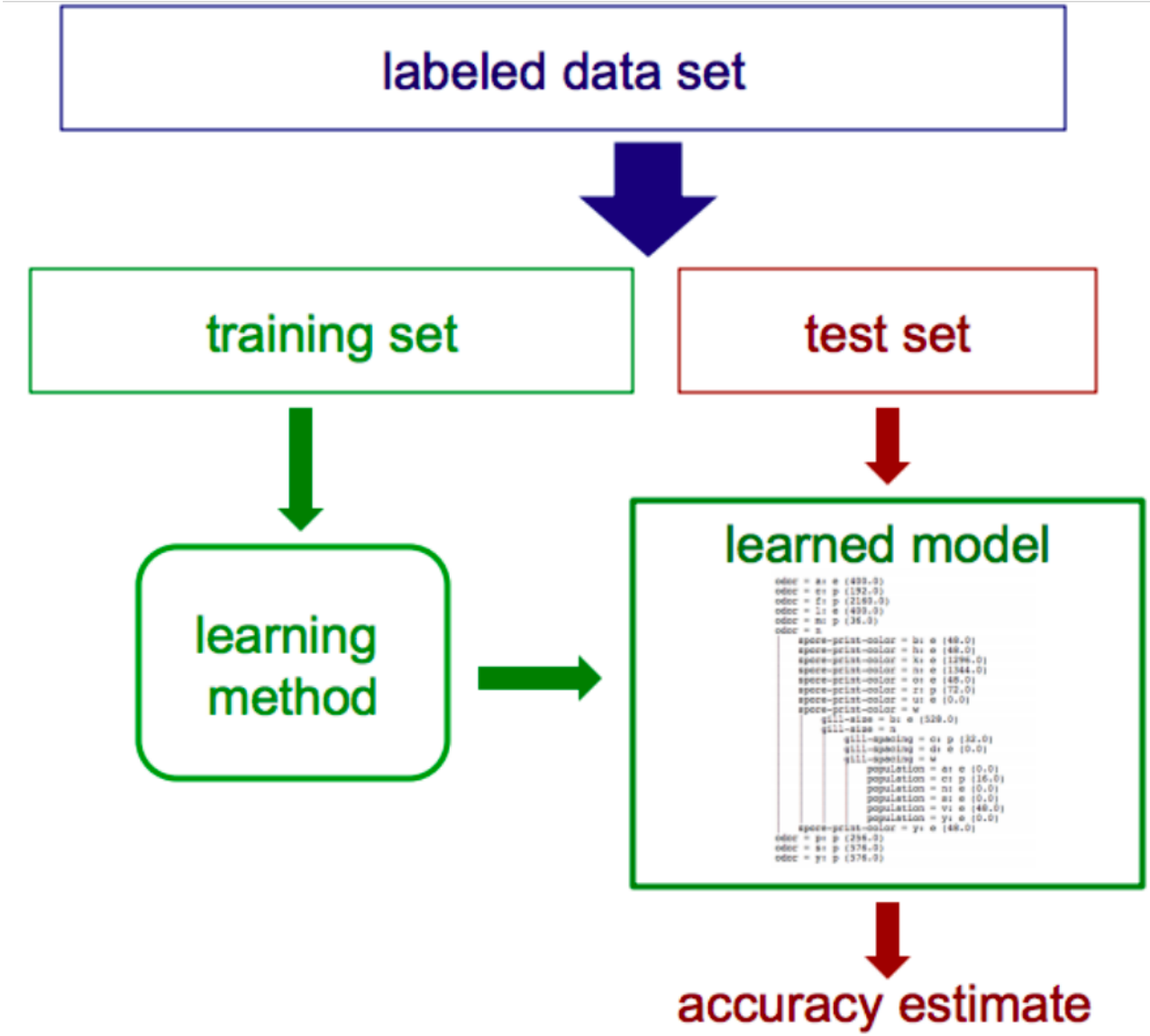
- model input is a representation of the example photo
- label is either '0' (cat) or '1' (bird)

testing data: labeled pictures of cats and birds (5K each)

ML steps:

1. create learned model from examples in training data
   - implement ML algorithm and apply to examples
2. predict on previously unseen examples
   - apply learned model to testing data
3. compare model predictions (0 or 1) against known labels
   - calculate accuracy measure

Evaluating ML algorithms #2



Python's scikit-learn module

Over the next two lectures

- we're going to perform some basic machine learning
- using Python's scikit-learn module

scikit-learn API:
`http://scikit-learn.org/stable/modules/classes.html`

scikit-learn tutorials:
`http://scikit-learn.org/stable/`

Extra - reinforcement learning example

MarI/O - Machine Learning for Video Games
https://www.youtube.com/watch?v=qv6UVOQ0F44