CPSC 540 - Machine Learning

Overview

Mark Schmidt

University of British Columbia

Fall 2014

Motivation

- We are entering the era of big data:
  - Tens of billions of webpages.
  - 100s of hours of YouTube videos every minute.
  - Sequenced genomes of 1000s of people, each containing billions of base-pairs.
  - Over 200 million products on Amazon.
  - Over 300 trillion experiments in the large hadron collider.

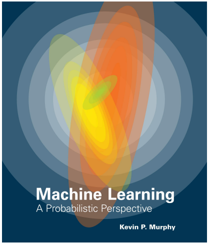
Motivation

- We are entering the era of big data:
  - Tens of billions of webpages.
  - 100s of hours of YouTube videos every minute.
  - Sequenced genomes of 1000s of people, each containing billions of base-pairs.
  - Over 200 million products on Amazon.
  - Over 300 trillion experiments in the large hadron collider.
- We need automated data analysis.

Motivation

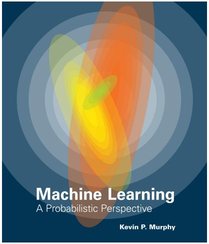
- We are entering the era of big data:
  - Tens of billions of webpages.
  - 100s of hours of YouTube videos every minute.
  - Sequenced genomes of 1000s of people, each containing billions of base-pairs.
  - Over 200 million products on Amazon.
  - Over 300 trillion experiments in the large hadron collider.
- We need automated data analysis.
- This overview roughly follows Chapter 1 of MLAPA.

Machine Learning



- *Study of using computers to automatically detect patterns in data, and use these to make predictions or decisions.*

Machine Learning

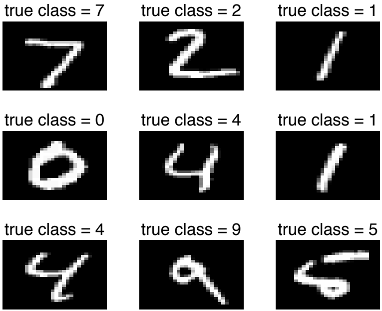


- *Study of using computers to automatically detect patterns in data, and use these to make predictions or decisions*.

- One of the fastest-growing areas of science/engineering.
- Recent successes: Kinect, book/movie recommendation, spam detection, credit card fraud detection, face recognition, speech recognition, object recognition, self-driving cars.
- Many more applications to be discovered!

Types of Machine Learning

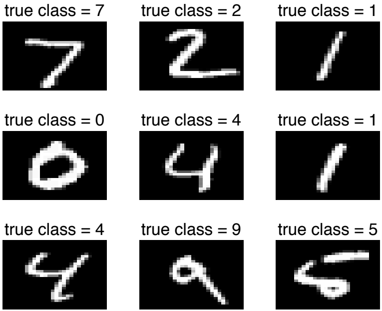
- Supervised learning:



  - Given input and output examples.
  - Build a model that predicts the output from the inputs.
  - You can use the model to predict the output on new inputs

Types of Machine Learning

- Supervised learning:



  - Given input and output examples.
  - Build a model that predicts the output from the inputs.
  - You can use the model to predict the output on new inputs

- Called regression with continous outputs,
  and classification with discrete outputs.

Types of Machine Learning

- Comments on supervised learning:
  - Most common type of machine learning.

Types of Machine Learning

- Comments on supervised learning:
  - Most common type of machine learning.
  - Useful you have a well-defined pattern recognition problem but don't know how to solve it,

Types of Machine Learning

- Comments on supervised learning:
  - Most common type of machine learning.
  - Useful you have a well-defined pattern recognition problem but don't know how to solve it,
  - And you have lots of labeled data.

Types of Machine Learning

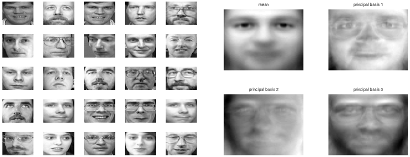
- Comments on supervised learning:
  - Most common type of machine learning.
  - Useful you have a well-defined pattern recognition problem but don't know how to solve it,
  - And you have lots of labeled data.
  - Many variants:
    - Multiple regression and multi-label classificaiton.
    - Multi-task variants.
    - Collaborative filtering.
    - Structured prediction.

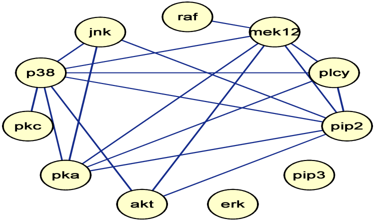
Types of Machine Learning

- Comments on supervised learning:
  - Most common type of machine learning.
  - Useful you have a well-defined pattern recognition problem but don't know how to solve it,
  - And you have lots of labeled data.
  - Many variants:
    - Multiple regression and multi-label classificaiton.
    - Multi-task variants.
    - Collaborative filtering.
    - Structured prediction.
  - Key reason for machine learning's popularity and success.
  - Major focus of this course.
  - But, unsupervised variants based on similar principles.

Types of Machine Learning

- Unsupervised learning:





- Given data, discover ‘patterns’.
- Could be simple model that reproduces data.
- Could be relationships between the variables.
- Could be relationships between the data.

Types of Machine Learning

- Reinforcement learning:
  - We have an agent that can perform actions.
  - We give it rewards and punishments.
  - Not covered in this course.
  - But supervised/unsupervised methods often key component.

Basic Concepts

- Parametric models:
  - have a fixed number of parameters.
  - Examples: naive Bayes, linear regression.
- Non-parametric models:
  - number of parameters grows with the data size.
  - Examples: k-nearest neighbors, kernel regression.

Basic Concepts

- Parametric models:
  - have a fixed number of parameters.
  - Examples: naive Bayes, linear regression.
- Non-parametric models:
  - number of parameters grows with the data size.
  - Examples: k-nearest neighbors, kernel regression.
- Generalization error:
  - in machine learning focuses on predictions for new data.
  - we can estimate this using a validation set.
- Over-fitting:
  - if we have too many parameters, we may fit noise in the data.
  - we can combat this with model selection or regularization.

Basic Concepts

- Common models for classification:
  - Naive Bayes.
  - k-nearest neighbors.
  - Logistic regression.
  - Support vector machines.
  - Random Forests.
  - Gaussian processes.
  - Neural networks.
- No free lunch theorem:
  - There is no single best model that works optimally for all kinds of problems [Wolpert, 1996].
  - Model that works in one domain may work poorly in another.
  - In this course we'll look at a variety of models/assumptions.
  - "All models are wrong, but some are useful" - Box.

Basic Concepts

- How should we evaluate a classifier?

Basic Concepts

- How should we evaluate a classifier?
  - Empirically estimate generalization error.

Basic Concepts

- How should we evaluate a classifier?
  - Empirically estimate generalization error.
  - Number of parameters.

Basic Concepts

- How should we evaluate a classifier?
  - Empirically estimate generalization error.
  - Number of parameters.
  - Training time.
  - Testing time.

Basic Concepts

- How should we evaluate a classifier?
  - Empirically estimate generalization error.
  - Number of parameters.
  - Training time.
  - Testing time.
  - Model flexibility.
  - How much data is needed?