Machine Learning and Data Mining I: Lecture 1

Morgan McCarty

03 July 2023

1 Why Machine Learning?

1.1 An Example of Machine Learning in Action

- The Kakapo is an endangered bird with a genetic disposition for disease
- Machine Learning can be used to predict whether future Kakapos will be susceptible to disease by analyzing their genome

1.2 The Use of Machine Learning

- Data is ubiquitous
- Predictions can be made using Data
- Past situations can be analyzed to predict future situations
- Machine Learning is a set of tools for understanding data

2 Machine Learning: High Level

2.1 Questions

- Can we really make our machines learn?
  - With lots of Data
  - Rather than program expertise, we program the ability to learn from Data
  - We can build machine learning models and use them to make remarkably accurate predictions

2.2 How does it work?

- The end goal is prediction
- Most machine-learning models are trained to make predictions
- This can be a simple model that uses a single variable (e.g. location) or a more complex model that uses many variables (e.g. location, time, weather, etc.)

2.3 The Three Main Approaches

- Classification (e.g. spam or not spam) [SVM, nearest neighbors, etc]
  Identify which category an object belongs to
- Regression (e.g. price of a house) [linear regression, random forest, SVR, etc]
  Predicting a continuous-valued attribute associated with an object
- Clustering (e.g. customer segmentation) [k-means, spectral clustering, a-priori, etc]
  Automatic grouping of similar objects into sets

2.4 Types of Machine Learning

- **Supervised Learning**
  A known input and output is supplied and the algorithm learns a general rule to map the input to the output
- **Unsupervised Learning**
  The model arrives at conclusions and determines patters through unlabeled data
- Semi-Supervised Learning
  The model is built with a mix of labeled and unlabeled data - e.g. sets of categories, suggestions, and example labels
- Reinforcement Learning
  A system is used to cause the model to learn through trial and error using rewards and punishments

2.5 Focus of ML1

- The main focus of this course is on supervised learning (classification and regression) with potential to additionally cover dimensionality reduction

2.6 ML Recipe

- Find a dataset
- Explore the dataset for possible patterns or ideas
- Split the dataset into training and testing sets
- Build a model
- Train the model

2.7 Main Topics

- Linear Regression for predictions
- Classification (probabilistic and non-probabilistic)
- Decision Trees
- Neural Networks
- Model selection, timing, optimization

- Python for Machine Learning (Algorithms, tools, packages)
- Ethics for Machine Learning

2.8 Goals

- Understand core machine-learning concepts and Approaches
- Implemeents algorithms from scratch
- Evaluate models using metrics appropriate for the given task/problem
- Select and apply appropriate techniques
- Understand the standard approaches
- Understand potential ethical issues and make sure decisions are appropriate for the task

3 History of Machine Learning

3.1 Timeline

- AI $\approx 1950 - 1980$
- ML $\approx 1980 - 2010$
- Deep Learning $\approx 2010 - \text{Now}$
- Began with Bayes Theorem
- Present day is dominated by Generative AI

4 Methods

4.1 Linnear Regression

- Introduced by Sir Francis Galton (1822-1911) in 1886
- "Regresson towards mediocrity" (regression to the mean)
- $y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$
- Where for $i = 1, \ldots, n$ observations:
  - y_i is the response variable (dependent variable)
  - $x_i^1, \ldots, x_i^p$ are the predictor variables (independent variables) / explanatory variables
  - $\beta_0, \ldots, \beta_p$ are the parameters
  - β_0 is the intercept
  - β_p is the slope
  - ϵ_i is the error term (aka the residual)
- e.g. for house prices:

  - y_i is the price of the house
  - x_i^1 is the size of the house
  - x_i^2 is the number of rooms
  - x_i^3 is the zip-code
- The goal is to find the best fit line (i.e the line that minimizes the sum of the squared errors) to predict future values

4.2 Classification

- Input new data with a set of categories (e.g. a 2D dataset with shapes)
- Pass through a classifier to determine which category the data belongs to
- Output the type of the datapoint
- e.g. Orchids (Iris Dataset)
  - x_i^1 is the sepal length
  - x_i^2 is the sepal width
  - x_i^3 is the petal length
  - x_i^4 is the petal width
  - The output is the type of orchid (setosa, versicolor, or virginica)

4.3 Clustering

- Input new data with no set of categories, but that has some type of grouping
- Pass through a clustering algorithm to determine which group the data belongs to
- e.g. anomaly detection
- Uses a distance or simularity metric to determine which group the data belongs to
- Outputs labels for the data that were not predefined

4.4 Neural Networks and Deep Learning

- Introduced with the perceptron
- Inputs are passed through a series of layers
- Weights are applied to the inputs
- The inputs are summed and passed through an activation function
- An output is produced
- The perceptron evolved into the neural network and, later, deep learning through the addition of more layers (multilayer perceptron)

4.5 Genetic Algorithms

- Inspired by cellular reproduction and is used to optimize and train
- A population of solutions is created and evaluated with a fitness function
- The best solutions are selected and used to create a new population
- This process is repeated until a solution is found or the maximum number of generations is reached

4.6 Ethics of Machine Learning

- Hidden variables can lead to bias when models are trained using data that is either not representative of the population or that is not diverse enough (or the population is not diverse enough and the model is applied to a diverse population)
- As such it is important to consider the ethical implications of the data used to train models and the models themselves