Practical Machine Learning in R

Introduction

Lars Kotthoff$^{1,2}$
larsko@uwyo.edu

$^1$with slides from Bernd Bischl and Michel Lang
$^2$slides available at http://www.cs.uwyo.edu/~larsko/ml-fac
What is Machine Learning?

▷ “gives computes the ability to learn without being explicitly programmed” (Wikipedia)
What is Machine Learning?

▷ “gives computes the ability to learn without being explicitly programmed” (Wikipedia)
▷ “A computer program is said to learn from experience \( E \) with respect to some class of tasks \( T \) and performance measure \( P \) if its performance at tasks in \( T \), as measured by \( P \), improves with experience \( E \).” (Tom Mitchell)
Examples
Examples
Examples

https://pythonprogramming.net/forecasting-predicting-machine-learning-tutorial/
Examples

Supervised Learning

- learn the relationship between input $x$ and output $y$
- training data with labels available — $y$ known for given $x$
- can see this as function approximation — find an $f$ such that

$$y \approx f(x)$$
Supervised Learning

- $x$ are features or attributes
- $y$ is the ground truth
- denote predictions $f(x) = \hat{y}$
- loss function $L(y, \hat{y})$ measures how good predictions are, e.g.
  \[
  L(y, \hat{y}) = (y - \hat{y})^2
  \]
- want to minimize loss given training data $X_{\text{train}} = \{(x_i, y_i)\}^n$:
  \[
  \arg \min \sum_{i=1}^{n} L(y_i, \hat{y}_i)
  \]
Supervised Learning

▷ want to learn a general function that is predictive on new data
▷ second set $X_{test}$ that is not used in training to test generalization performance:

$$\sum_{i=1}^{n} L(y_i, \hat{y}_i)$$

▷ usually full data set $X$ is split into non-overlapping train and test sets:

$$X_{train} \cup X_{test} = X$$
$$X_{train} \cap X_{test} = \emptyset$$
**Supervised Classification**

**Goal**: Predict a class (discrete quantity), or membership probabilities
Supervised Regression

Goal: Predict a continuous quantity
Unsupervised Learning

- no ground truth $y$ available
- determine group membership or assign labels
- loss function measures properties of groups, e.g. homogeneity wrt. features
- still want to minimize loss given training data and generalize
Unsupervised Clustering

Goal: Group data by similarity, or estimate membership probabilities
In this Course

▷ classification
▷ regression
▷ clustering
▷ data preprocessing (missing values, dimensionality reduction)
▷ performance evaluation
▷ parameter tuning
Not in this Course

- R tutorial
- details on particular methods
- deep learning
- time series
- Big Data
What you’ll need
Install RStudio

https://www.rstudio.com/products/rstudio/download/
Install mlr

▷ on the R console:

```r
install.packages("mlr")
```

▷ or see http://derekogle.com/IFAR/supplements/installations/InstallPackagesRStudio.html

Format

- meetings roughly every week
- half lecture, half practical exercises
- happy to discuss specific problems