What is machine learning and what can it teach us about prostate cancer?

Machine learning & neural networks

Anne Helby Petersen

Today's program

L1 What is machine learning and what can it teach us about prostate cancer?

Today's program

- L1 What is machine learning and what can it teach us about prostate cancer?
- E1 Make your first (stupid) machine learning models

Today's program

L1 What is machine learning and what can it teach us about prostate cancer?

E1 Make your first (stupid) machine learning models

L2 A slightly smarter machine: Using logistic regression

Today's program

- L1 What is machine learning and what can it teach us about prostate cancer?
- E1 Make your first (stupid) machine learning models
- L2 A slightly smarter machine: Using logistic regression
- E2 Training with logistic regression

Today's program

- L1 What is machine learning and what can it teach us about prostate cancer?
- E1 Make your first (stupid) machine learning models
- L2 A slightly smarter machine: Using logistic regression
- E2 Training with logistic regression
- L3 Introduction to neural networks

Today's program

- L1 What is machine learning and what can it teach us about prostate cancer?
- E1 Make your first (stupid) machine learning models
- L2 A slightly smarter machine: Using logistic regression
- E2 Training with logistic regression
- L3 Introduction to neural networks

  Lunch ($\simeq$ 11.15-12.00)

Today's program

- L1 What is machine learning and what can it teach us about prostate cancer?
- E1 Make your first (stupid) machine learning models
- L2 A slightly smarter machine: Using logistic regression
- E2 Training with logistic regression
- L3 Introduction to neural networks
- Lunch ($\simeq$ 11.15-12.00)
- L3 Introduction to neural networks (continued)
- E3 Train neural networks
- L4 Introduction to deep learning: More tools for NNs
- E4 Train more neural networks with your brand new tools
- L5 What can you do next?

A few pointers for today

- More **workshop** than "classical" course: Focus on you trying stuff out in practice, not on theory.
  - A lot of exercises - decide for yourself if you want to focus on a few or go quicker through more.

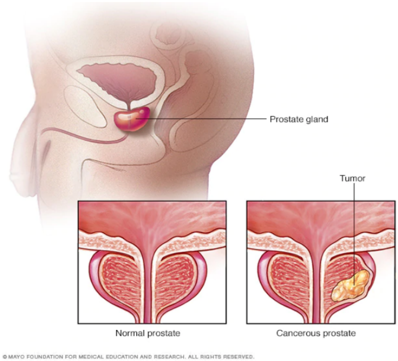
A few pointers for today

- More **workshop** than "classical" course: Focus on you trying stuff out in practice, not on theory.
  - A lot of exercises - decide for yourself if you want to focus on a few or go quicker through more.
- You will be working in R, sometimes with **semi-advanced code**.
  - I don't expect you to be programmers!
  - Try to see if you can make sense of the code.
  - Ask questions!

A few pointers for today

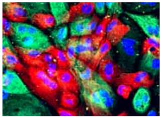
- More **workshop** than "classical" course: Focus on you trying stuff out in practice, not on theory.
  - A lot of exercises - decide for yourself if you want to focus on a few or go quicker through more.
- You will be working in R, sometimes with **semi-advanced code**.
  - I don't expect you to be programmers!
  - Try to see if you can make sense of the code.
  - Ask questions!
- We will work on a **difficult real life classification problem**
  - This is not a textbook example.
  - It may be challenging to get anywhere.
  - I cannot promise you that you will be making very clever machines today.

Our problem for today: Predict prostate cancer patient survival



- Data from comparison arms from 3 phase III clinical trials for metastatic castrate resistant prostate cancer patients.
- A total of 1495 patients with 95 measured variables.
- **Goal: Predict whether a patient dies within 2 years.**

Prostate cancer DREAM challenge



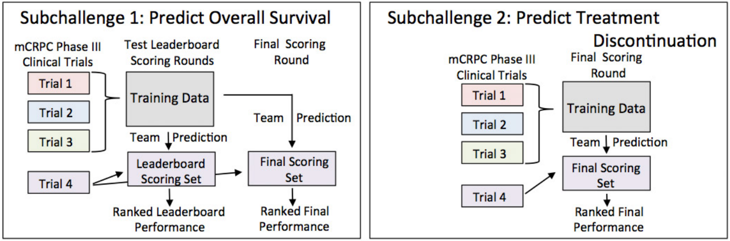
DREAM 9.5 Prostate Cancer DREAM Challenge

DREAM 9.5 Prostate Cancer DREAM Challenge

March 16- July 27, 2015

This challenge focused on predicting survival for prostate cancer patients based on patients' clinical variables.

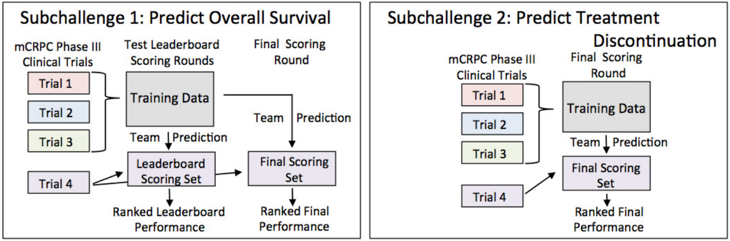
A prediction competition



Two subchallenges:

1. Predicting whether a patient is registered as "dead" within 2 years of the study *(tricky)*
2. Predicting whether a patient's treatment is discontinued within 3 months of the study due to adverse effects (Subchallenge 2 outcome) *(trickier)*

A prediction competition



Two subchallenges:

1. Predicting whether a patient is registered as "dead" within 2 years of the study *(tricky)*
2. Predicting whether a patient's treatment is discontinued within 3 months of the study due to adverse effects (Subchallenge 2 outcome) *(trickier)*

Results from the DREAM challenges

Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data

Justin Guinney*, Tao Wang*, Teemu D Laajala*, Kimberly Kanigel Winner, J Christopher Bare, Elias Chaibub Neto, Suleiman A Khan, Gopal Peddinti, Antti Airola, Tapio Pahikkala, Tuomas Mirtti, Thomas Yu, Brian M Bot, Liji Shen, Kald Abdallah, Thea Norman, Stephen Friend, Gustavo Stolovitzky, Howard Soule, Christopher J Sweeney, Charles J Ryan, Howard I Scher, Oliver Sartor, Yang Xie†, Tero Aittokallio†, Fang Liz Zhou†, James C Costello†, and the Prostate Cancer Challenge DREAM Community‡

Lancet Oncol 2017; 18: 132–42
Published Online November 15, 2016
http://dx.doi.org/10.1016/S1470-2045(16)30560-5
See Comment page 15
*Contributed equally as first

Summary

Background Improvements to prognostic models in metastatic castration-resistant prostate cancer have the potential to augment clinical trial design and guide treatment strategies. In partnership with Project Data Sphere, a not-for-profit initiative allowing data from cancer clinical trials to be shared broadly with researchers, we designed an open-data, crowdsourced, DREAM (Dialogue for Reverse Engineering Assessments and Methods) challenge to not only identify a better prognostic model for prediction of survival in patients with metastatic castration-resistant prostate cancer but also engage a community of international data scientists to study this disease.

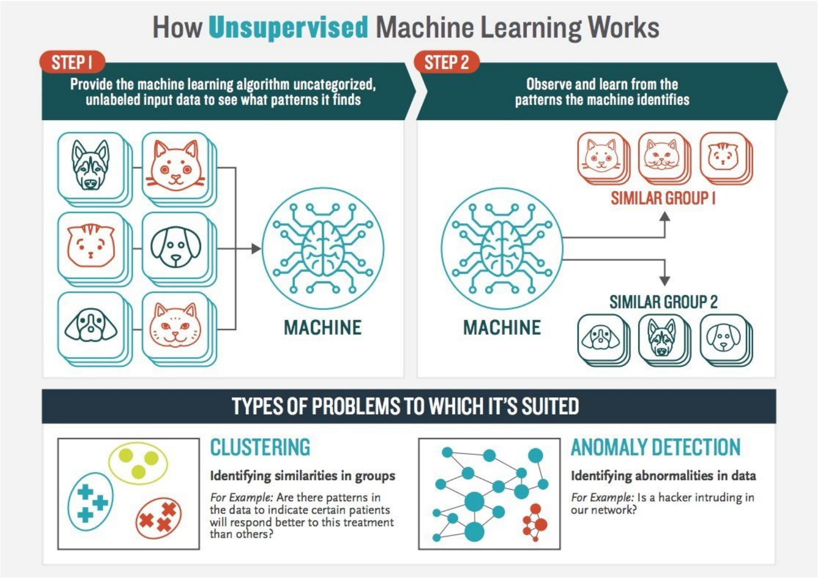
original report

A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer

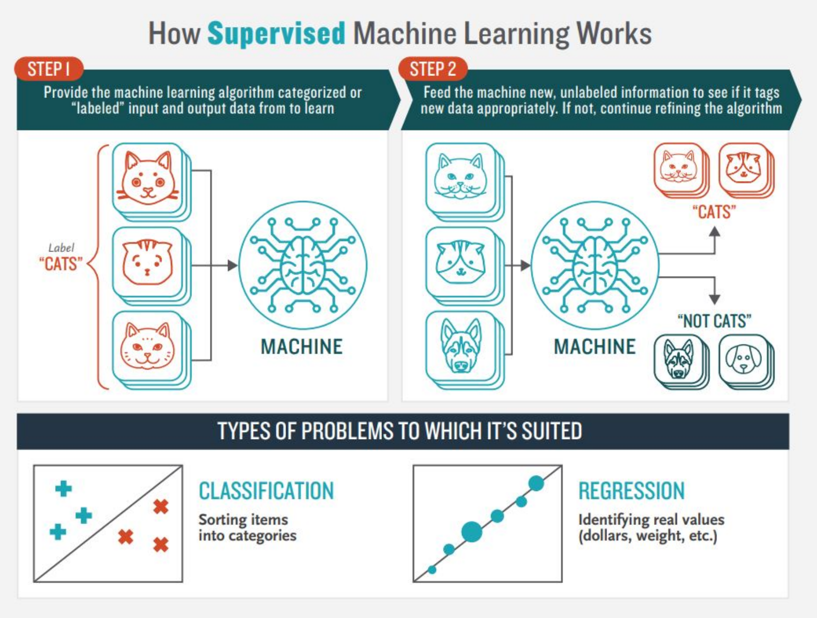
abstrac

Purpose Docetaxel has a demonstrated survival benefit for patients with metastatic castration-resistant prostate cancer (mCRPC); however, 10% to 20% of patients discontinue docetaxel prematurely because of toxicity-induced adverse events, and the management of risk factors for toxicity remains a challenge.

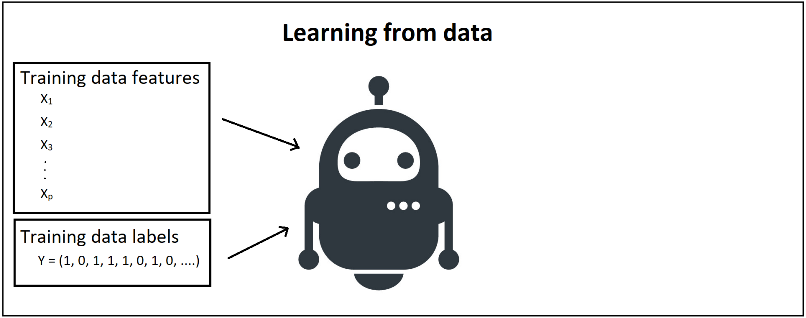
Supervised learning vs. unsupervised learning



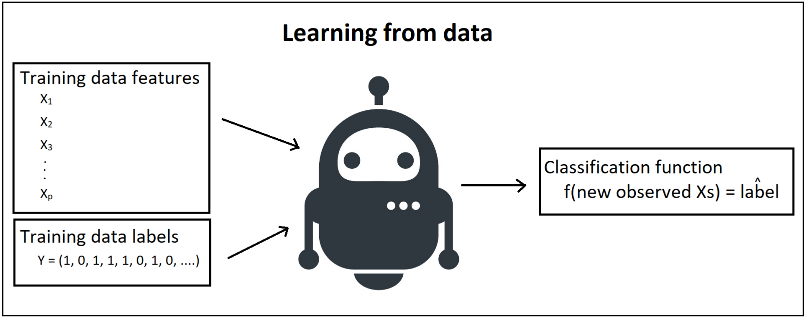
Supervised learning vs. unsupervised learning



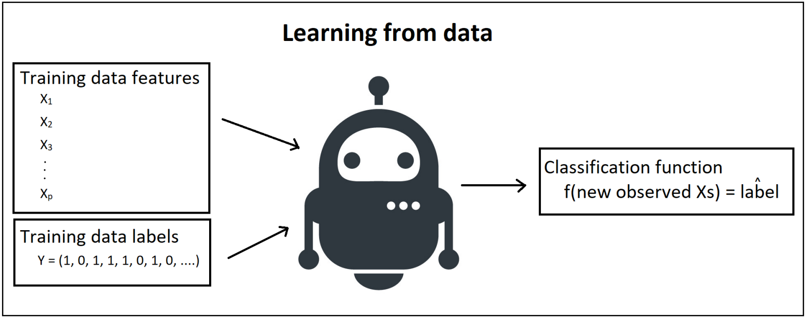
Supervised machine learning workflow

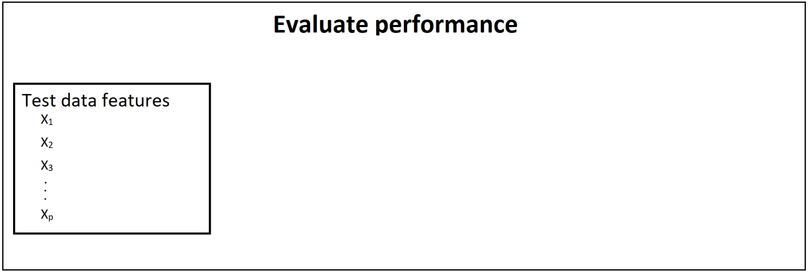


Supervised machine learning workflow

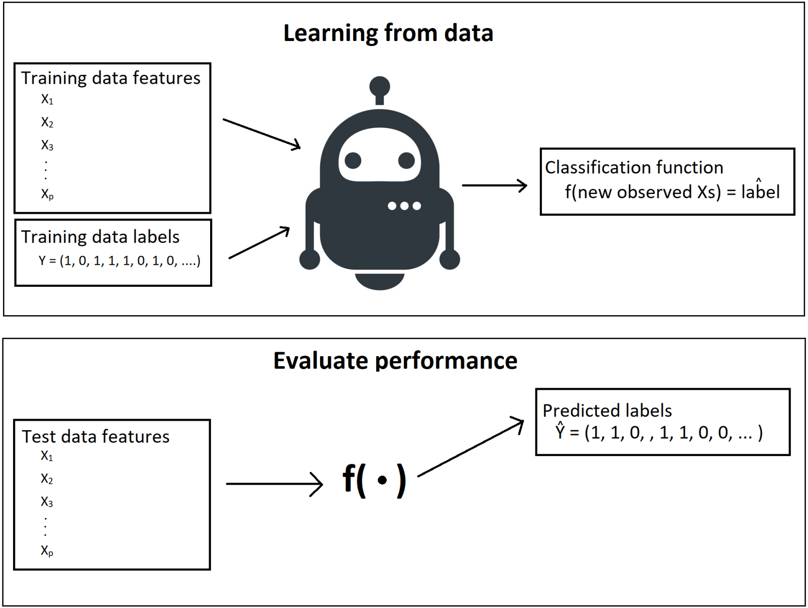


Supervised machine learning workflow

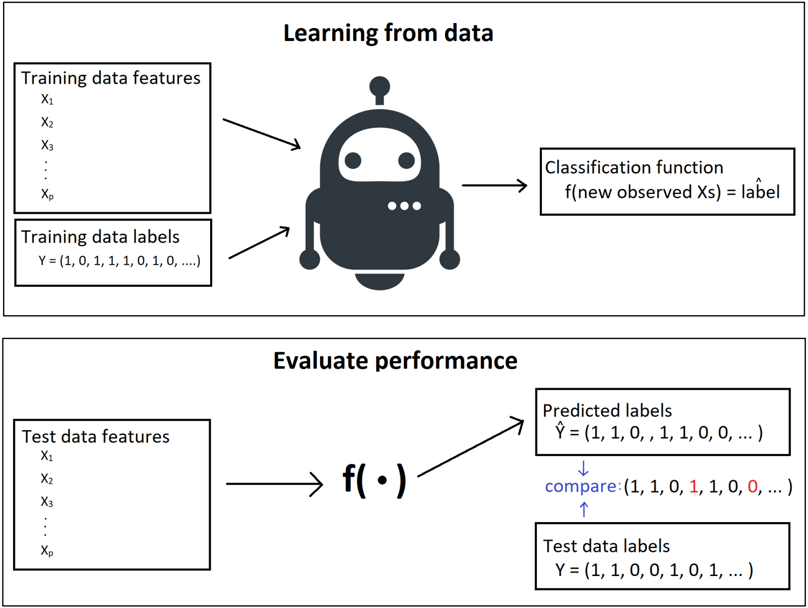




Supervised machine learning workflow



Supervised machine learning workflow



Supervised learning: evaluating performance

- Performance can be evaluated e.g. by looking at the *accuracy*:

$$\frac{\#Y \text{ from test data is equal to } \hat{Y}}{\#\text{observations in test data}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} 1_{(Y_i = \hat{Y}_i)}$$

- Note: This is very different from "classical" statistics. **There is a true answer and we know it!**

Supervised learning: 2 rules for building your machine

Supervised learning: 2 rules for building your machine

1. Do not touch the test data when training the machine

Supervised learning: 2 rules for building your machine

1. Do not touch the test data when training the machine
2. **Do not touch the test data when training the machine**

Machine learning 101

Machine learning for prostate cancer survival

- Goal: Given data on a *new* patient you haven't seen before, predict whether he will die within 2 years.
- We will measure performance primarily in terms of accuracy. **The best machine is the one that achieves the highest accuracy on the test data.**

Our first machine: George Sr.

Machine name:

Evaluation of machine

Died in real life?										
Died according to machine?										

$$\text{Accuracy} = \frac{number\ correctly\ classified}{number\ of\ patients\ in\ test\ data} = \frac{\quad}{10} =$$

A digital machine: George Jr.

```
george <- function(data_x, y) {
  #Learning from data would have happened here
  #if George had bothered doing it

  predictFunction <- function(newdata) {
    #George flips a coin for each observation in "newdata"
    #to see if he should return 1 or 0 as their label
    ys <- sample(c(1,0), size = nrow(newdata),
                 prob = c(0.5, 0.5),
                 replace = TRUE)
    return(ys)
  }
  #return the prediction function
  return(predictFunction)
}
```

Load and look at the mCRPC data

```r
load("./data/andata.rda")
```

```r
dim(traindata_x)
```

```
## [1] 1203   91
```

```r
dim(testdata_x)
```

```
## [1] 292  91
```

```r
table(traindata_DEATH2YRS)
```

```
## traindata_DEATH2YRS
##   0   1
## 769 434
```

Evaluate George's performance

```r
#train George
george_predict <- george(traindata_x, traindata_DEATH2YRS)

#predict labels for test data
george_guesses <- george_predict(testdata_x)

#compute accuracy
mean(george_guesses == testdata_DEATH2YRS)
```

```
## [1] 0.4965753
```

More information about the data

- Data from comparison arms from 3 phase III clinical trials for metastatic castrate resistant prostate cancer patients.
- I have prepared ready to go data (e.g. no missing information): `andata.rda`.
- Look in the codebook (`codebook_mCRPCdata.pdf`) for more information about the features you can use.
- Note: Categorical variables (`ECOG` and `AGEGRP`) are coded as dummies:

```
table(ECOG1 = traindata_x$ECOG_1); table(ECOG2 = traindata_x$ECOG_2)
```

```
## ECOG1
##   0   1
## 617 586
```

```
## ECOG2
##    0    1
## 1148   55
```

Time to build your first machine!

Go to the course website and find exercise session 1:

Exercise session 1

Machine learning & neural networks

Anne Helby Petersen

May 9, 2019

Overview

The goal of this exercise session is to:

- Get an overview of the data
- Try training your first machine to classify new observations

1.1. Load the data and look at it

1.1.1: Load the data into R.

If you're working within the `NNday` project, this can be done using the following line of code:

```
load("./data/andata.rda")
```

You will now have six object available in your workspace:

- `traindata_x` : the feature variables in the training dataset
- `traindata_DEATH2YRS` : the `DEATH2YRS` outcome variable from the training dataset