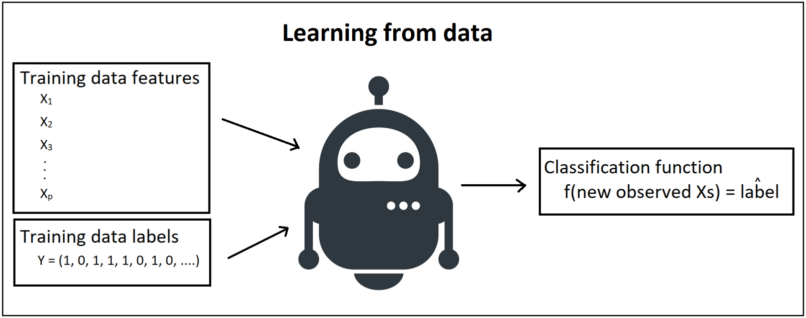
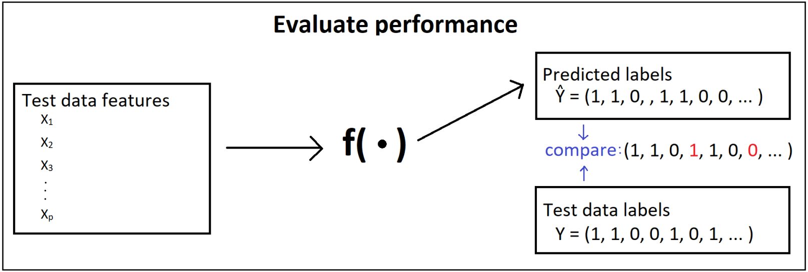
A slightly smarter machine: Using logistic regression

Machine learning & neural networks

Anne Helby Petersen

Going beyond manual machines



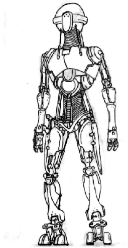


Going beyond manual machines

Learning from data

Training data features
X_1
X_2
X_3
⋮
X_p

Training data labels
Y = (1, 0, 1, 1, 1, 0, 1, 0, ...)



Classification function
f(new observed Xs) = $\hat{\text{label}}$

Evaluate performance

Test data features
X_1
X_2
X_3
⋮
X_p

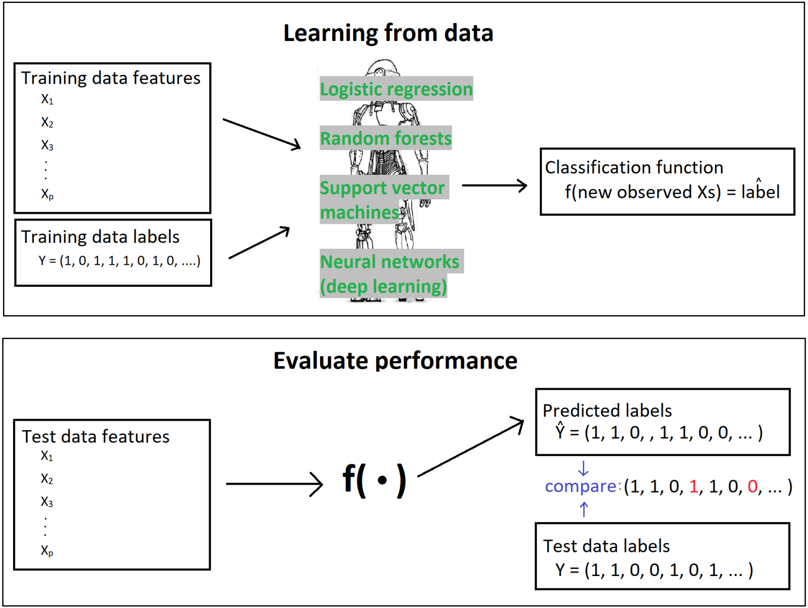
$f(\cdot)$

Predicted labels
$\hat{Y}$ = (1, 1, 0, , 1, 1, 0, 0, ...)

↓
compare: (1, 1, 0, 1, 1, 0, 0, ...)
↑

Test data labels
Y = (1, 1, 0, 0, 1, 0, 1, ...)

Going beyond manual machines



Logistic regression

A logistic regression model for $Y = \text{DEATH2YRS}$:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_d X_d$$

and we can *learn* (estimate) the values of $\alpha, \beta_1, ..., \beta_d$ from the training data by use of maximum likelihood estimation.

Logistic regression: fitting in R

We fit the model

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \alpha + \beta_1 \text{AST}$$

as follows (AST: Aspartate aminotransferase, measure of liver damage):

```
m1 <- glm(traindata_DEATH2YRS ~ AST, data = traindata_x,
          family = "binomial")

coef(m1)
```

```
## (Intercept)         AST
##  -1.3497249   0.0276447
```

and we see that $\hat{\alpha} = -1.35$ and $\hat{\beta} = 0.03$.

Logistic regression: Making predictions for new patient with AST $= 14$

We can predict log odds of Y by inserting $\hat{\alpha}, \hat{\beta}, 14$ in the model:

$$\log\left(\frac{\hat{P}(Y = 1 \,|\, \text{AST} = 14)}{1 - \hat{P}(Y = 1 \,|\, \text{AST} = 14)}\right) = \hat{\alpha} + \hat{\beta}_1 \cdot 14$$

Logistic regression: Making predictions for new patient with AST = 14

We can predict log odds of Y by inserting $\hat{\alpha}, \hat{\beta}, 14$ in the model:

$$\log\left(\frac{\hat{P}(Y = 1 \mid \text{AST} = 14)}{1 - \hat{P}(Y = 1 \mid \text{AST} = 14)}\right) = \hat{\alpha} + \hat{\beta}_1 \cdot 14$$
$$= -1.35 + 0.03 \cdot 14$$

Logistic regression: Making predictions for new patient with AST $= 14$

We can predict log odds of Y by inserting $\hat{\alpha}, \hat{\beta}, 14$ in the model:

$$\log\left(\frac{\hat{P}(Y = 1 \,|\, \text{AST} = 14)}{1 - \hat{P}(Y = 1 \,|\, \text{AST} = 14)}\right) = \hat{\alpha} + \hat{\beta}_1 \cdot 14$$
$$= -1.35 + 0.03 \cdot 14$$
$$= -0.96$$

These are the predicted *log odds* for the new patient.

From log odds to probabilities

Note that

$$\log\left(\frac{p}{1-p}\right) = k \qquad \Leftrightarrow \qquad p = \frac{\exp(k)}{1+\exp(k)}$$

From log odds to probabilities

Note that

$$\log\left(\frac{p}{1-p}\right) = k \qquad \Leftrightarrow \qquad p = \frac{\exp(k)}{1+\exp(k)}$$

so we find that

$$\begin{aligned}\hat{P}(Y = 1 \,|\, \text{AST} = 14) &= \frac{\exp(-0.96)}{1+\exp(-0.96)} \\ &= 0.28\end{aligned}$$

which is the *predicted probability* of the new patient dying within 2 years.

From probabilities to labels

We can make a classifier by using a *decision rule*:

$$\hat{Y}_{\text{new}} = f(\text{AST}_{\text{new}}) = \begin{cases} 1 & \text{if } \hat{P}(Y = 1 \,|\, \text{AST} = \text{AST}_{\text{new}}) > 0.5 \\ 0 & \text{if } \hat{P}(Y = 1 \,|\, \text{AST} = \text{AST}_{\text{new}}) \leq 0.5 \end{cases}$$

From probabilities to labels

We can make a classifier by using a *decision rule*:

$$\hat{Y}_{\text{new}} = f(\text{AST}_{\text{new}}) = \begin{cases} 1 & \text{if } \hat{P}(Y = 1 \,|\, \text{AST} = \text{AST}_{\text{new}}) > 0.5 \\ 0 & \text{if } \hat{P}(Y = 1 \,|\, \text{AST} = \text{AST}_{\text{new}}) \leq 0.5 \end{cases}$$

For $\text{AST}_{\text{new}} = 14$, where $\hat{P}(Y = 1 \,|\, \text{AST} = 14) = 0.28$, we thus guess at the label

$$\hat{Y}_{\text{new}} = f(14) = 0$$

so we *don't* think the new patient with AST level 14 will die within 2 years.

Using R for predictions and classifciation based on logistic regression model

```
preds <- predict(m1,
            newdata = testdata_x[, "AST", drop = FALSE],
            type = "response")
```

Using R for predictions and classifciation based on logistic regression model

```
preds <- predict(m1,
             newdata = testdata_x[, "AST", drop = FALSE],
             type = "response")
```

```
head(round(preds,2), 10)
```

```
##    1    2    3    4    5    6    7    8    9   10
## 0.42 0.37 0.32 0.66 0.32 0.37 0.52 0.32 0.52 0.39
```

Using R for predictions and classifciation based on logistic regression model

```
preds <- predict(m1,
            newdata = testdata_x[, "AST", drop = FALSE],
            type = "response")
```

```
head(round(preds,2), 10)
```

```
##    1    2    3    4    5    6    7    8    9   10
## 0.42 0.37 0.32 0.66 0.32 0.37 0.52 0.32 0.52 0.39
```

```
labels <- rep(0, nrow(testdata_x))
labels[preds > 0.5] <- 1
```

Using R for predictions and classifciation based on logistic regression model

```
preds <- predict(m1,
             newdata = testdata_x[, "AST", drop = FALSE],
             type = "response")
```

```
head(round(preds,2), 10)
```

```
##    1    2    3    4    5    6    7    8    9   10
## 0.42 0.37 0.32 0.66 0.32 0.37 0.52 0.32 0.52 0.39
```

```
labels <- rep(0, nrow(testdata_x))
labels[preds > 0.5] <- 1
```

```
head(labels, 10)
```

```
##  [1] 0 0 0 1 0 0 1 0 1 0
```

Evaluating the performance of the logistic regression machine

```
#Confusion matrix
table(labels, testdata_DEATH2YRS)
```

```
##       testdata_DEATH2YRS
## labels   0   1
##      0 191  84
##      1   5  12
```

```
#Accuracy
mean(labels == testdata_DEATH2YRS)
```

```
## [1] 0.6952055
```

Time to see if logistic regression is the answer

Go to the course website and find exercise session 2:

Exercise session 2

Machine learning & neural networks

Anne Helby Petersen

May 9, 2019

Overview

The goal of this exercise session is to:

- Use logistic regression to train a machine
- Experiment more with different performance measures (accuracy, AUC, AUPRC)

2.1: A simple logistic regression machine: Pablo

Below, we define a machine that uses logistic regression in its training step, and we name him Pablo.

Pablo's wants to use a logistic regression model to predict labels for new observations and he will use the `ECOG_1` and `ECOG_2` variables as well as the `HB` and `AST` variables. That means that he will fit the following model on the training data:

$$\log\left(\frac{P(\text{DEATH2YRS} = 1)}{1 - P(\text{DEATH2YRS} = 1)}\right) = \alpha + \beta_1 \cdot \text{ECOG\_1} + \beta_2 \text{ECOG\_2} + \beta_3 \text{HB} + \beta_4 \text{AST}$$