Statistical learning and Visualization:

Supervised learning - classification (1/2)

Erik-Jan van Kesteren

Department of Methodology and Statistics

Universiteit Utrecht

Applied Data Science

About me

- Assistant professor of data science @ UU M&S
- Team lead for the Social Data Science Team (ODISSEI national consortium)
- Background in statistics
- I will teach two classification weeks in this course
- I will coordinate the `INFOMDA2` course!

Topics this week

- Classification
- KNN
- Logistic regression
- Linear discriminant analysis
- Generative vs discriminative
- Trees
- Confusion matrix

Classification

The thing you're trying to predict is *discrete*:

- *Titanic*: Survival/Nonsurvival
- Banking data: Default on/payment of debt
- GPS/Accelerometer data: Work/Home/Friend/Parking/Other
- Imagenet: gazelle/tank/pirate/sea lion/tandem bicycle/. . .
- Etc.

KNN

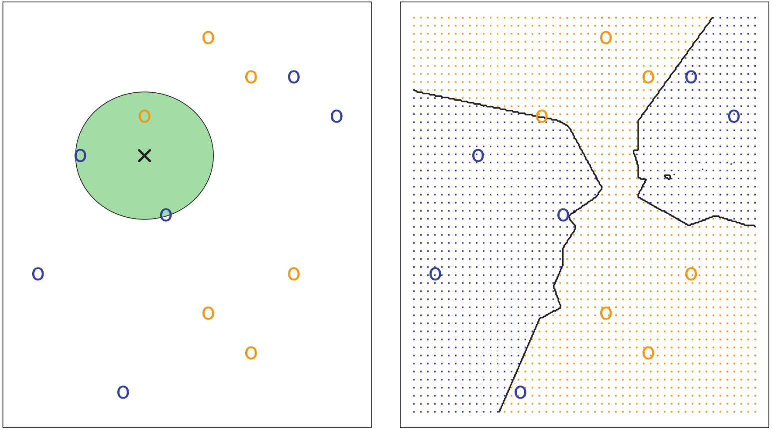


FIGURE 2.14. *The KNN approach, using* $K = 3$*, is illustrated in a s*

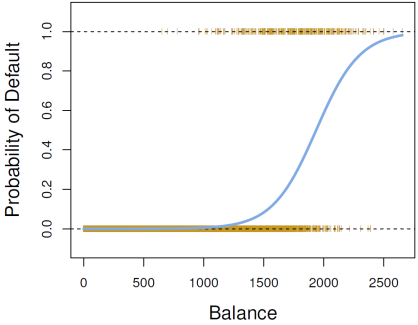
Discriminative classifier

Directly model $p(Y = k|X)$ as a function of X.

$$p(Y = k|X) = f(X)$$

Logistic regression

$$p(Y = 1|X) = logit^{-1}(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$



$$\beta_0 = -10.65, \beta_1 = 0.0055$$

Logistic regression

Turning this function around:

$$log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X$$

Get comfortable with odds, log-odds, the logit, and the inverse logit!

Logistic regression

$$log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right)=\beta_0+\beta_1 X$$

If $\beta_0 = 0; \beta_1 = 2$: Interpretation for log-odds?
When X increases by 1, the log-odds of $Y = 1$ increase by 2.

Logistic regression

$$\frac{p(Y=1|X)}{1-p(Y=1|X)} = e^{\beta_0+\beta_1 X}$$

If $\beta_0 = 0; \beta_1 = 2$: Interpretation in odds?
When X increases by 1, the odds of $Y = 1$ multiply by $e^2 = 7.39$

Logistic regression

$$p(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

If $\beta_0 = 0; \beta_1 = 2$: Interpretation in probabilities?

- When X increases from 0 to 1, $Pr(Y = 1)$ increases from $logit^{-1}(0 + 2 \cdot 0) = 0.5$ to $logit^{-1}(0 + 2 \cdot 1) \approx 0.88$
- When X increases from 1 to 2, $Pr(Y = 1)$ increases from $logit^{-1}(0 + 2 \cdot 1) \approx 0.88$ to $logit^{-1}(0 + 2 \cdot 2) \approx 0.98$

Tip: use predicted probabilities (`predict(model, type = "response")` function in R)

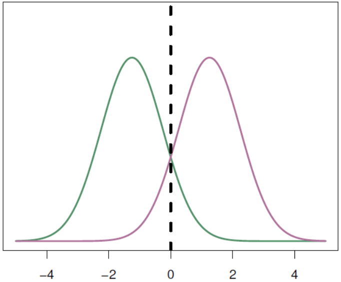
Generative classifier

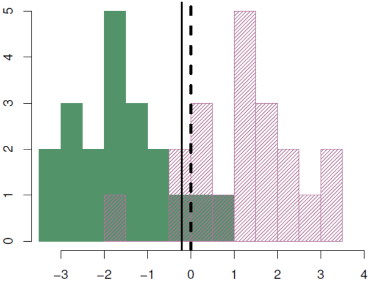
Use Bayes' rule to get to $p(Y = k|X)$.

$$p(Y = k|X) = \frac{\pi_k \cdot p(X|Y = k)}{\sum_{k=1}^{K} \pi_k \cdot p(X|Y = k)}$$

Linear discriminant analysis

- π_k is the proportion of observations in class k
- $p(X|Y = x)$ is a normal distribution with mean μ_k and common variance σ^2





Linear discriminant analysis

Advantages over logistic regression:

- Easy to extend to K > 2 classes
- Really easy to estimate (analytic solution for μ_k and σ^2). You can program it yourself!
- You can generate new *X* from the model (generative model).

Disadvantages:

- Assumption that *X* is normally distributed within each class *k* (categorical predictors???)
- Assumption that the variance of each normal distribution is the same!

Linear discriminant analysis

Discriminative classifiers

- Directly model $p(Y = k|X)$, for example using the logit link function.

Generative classifiers

- Estimate $p(X|Y = k)$ and π_k
- Use Bayes' rule to turn this into $p(Y = k|X)$:

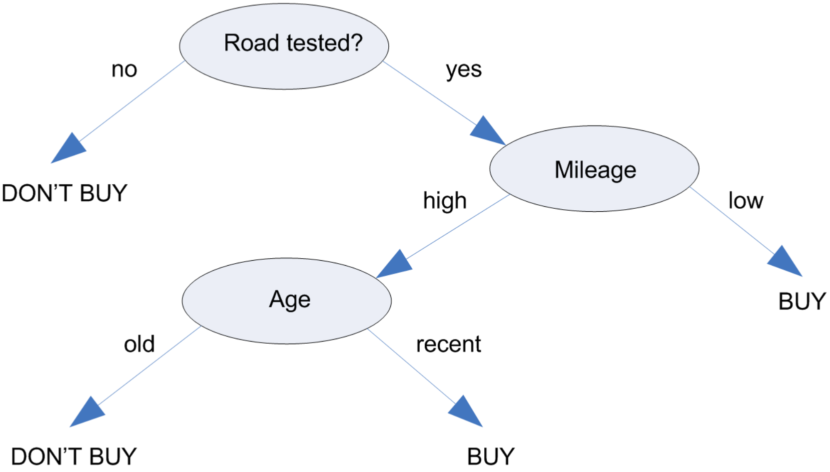
$$p(Y = k|X) = \frac{\pi_k \cdot p(X|Y = k)}{\sum_{k=1}^{K} \pi_k \cdot p(X|Y = k)}$$

Break

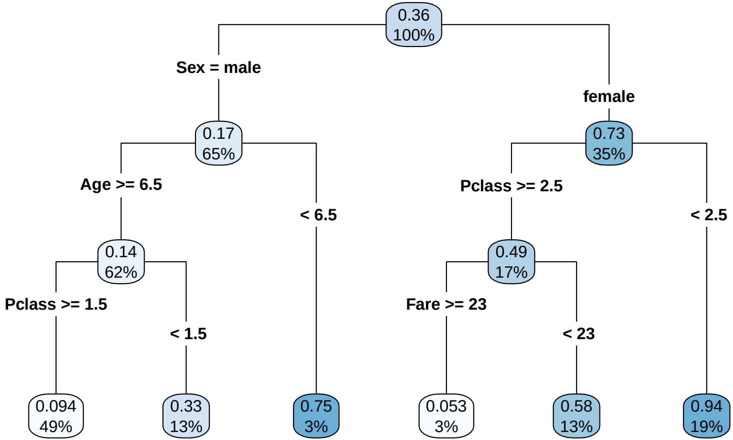
Trees!

Using decision trees for prediction

Decision tree: should I buy a car?



Prediction tree: wood you survive the *Titanic*?



Growing decision trees from data

Recursive partitioning

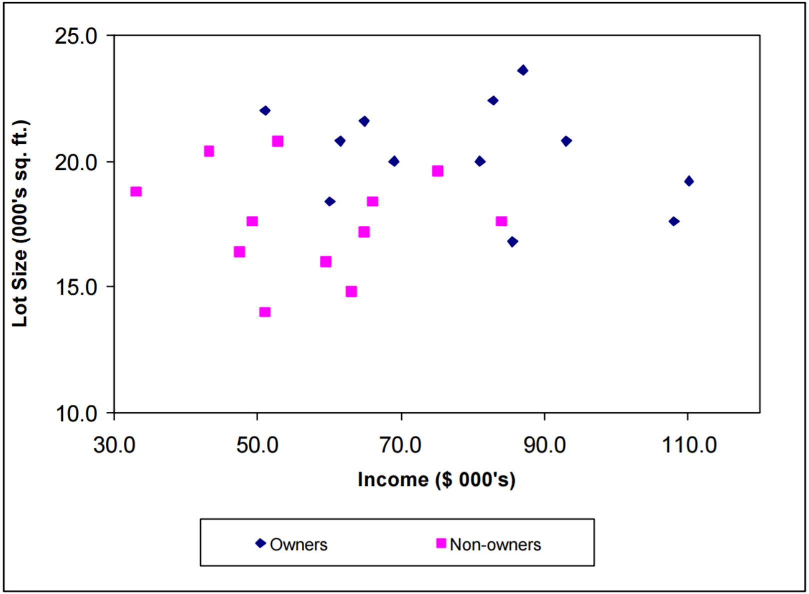
1. Find the split that makes observations as similar as possible on the outcome within that split;
2. Within each resulting group, do (1).

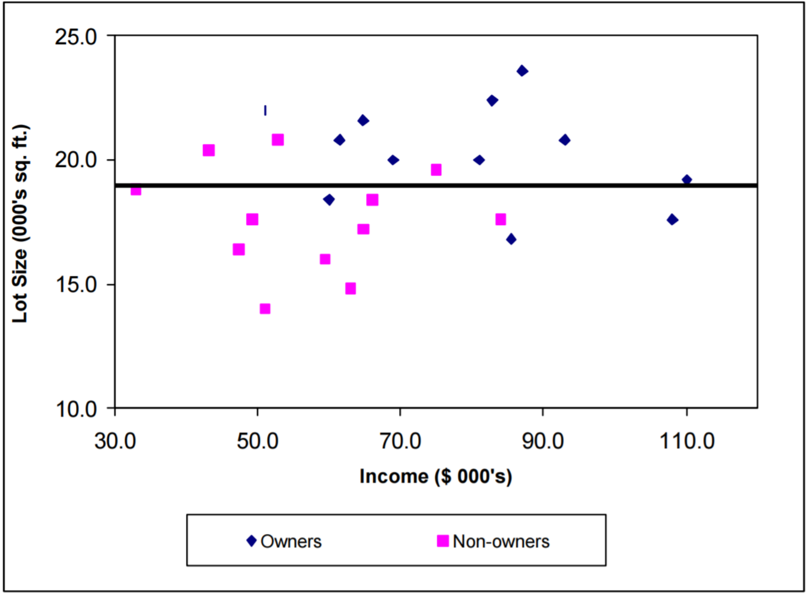
Recursive partitioning

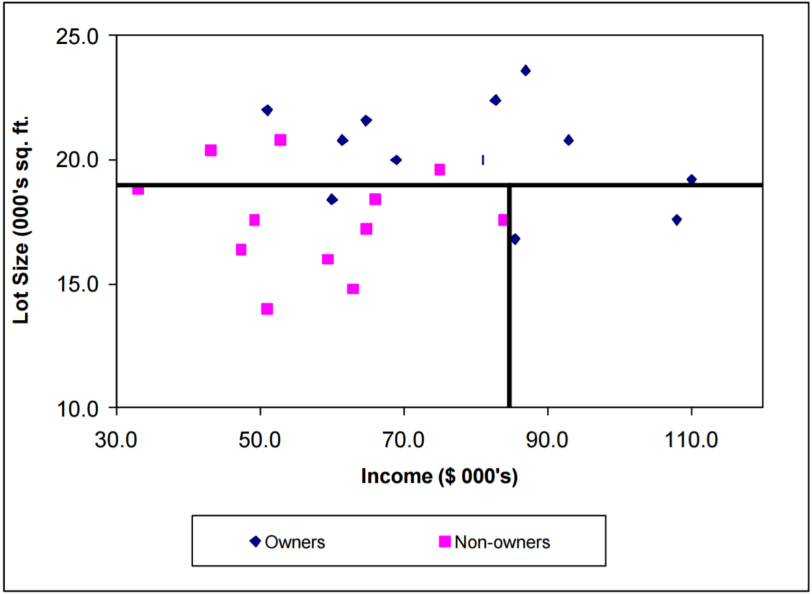
1. Find the split that makes observations as similar as possible on the outcome within that split;
2. Within each resulting group, do (1).

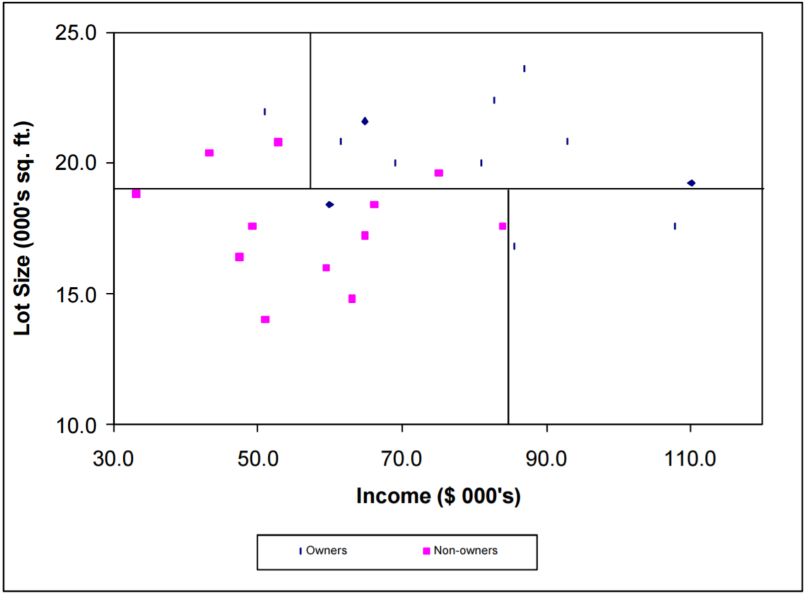
- Criteria for "as similar as possible": Purity, Reduction in MSE, ...
- Early stopping: add after (2):
  - "unless there are fewer than n_{min} observations in the group" (typically 10);
  - "unless the total complexity of the model becomes more than *cp*" (typically 0.05);

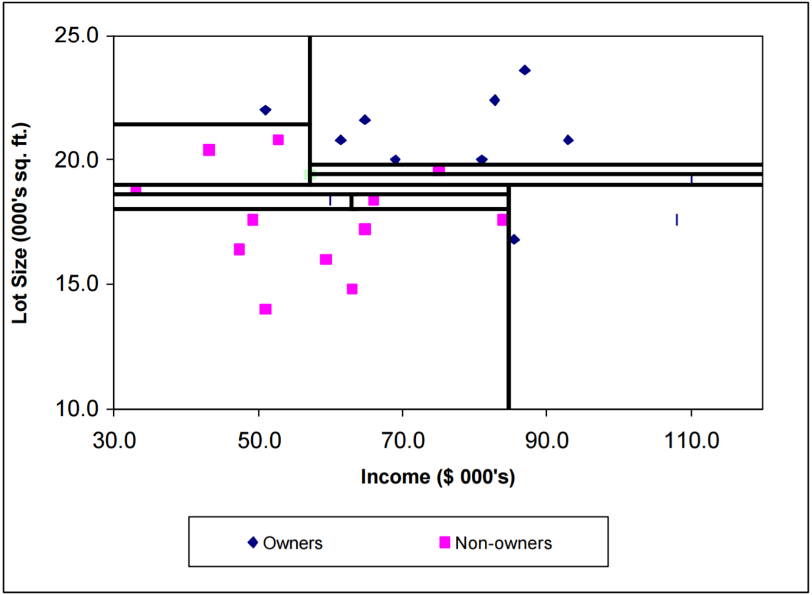
Simple example

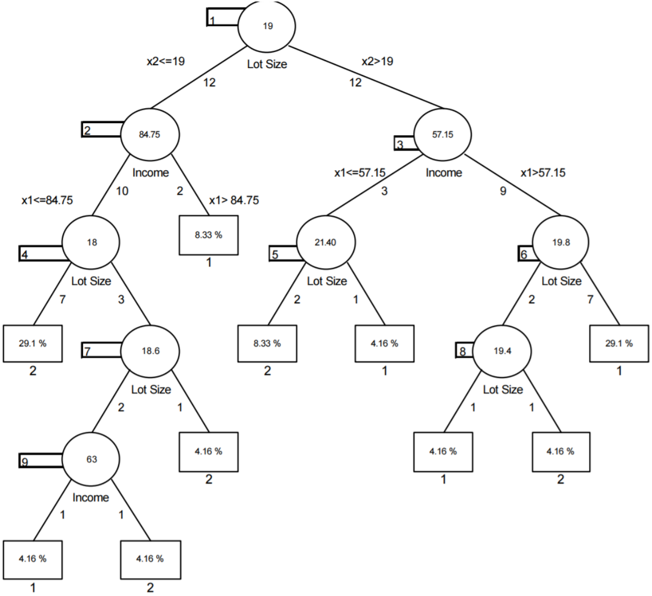

25.0
20.0
15.0
10.0
Lot Size (000's sq. ft.)
30.0
50.0
70.0
90.0
110.0
Income ($ 000's)
Owners
Non-owners


25.0
20.0
15.0
10.0
Lot Size (000's sq. ft.)
30.0
50.0
70.0
90.0
110.0
Income ($ 000's)
Owners
Non-owners


25.0
20.0
15.0
10.0
Lot Size (000's sq. ft.
30.0
50.0
70.0
90.0
110.0
Income ($ 000's)
Owners
Non-owners

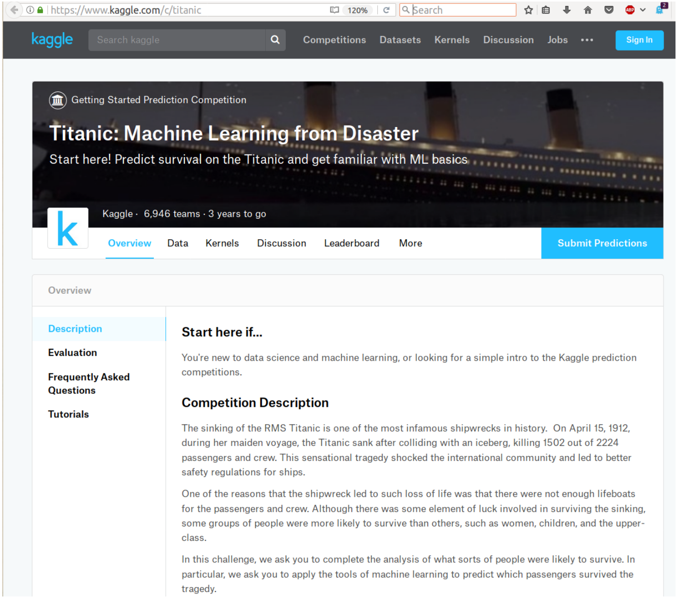

25.0
20.0
15.0
10.0
Lot Size (000's sq. ft.)
30.0
50.0
70.0
90.0
110.0
Income ($ 000's)
Owners
Non-owners


25.0
20.0
15.0
10.0
Lot Size (000's sq. ft.)
30.0
50.0
70.0
90.0
110.0
Income ($ 000's)
Owners
Non-owners



1
19
Lot Size
x2<=19
x2>19
12
12
2
84.75
Income
3
57.15
Income
x1<=84.75
10
2
x1> 84.75
x1<=57.15
3
9
x1>57.15
4
18
Lot Size
8.33 %
1
5
21.40
Lot Size
6
19.8
Lot Size
7
3
2
1
2
7
29.1 %
2
7
18.6
Lot Size
8.33 %
2
4.16 %
1
8
19.4
Lot Size
29.1 %
1
2
1
1
1
9
63
Income
4.16 %
2
4.16 %
1
4.16 %
2
1
1
4.16 %
1
4.16 %
2

More interesting example


https://www.kaggle.com/c/titanic
kaggle
Search kaggle
Competitions Datasets Kernels Discussion Jobs
Sign In
Getting Started Prediction Competition
Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Kaggle · 6,946 teams · 3 years to go
Overview Data Kernels Discussion Leaderboard More
Submit Predictions
Overview
Description
Evaluation
Frequently Asked Questions
Tutorials
Start here if...
You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.
Competition Description
The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.
In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Getting the Titanic data from Kaggle

```
# Import the Titanic data from Kaggle
train_url <-
"http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv"
titanic_df_kaggle <- read.csv(train_url)

# Make sure the results are reproducible
set.seed(1027)

# Randomize the rows
nobs <- nrow(titanic_df_kaggle) # Number of rows
idx_df <- 1:nobs # Indices of rows
titanic_df_kaggle <- titanic_df_kaggle[sample(idx_df), ] # Randomize

# Split the data into 70% train and 30% validation data
train_idx <- seq(1, nobs * 0.7) # Training data indices
val_idx <- seq((max(train_idx) + 1), nobs) # Validation data indices

train_df <- titanic_df_kaggle[train_idx, ] # Training data
val_df <- titanic_df_kaggle[val_idx, ] # Validation data
```

```
> head(train_df)
    PassengerId Survived Pclass                                          Name    Sex  Age SibSp Parch     Ticket    Fare Cabin Embarked
133         133        0      3 Robins, Mrs. Alexander A (Grace Charity Laury) female 47.0     1     0 A/5. 3337 14.5000                S
184         184        1      2                        Becker, Master. Richard F   male  1.0     2     1     230136 39.0000    F4        S
687         687        0      3                         Panula, Mr. Jaako Arnold   male 14.0     4     1    3101295 39.6875                S
178         178        0      1                       Isham, Miss. Ann Elizabeth female 50.0     0     0   PC 17595 28.7125   C49        C
880         880        1      1  Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) female 56.0     0     1      11767 83.1583   C50        C
204         204        0      3                             Youseff, Mr. Gerious   male 45.5     0     0       2628  7.2250                C
```

Fitting a classification tree in R

```
library(rpart)

titanic_tree <-
  rpart(
    Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
    data = train_df,
    control = list(cp = 0.02)
  )
```

Evaluating classifiers

THE INTERNATIONAL JOURNAL OF ROBOTICS RESEARCH / January 2007

Table 5. Place Confusion Matrix

Truth	Inferred labels					FN
	Work	Home	Friend	Parking	Other	
Work	5	0	0	0	0	0
Home	0	4	0	0	0	0
Friend	0	0	3	0	2	0
Parking	0	0	0	8	0	2
Other	0	0	0	0	28	1
FP	0	0	1	1	2	-

More on this next week.
Wednesday: Q&A session for practical.

Have a nice day!