



MEDNARODNA  
PODIPLOMSKA ŠOLA  
JOŽEFA STEFANA

INFORMATION AND COMMUNICATION TECHNOLOGIES  
PhD study programme

# Data Mining and Knowledge Discovery Homework assignment 2

Petra Kralj Novak  
December 11, 2019

[http://kt.ijs.si/petra\\_kralj/dmkd3.html](http://kt.ijs.si/petra_kralj/dmkd3.html)

# Assignment 1: Home reading

Read: Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning, Second edition*. New York: Springer series in statistics. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>

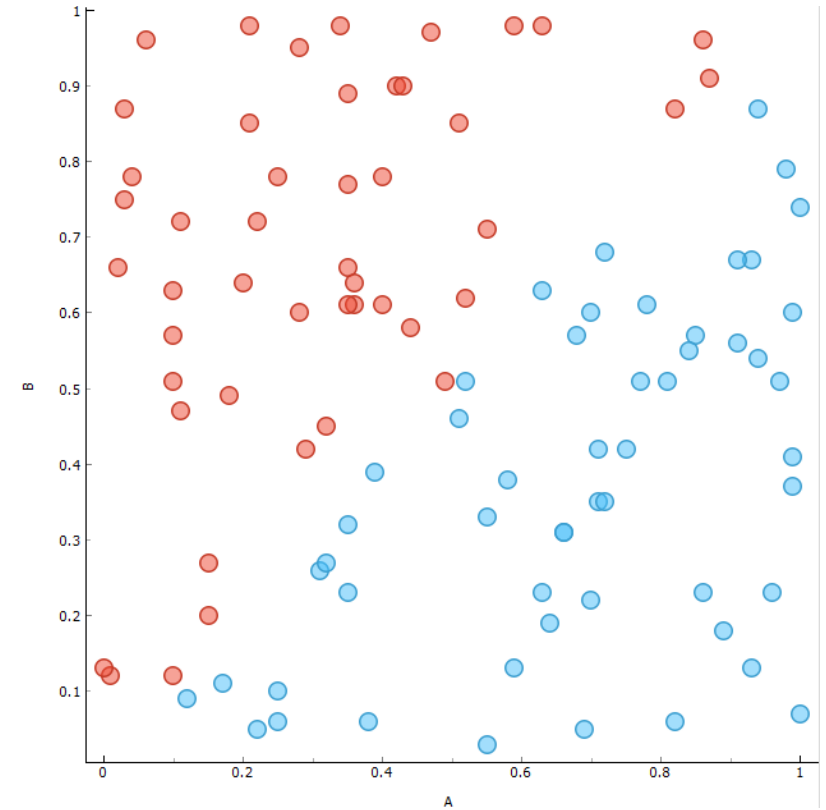
Pages 9 – 18:

2.1 Introduction . . . . .	9
2.2 Variable Types and Terminology . . . . .	9
2.3 Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors . . . . .	11
2.3.1 Linear Models and Least Squares . . . . .	11
2.3.2 Nearest-Neighbor Methods . . . . .	14
2.3.3 From Least Squares to Nearest Neighbors . . . . .	16

# Assignment 2: Decision boundary

- What is a decision boundary like for KNN?
  - K=1
  - K=3
  - K=10

You can do it by hand, in Orange or in SciKit.



# Assignment 3: Confusion matrix

		Predicted		$\Sigma$
		no	yes	
Actual	no	1364	126	1490
	yes	362	349	711
$\Sigma$		1726	475	2201

		Predicted				$\Sigma$
		unacc	acc	good	v-good	
Actual	unacc	1154	54	2	0	1210
	acc	94	276	14	0	384
	good	0	44	22	3	69
	v-good	0	25	0	40	65
$\Sigma$		1248	399	38	43	1728

	Titanic	Car
Number of examples		
Number of classes		
Number of examples in each class		
Number of examples classified in individual classes		
Number of misclassified examples		
Classification accuracy		

# Assignment 4: F1

- Express F1 in terms of TP, FP, TN, FN

$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

		Predicted class		Total instances
		+	-	
Actual class	+	TP	FN	P
	-	FP	TN	N