Klasifikacija – Nearest Neighbor

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

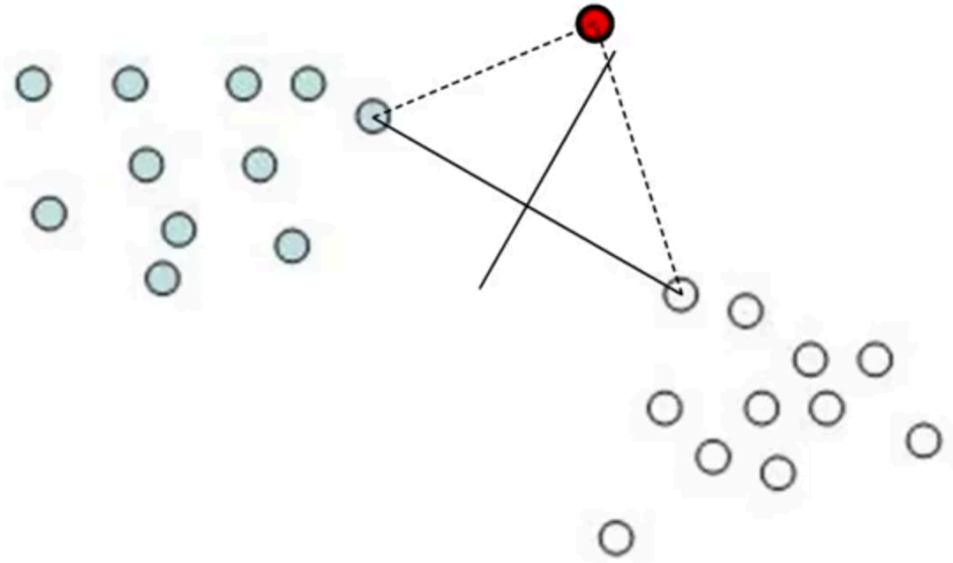
URL: http://nikola.milikic.info

Najbliži sused (eng. Nearest Neighbor)

- Nearest Neighbor pretražuje trening set i traži najsličniju instancu
  - instance u trening setu predstavljaju "znanje"
  - lenjo učenje ("lazy learning") – ništa ne radi do momenta kada treba da da predikcije

- Jedan od najjednostavnijih algoritama mašinskog učenja
- Instance-based learning = nearest neighbor learning

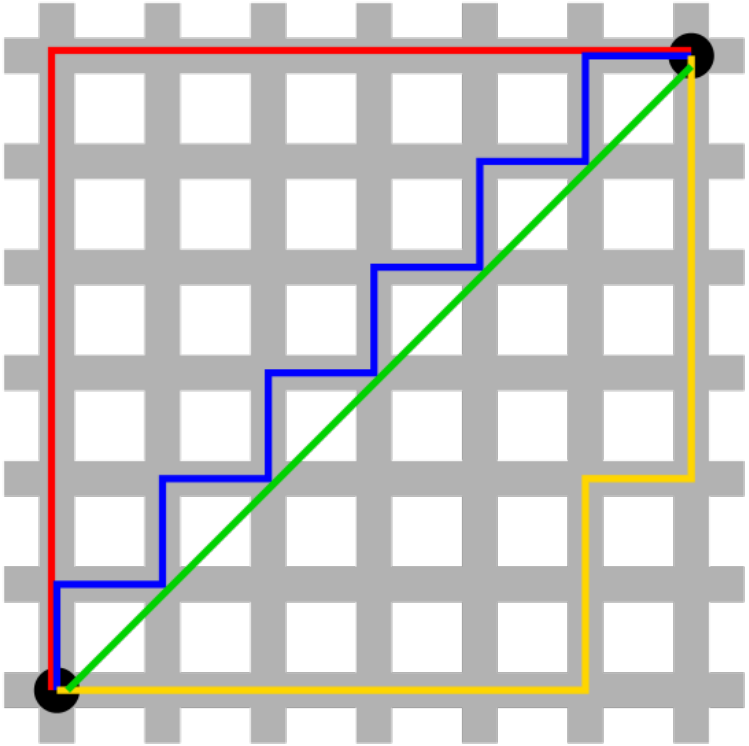
Primer klasifikacije



- Nepoznatu instancu klasifikujemo na osnovu klase najbliže instance

Mera sličnosti

- Euklidska udaljenost (suma razlike kvadrata)
- Menhetn udaljenost (suma apsolutnih razlika)
- Normalizacija atributa ako su na skalama koje se dosta razlikuju
- Nominalni atributi? Obično se uzima ako su različite vrednosti, udaljenost = 1; ako su iste vrednosti, udaljenost = 0



https://en.wikipedia.org/wiki/File:Manhattan_distance.svg

Broj suseda

- k-nearest neighbor – od k najbližih suseda, odabrati klasu koju ima većina suseda
- Ako je u pitanju binarna klasifikacija, za k se najčešće bira neparan broj
- Ako su podaci *noisy*, uzeti u obzir više suseda
- Ako su podaci *skewed*, onda instance klase koja se češće pojavljuje mogu dominirati
- Ako je k malo, postoji veća sklonost ka overfitting-u

Dodeljivanje težinskih faktora

- Kako bi se uzela u obzir udaljenost suseda od nepoznate instance, često se dodeljuju težinski faktori
- Najčešće se svakom susedu dodeljuje težinski faktor $1/d$, gde je d udaljenost suseda od nepoznate instance

Kada koristiti KNN?

- Manje od 20 atributa
- Dosta podataka za trening

Prednosti:

- Treniranje je brzo
- Može naučiti kompleksne funkcije
- Nema gubitaka informacija

Nedostaci:

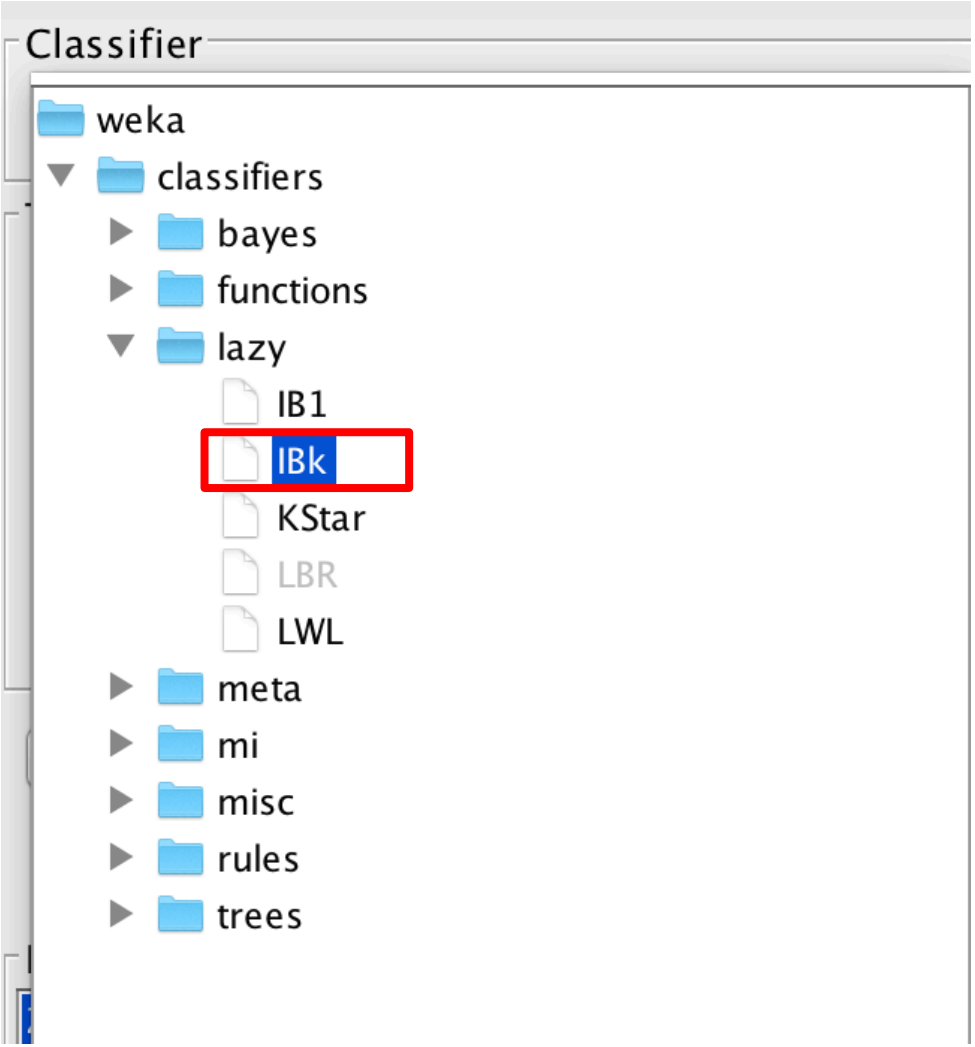
- Spor odziv
- Nerelevantni atributi unose veliku grešku

Primer 1 – Skup podataka “Dijabetes”

diabetes.arff

- Skup podataka “Pima Indians Diabetes Database” sadrži podatke o ženama koje pripadaju grupi Pima Indijanaca koje imaju najmanje 21 godinu i testirane su na dijabetes. Dataset je kreirao Johns Hopkins University, Merilend, SAD.
- Postoji ukupno 768 instanci sa 8 atributa numeričke vrednosti koje opisuju karakteristike pacijenata i imaju podatak da li su osobe bile pozitivne ili negativne na dijabetes.
- Naš cilj je da predvidimo da li je neispitani pacijent pozitivan na dijabetes ili ne.

KNN u Weka-i


Classifier
weka
classifiers
bayes
functions
lazy
IB1
IBk
KStar
LBR
LWL
meta
mi
misc
rules
trees

Kako se računa weighted average?

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.834     0.459       0.772    0.834     0.802      0.773    tested_negative
                   0.541     0.166       0.636    0.541     0.585      0.773    tested_positive
Weighted Avg.      0.732     0.357       0.725    0.732     0.726      0.773

=== Confusion Matrix ===

   a   b   <-- classified as
 417  83 |   a = tested_negative
 123 145 |   b = tested_positive
```

$$\frac{0.802 \cdot (417 + 83) + 0.585 \cdot (123 + 145)}{417 + 83 + 123 + 145} = 0.726$$

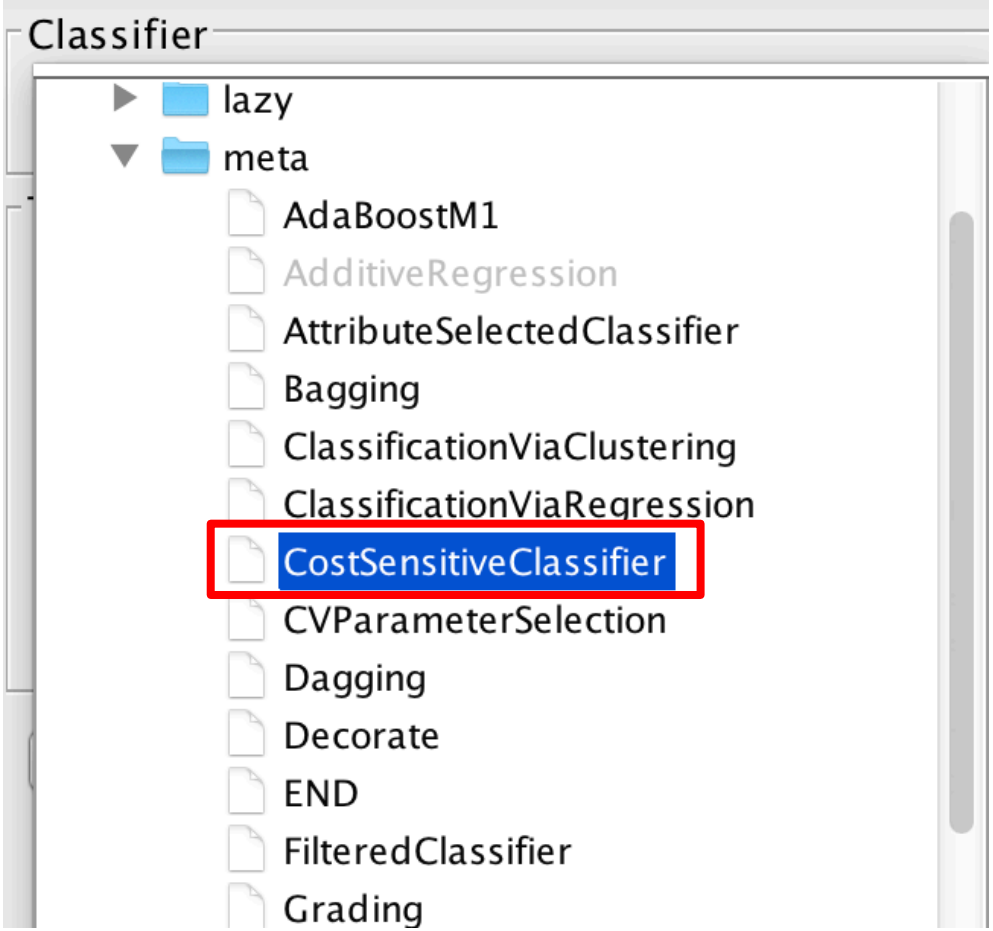
Cost Sensitive klasifikacija

- Primenjuje se kada imamo izrazito nebalansiran dataset (eng. skewed dataset) po klasi
  - npr. u datasetu sa 10000 instanci sa dve moguće klase, imamo 100 instanci sa prvom klasom, a preostalih 9990 sa drugom klasom
- Ovo može uticati na precision, recall i f-measure
- Cost Sensitive klasifikacija kažnjava FP (false positive) ili FN (false negative)

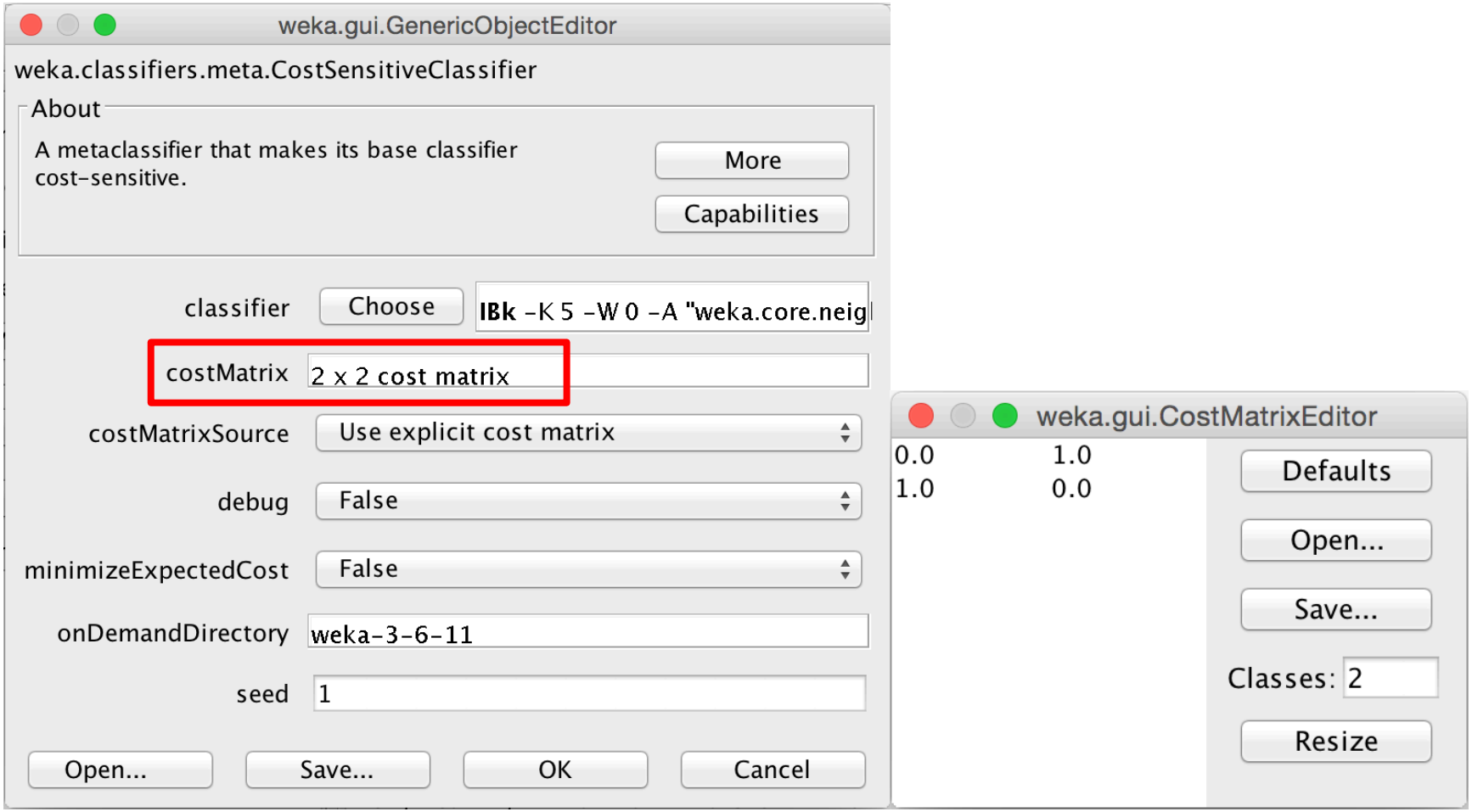
Confusion Matrix

		Algoritam predvideo	
		Klasa 1	Klasa 2
Stvarna klasa	Klasa 1	**TP**	**FN (2)**
	Klasa 2	**FP (1)**	**TN**

Cost Sensitive klasifikator u Weka-i



Cost Sensitive klasifikator u Weka-i



Preporuke i zahvalnice

Weka Tutorials and Assignments @ The Technology Forge

- http://www.technologyforge.net/WekaTutorials/

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- https://www.youtube.com/user/WekaMOOC/

”Weka Tutorials”, Learn with Rashdi.

- https://www.youtube.com/channel/UCa8nqCmiWvaA8rnrRCySQsw

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I

Pitanja?

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

URL: http://nikola.milikic.info