Classification – Nearest Neighbor

UROŠ KRČADINAC

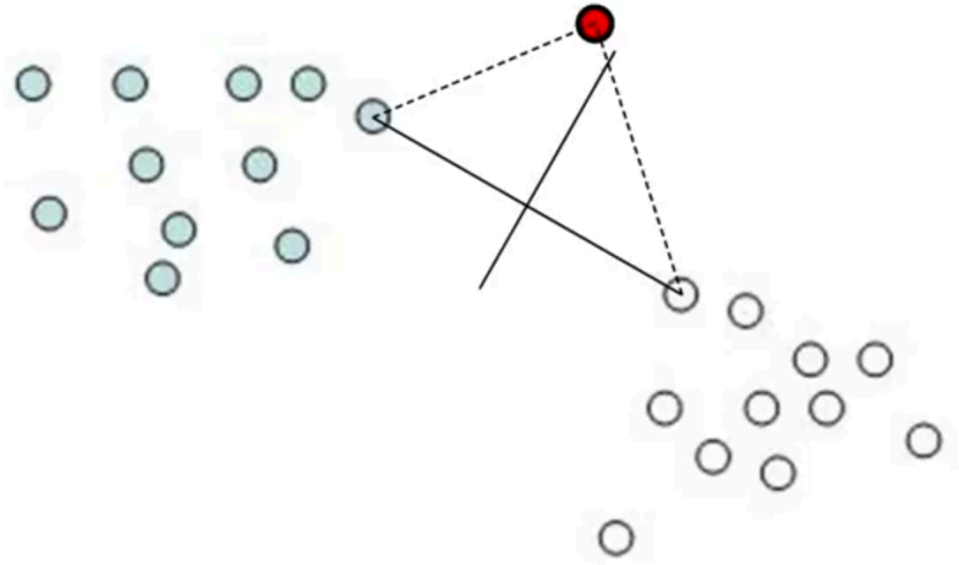
EMAIL: uros@krcadinac.com

URL: http://krcadinac.com

Nearest Neighbor

- Nearest Neighbor is searching the training set looking for the most similar instance
  - instances in training set are representing the “knowledge”
  - “lazy learning” – does nothing until the moment it nneds to make a prediction
- One of the most simplest machine learning algorithms
- Instance-based learning = nearest neighbor learning

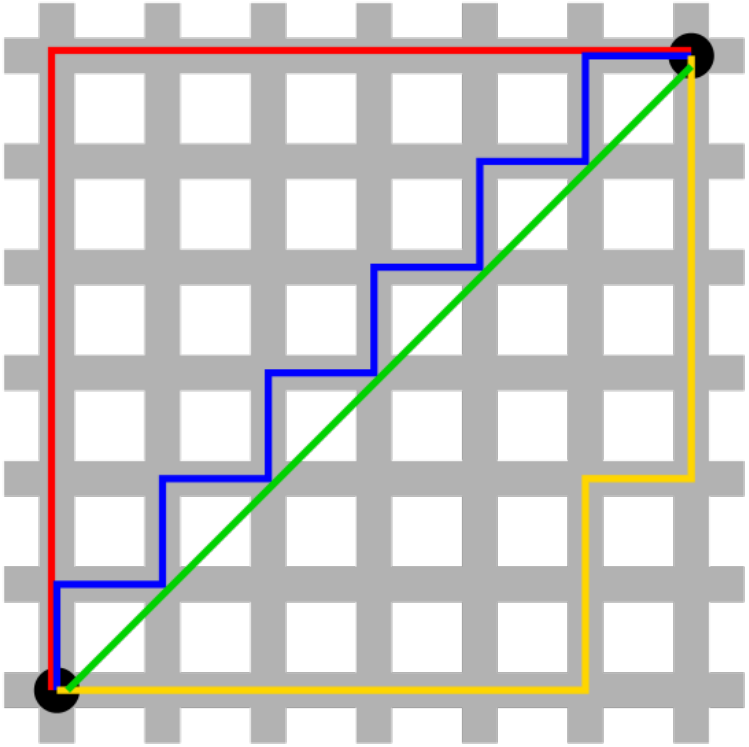
Classification Example



- Unknown instance is classified based on the nearest instance class

Similarity Measure

- Euclidian distance (sum of squared errors)
- Manhattan distance (sum of absolute errors)
- Attribute normalization if scales are different
- Nominal attributes? Usually if values are different, distance is 1. If values are the same, distance is 0.



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

Number of Neighbors

- k-nearest neighbor – from k nearest neighbors, choose the majority class
- K is usually odd number
- If data is *noisy*, take into account more neighbors
- If k is too small, there is a tendency for overfitting

Distance Weighting

- In order to take into account distance between an unknown instance and a neighbor, add weight to the distance
- Usually each neighbor distance is weighted with $1/d$, where d is a distance from a neighbor

When to use KNN?

- Less then 20 attributes
- Enough training data

Advantages:

- Training is fast
- Can solve complex functions
- There is no data loss

Disadvantages:

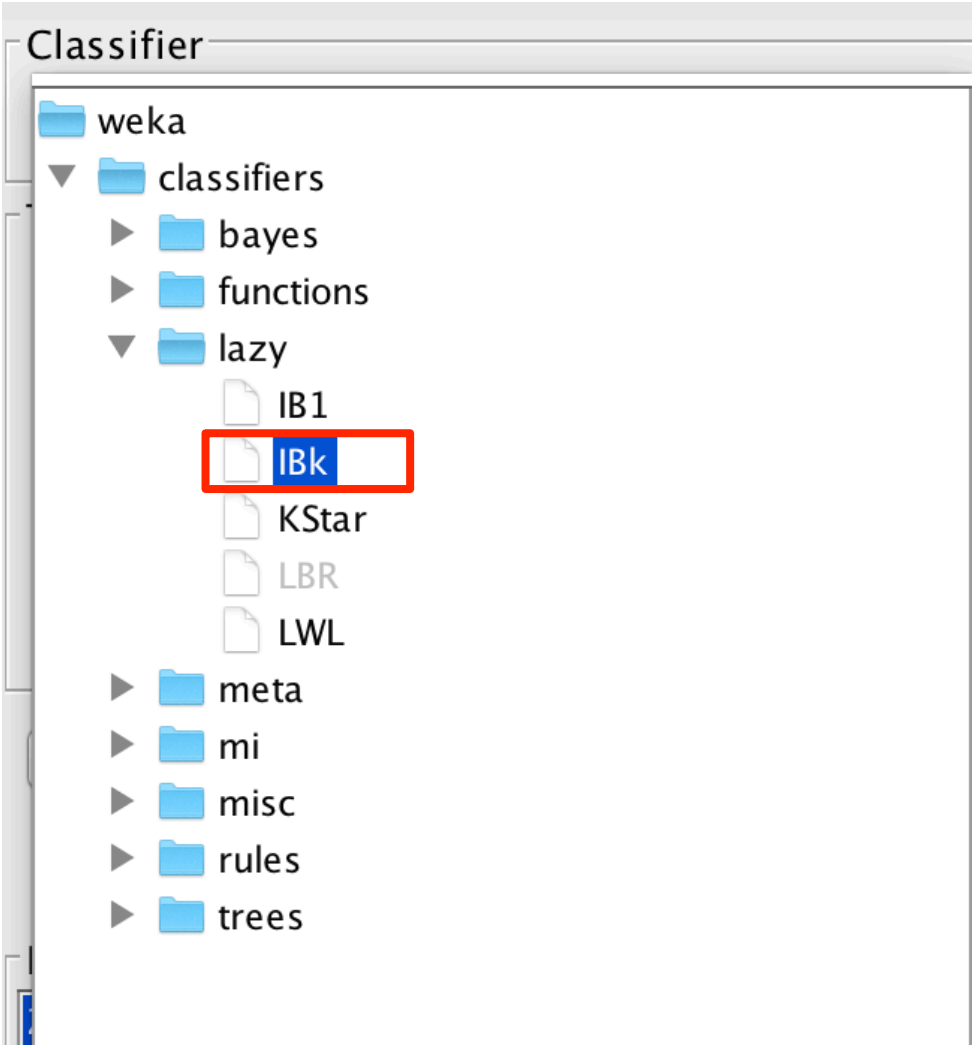
- Slow recall
- Irrelevant attributes introduce big error

Example 1 – "Diabetes" Dataset

diabetes.arff

- Dataset "Pima Indians Diabetes Database" contains data about female Pima Indians aged 21 years or higher and tested for diabetes. Dataset was donated by the Johns Hopkins University, Maryland, USA.
- There are total of 768 instances described by 8 numerical attributes about patient conditions and annotated with a class determining whether patients were positive or negative for diabetes.
- Our goal is to predict whether a new patient will be diagnosed positive or negative.

KNN in Weka


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

How to calculate weighted average?

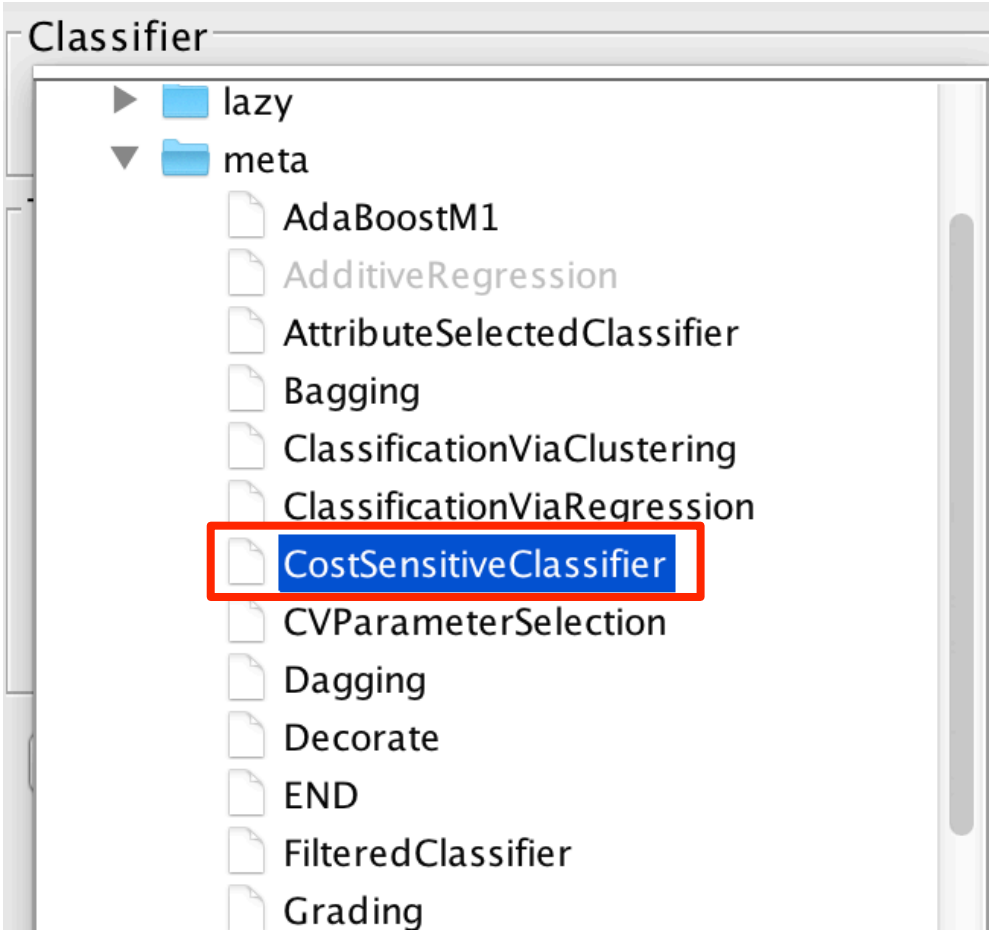
```
=== Detailed Accuracy By Class ===

                 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
```

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

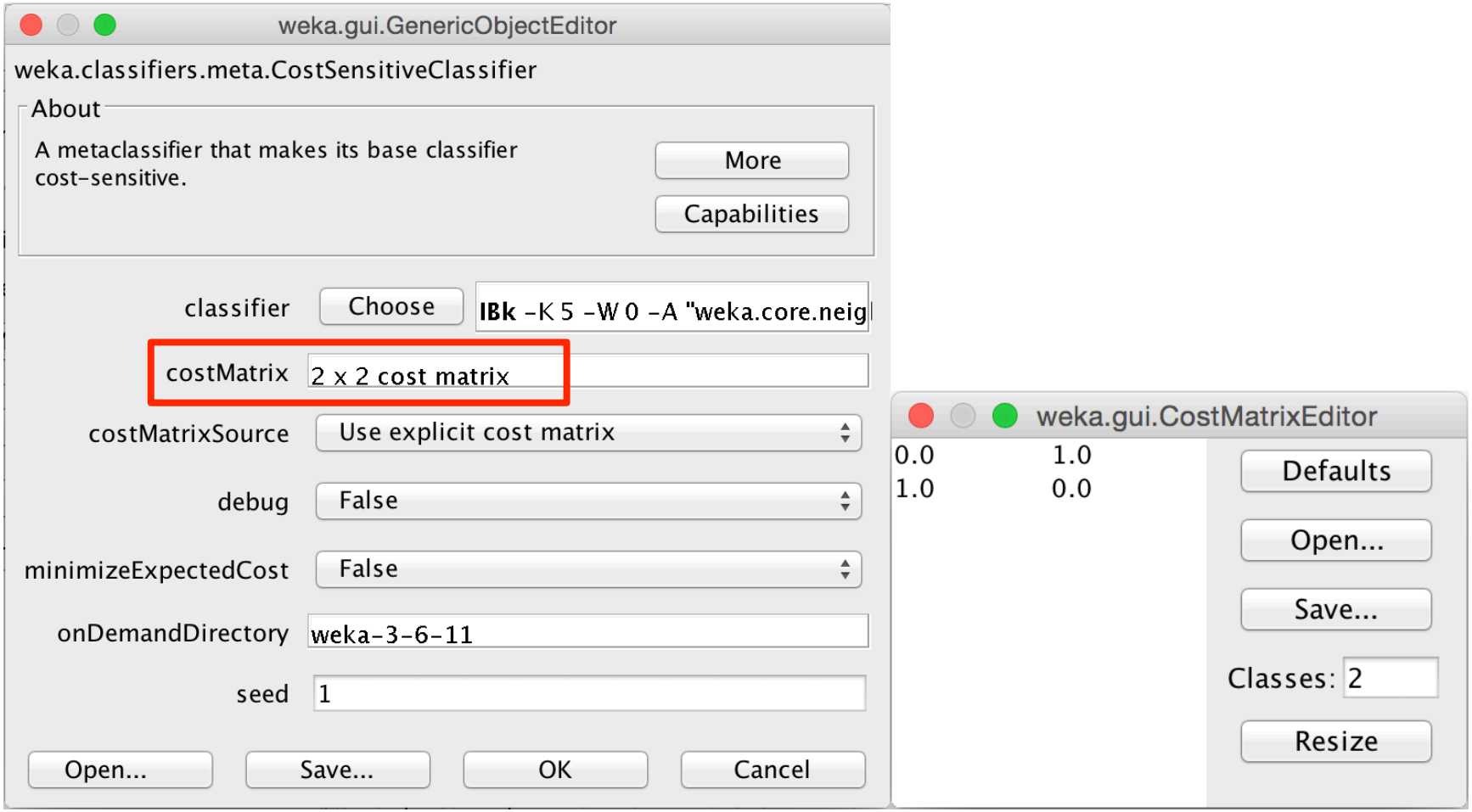
Cost Sensitive Classification

- Skewed dataset
  - e.g. in dataset with 10000 instances with two possible classes, there are 100 instances with first class, and other 9990 instances with second class
- This can influence precision, recall and f-measure
- Cost Sensitive classification punishes FP (false positive) or FN (false negative)

Cost Sensitive Classification in Weka



Cost Sensitive Classification in Weka



Recommendations and credits

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

(Anonymous) survey for your comments and suggestions:
http://goo.gl/cqdp3l

ANY QUESTIONS?

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: http://krcadinac.com