Rozpoznávanie obrazcov - 10. cvičenie
Decision trees

Viktor Kocur
viktor.kocur@fmph.uniba.sk

DAI FMFI UK

27.4.2020

Evaluation

Multiple classes

So far we mostly had binary classification tasks. Some classifiers (NB, kNN), which we tried can already do multiclass classification. For binary classifier it is necessary to use multiple of them to obtain a multiclass classifier.

fitceocc

Mdl = fitcecoc(X, y) - returns a multiclass SVM classifier

Accuracy

Is accuracy sufficient?

Accuracy is defined as the fraction of correctly classified examples and total examples. This metric can be deceptive. Imagine a situation where we have class imbalance and 90% of examples are from one class and 10% from the other. Then a classifier which blindly selects the first class will have an accuracy of 90%, but it is not a good classifier.

Confusion matrix

Confusion matrix

One of the ways to evaluate a classifier is to use the confusion matrix. Element on i-th row and j-th column is the amount of examples which are from the i-th class, but were classified as the j-th class.

confusionmat

C = confusionmat(g1,g2) - returns the confusion matrix for correct labels g1 and predicted labels g2.

confusionchart

cm = confusionchart(g1,g2) - plots the confusion matrix with colors

True/False Positive/Negative

We will use some terms for every class:

- True Positive - TP
  classifier predicted the class and it is correct
- False Positive - FP
  classifier predicted the class and it is incorrect
- True Negative - TN
  classifier did not predict this class and it is correct
- False Negative - FN
  classifier did not predict this class and it is incorrect

Precision a Recall

Precision

We define precision as $\frac{TP}{TP+FP}$. The difference between precision and accuracy is usually in a fact that accuracy is a mean of precisions across all classes. When talking about precision we usually consider it for one class.

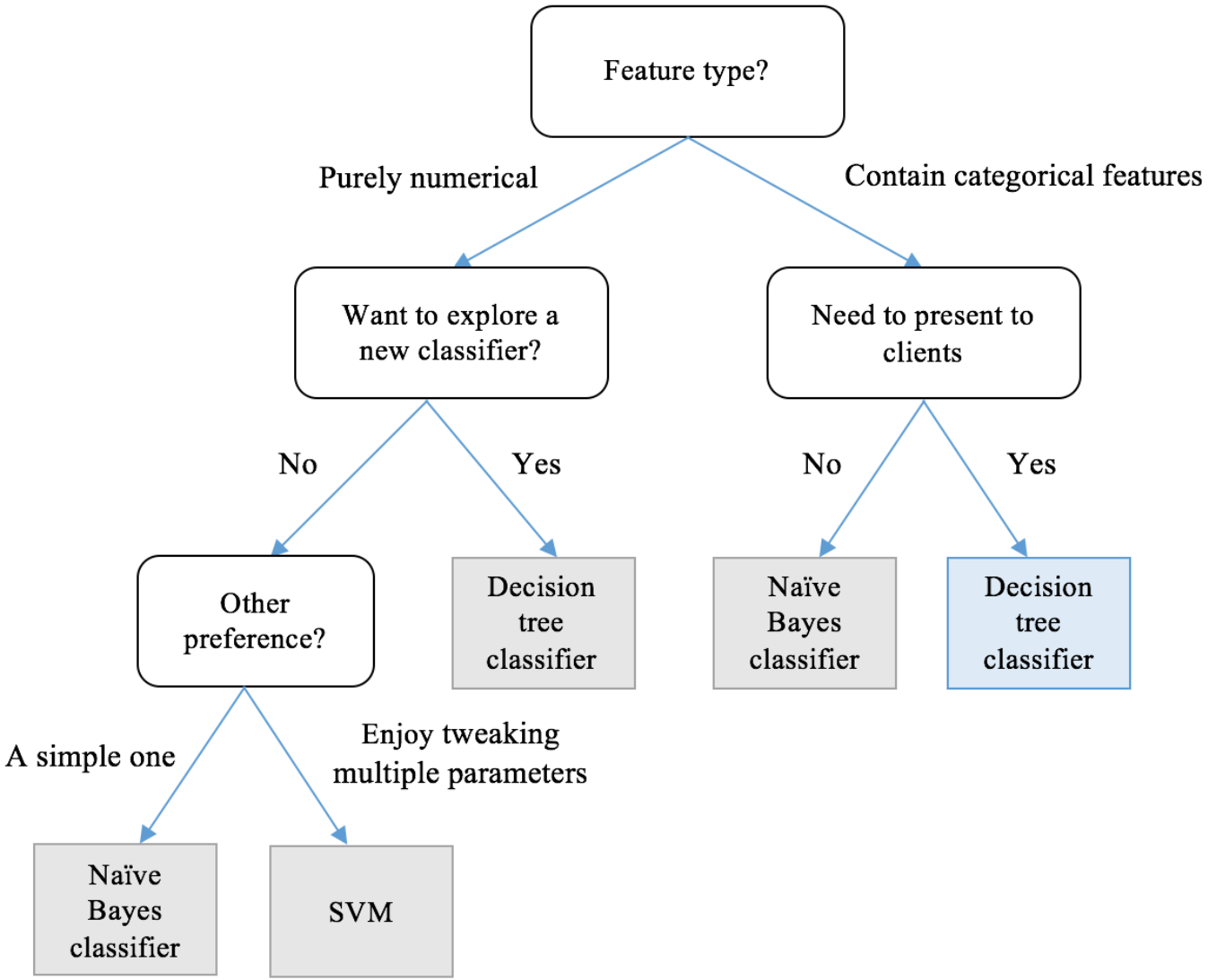
Recall

Recall is defined as $\frac{TP}{TP+FN}$, e.g. what portion of the examples in the class has the classifier correctly classified.

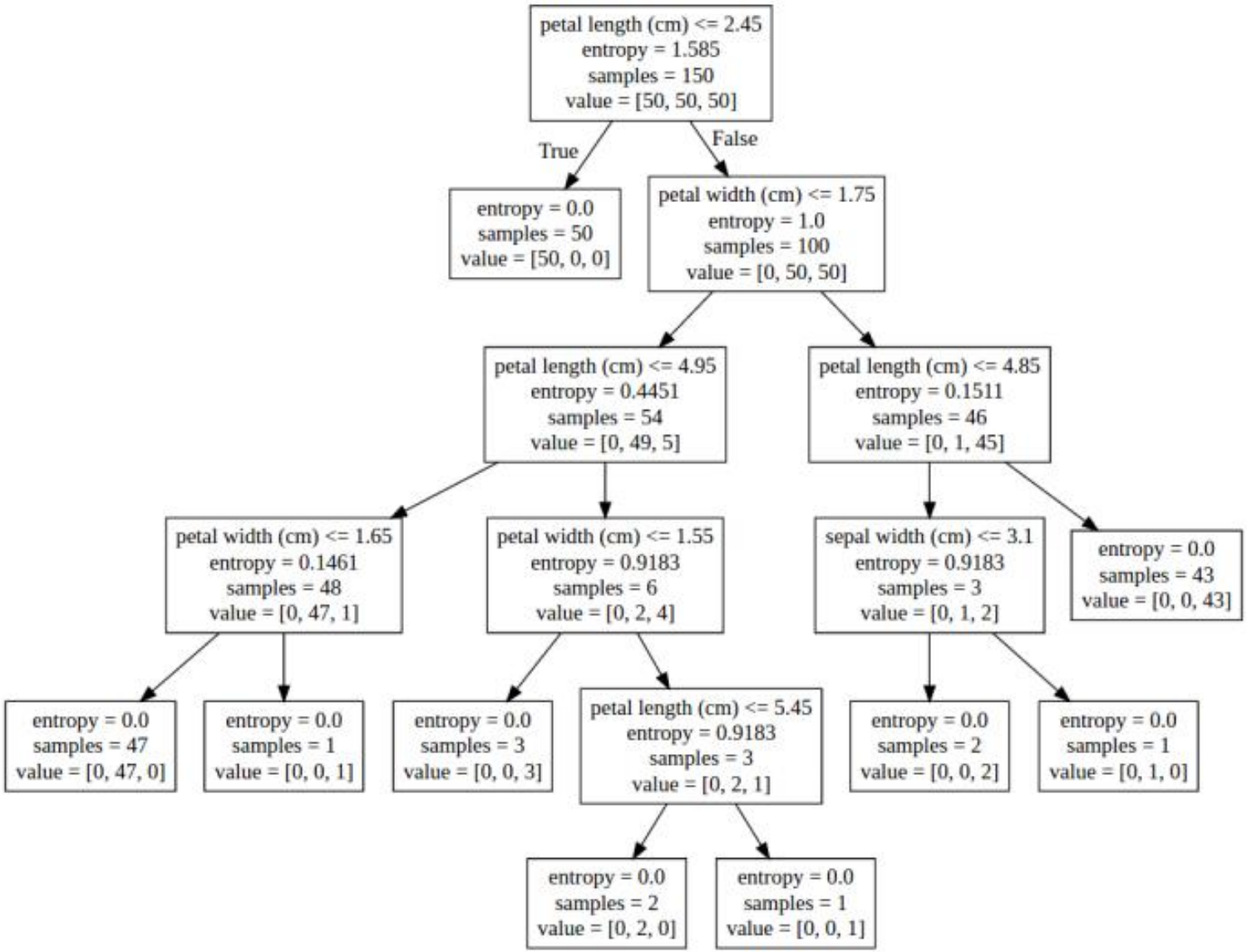
Exercise

Train a classifier on the fisheriris dataset and calculate the confusion matrix. Calculate the precision and recall as well.

Decision tree



Decision tree



Constructing trees

Splitting rules

The tree is constructed by selecting a feature and a value based on which we split the set of elements into two parts. This process is repeated with both subsets until some stopping criterion is fulfilled.

Stopping criterion

Examples: each subset contains only one class, the tree reach a certain depth, fewer misclassifications than a certain thresholds, next best feature for selection is worse than some threshold.

á

ID3

We choose a feature with lowest entropy, e.g. a feature for which the information gain is the highest (mutual information with classes is the highest).

C4.5

Similar to ID3, but this time we optimize for highest normalized information gain. C4.5 can also work with numerical data.

Splitting rules - 4th lab theory

Entropy

$$H(Y) = \sum_{y \in \omega} -P(Y = y) \cdot log_2(P(Y = y))$$

Specific conditional entropy

$$H(Y|X = v) = H(Y), \text{len pre hodnoty} Y, \text{kde } X = x$$

Splitting rules - 4th lab theory

Mutual information, information gain

$$I(Y;X) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \omega} P(X = x) \cdot H(Y|X = x)$$

Normalized information gain

$$nI(Y;X) = \frac{I(Y;X)}{H(X)}$$

Examples

ID3

https://sefiks.com/2017/11/20/
a-step-by-step-id3-decision-tree-example/

C4.5

https://sefiks.com/2018/05/13/
a-step-by-step-c4-5-decision-tree-example/

Matlab

fitctree

Mdl = fitctree(X,y) - returns a tree classifier.

fitctree

Mdl = fitctree(T,property) - returns a tree classifier for table T and classification target in the property column of the table.

CART

Matlab uses the CART algorithm which is similar to ID3, but slightly different. It is not a part of the lecture so we will not deal with it now.

Matlab

predict

Mdl.predict(x) - returns model prediction

view

Mdl.view('Mode','graph') - displays the tree

Exercise

Create and display a tree for the fisheriris and census1994 database.

Pruning the trees

Pruning

The tree can be too complex which leads to overfitting. It is possible to prune the tree so that its subtrees which only provide marginal benefits are converted to leafs.

prune

MdlP = prune(Mdl,'Property', value) - returns a pruned tree based on the selected property.

Exercise

Prune the tree for the data in fisheriris and census 1994. Test various properties. Check if pruning helps the accuracy on the test set of census1994.