Open your mind. LUT.
Lappeenranta University of Technology

LUT Machine Vision and Pattern Recognition 2015-11-16

BM40A0700 Pattern Recognition

Lasse Lensu

Exercise 10: Decicion trees and practical issues

1. Decision trees (1 point): Construct a decision tree from the provided training data by using the Matlab constructor `T = classregtree(X,Y)` with Gini impurity (Matlab: Gini's diversity index). You can visualize the tree using the method `view`. Evaluate the tree with the provided test data by using the method `eval`.

   Try different split criterion instead of Gini impurity: can you observe any difference in results?

   Hints: The function `classregtree` expects `X` and `Y` in such a form that the samples are on rows, not columns.

   Additional files: `irisdata.mat`.

2. Random decision forests (2 bonus points): Construct a random decision forest from the provided training data by using the Matlab constructor `T=classregtree(X,Y)` with Gini impurity (Matlab: Gini's diversity index). Train several trees with different randomly selected subsets from the training set. You can use all the variables at each node to make a decision at that node so you do not need to mofidy the `classregtree` function. Branching has to be done for the whole tree without pruning.

   How do the classification results change when the decision forest is formed from ten, one hundred or one thousand decision trees?

   Additional files: `irisdata.mat`.

3. Bias-variance tradeoff (1 point): Examine the dependence of variance of estimated classification error to the number of training and test samples. For the experimentation, use a classification method of your choice and the given data.

   (a) Estimate the classification error with the whole test data (75 samples) and subsets of different size of the training data (15, 30, 45, 60, 75 samples) using 100 random subsets of the training data for each training subset. Calculate the mean and standard deviation of the estimated classification error for each subset. How does the training data size affect the mean and standard deviation?

   (b) Repeat the above experiment using the whole training data (75 samples) and subsets of different size of the test data (15, 30, 45, 60, 75 samples), again using 100 random subsets for each size. Calculate the mean and standard deviation for each subset size. How does the test data size affect the mean and standard deviation?

   Additional files: `irisdata.mat`

PO BOX 20, FI-53851 LAPPEENRANTA, FINLAND
E-mail: lasse.lensu@lut.fi