Home assignment 2

Exercise 1. k - nn

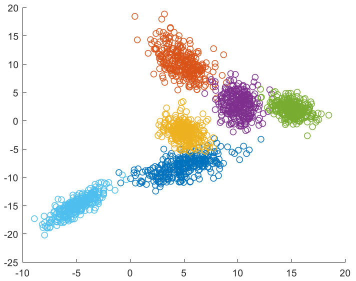
K-nn is a method for classification and regression.

1. Data is dividend into training and testing data.
2. Measuring distances from all points in testing data to all points in training data.
3. Getting K nearest neighbors for each point in testing data. Finding most occurring label for each point in testing data by getting most occurring class labels from k-nearest neighbors, most occurring class label will be the class label for the point in testing data.

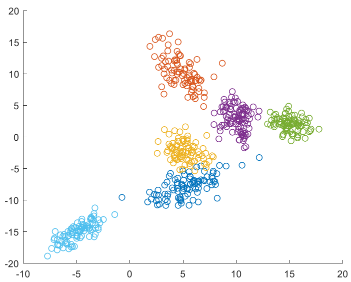
Results: Accuracy of classification with k-nn using different distance measures

	K=3	K=4	K=5	K=6	K=7
Manhattan distance	0.9870	0.9759	0.9833	0.9796	0.9815
Euclidean distance	0.9889	0.9796	0.9833	0.9852	0.9889
Chebyshev distance	0.9889	0.9778	0.9833	0.9815	0.9907
Canberra distance	0.9759	0.9704	0.9741	0.9685	0.9796

From results can be seen that most better results gives the Euclidean distance, but all others are not much behind. Also the results don't differ much when using different K-s, it shows that the clusters are very well distinguishable.



Joonis 1 Generated clusters

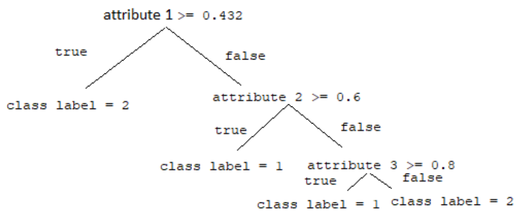


Joonis 2 Classified validation data

Exercise 2. Decision trees

Creating a decision tree starts with evaluating splits in dataset with help of Gini index. Gini index shows attributes measure of distribution, the attribute with the best Gini index is used for first split in decision tree. A split in the dataset involves one input attribute and one value for that attribute. It can be used to divide training patterns into two groups of rows. A Gini score gives an idea of how good a split is by how mixed the classes are in the two groups created by the split. A perfect separation results in a Gini score of 0. [1]

After finding the best splits, a decision tree is constructed. Each node in the tree has a value, this value determines the split point. While classifying, all attributes are given to the tree and for each attribute there is a node that decides, whether the attribute is bigger or smaller than the splitting value. After going through all nodes, the class label is known.



Joonis 3. Illustrative example of decision tree.

Data from my work consists of 1372 rows, each row has 4 attributes and a class label. 70% of rows are for training phase, 30% are for validation phase. The tree is constructed using training data. Accuracy can be validated with help of cross validation.

Result: 76.2136% accuracy

Exercise 3. Regression

A linear model has the form $Y = ax_i + b + \varepsilon_i$. The constant b_0 is called the intercept and the coefficient b_1 is the parameter estimate for the variable X. The ε is the error term. ε is the residual that can not be explained by the variables in the model. Most of the assumptions and diagnostics of linear regression focus on the assumptions of ε. [2] The goal of regression is to find estimates of the coefficients a and b sum squares of ε_i would be minimal.

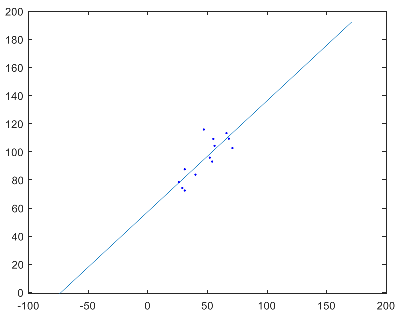


Figure 1 Example of linear regression with few points

In this work data consists of 4 predictors (independent variables) x1,..,x4 and response y (dependent variable).

Regression procedure:

1. Regressing y on x1, y on x2, y on x3 and y on x4.
2. Getting t-test P-values for each predictor, predictors with smallest P-value are used for next step.
3. Regressing y on x4 and x1, y on x4 and x2, y on x4 and x3, because x4 had the lowest P-value.
4. Enter x1 to our model, because it has the smallets P-value.
5. Regressing y on x4,x1 and x2 and y on x4,x1 and x3.
6. Removing x4, final model contains x1 and x2.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	52.577	2.2862	22.998	5.4566e-10
x1	1.4683	0.1213	12.105	2.6922e-07
x2	0.66225	0.045855	14.442	5.029e-08

The **p-values** help determine whether the relationships that you observe in your sample also exist in the larger population. The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is effect at the population level. [4]

Final model: y = 52.58 + 1.468x1 + 0.6623x2

Model is validated with least quares method: we take the squared value of our real data points minus the approximated values. After that we add those values up and divide them by the number of data points we have, taking the average. [3] $\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$

$R^2 = 0.6663$

Exercise 4. Gradient descent

Gradient descent is the first order optimization algorithm for finding the local minimum of a function.

Doing gradient decent means applying partial derivatives with respect to both m and b (y=mx+b) to the cost function to point us to the lowest point. When we reach close to, if not, zero with our derivatives, we also inevitably get the lowest value for our cost function. [3]

Gradient descent gives result as m and b values and cost.

Materials

[1] https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/

[2] https://support.sas.com/resources/papers/proceedings/proceedings/sugi22/STATS/PAPER267.PDF

[3] https://towardsdatascience.com/linear-regression-using-gradient-descent-in-10-lines-of-code-642f995339c0

[4] https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Data for excercise 3:
https://newonlinecourses.science.psu.edu/stat501/node/329/?fbclid=IwAR1QXo4IaWeaYU6QnCBP0aIRzL-wdY15EmufgIXlyrVEkDLXXupK3jtWWwM

Code:

https://gitlab.cs.ttu.ee/Tiina.Sumeri/iti8665-2019/tree/master/kodutoo2