Homework Assignment 3

2-INF-150: Machine Learning, Fall 2021

Deadline: 07.01.2022 23:55
(electronic submission through classroom)

1. Vapnik-Červonenkis dimension.

a) Consider a hypothesis space consisting of all circles on the $\mathbb{R}^2$ plane (all points inside the circle and at its boundary are classified as positive, all points outside are classified as negative). What is the VC dimension of the hypothesis set? Prove your answer.

b) In the lectures, we have shown that rectangles with axis-aligned boundaries in $\mathbb{R}^2$ plane (all points inside, including boundaries, are classified as positive and all points outside are classified as negative) have VC dimension 4. What is the VC dimension of axis-aligned multidimensional rectangles in $\mathbb{R}^d$ space? Prove your answer.

2. Average house prices. In this task, you will implement and use a simple function for PCA in python. You can base your implementation on the materials from both tutorials and lectures. You can use **only basic library functions**, such as functions to computer averages, standard deviations, eigenvalues, etc. If you implemented some parts of the code in the tutorial, you are allowed to reuse it. However, **you are not allowed to use library functions that directly compute PCA.**

a) Write function `myNormalize(x)` that normalizes input data in matrix x so that the average of each feature will be 0 and standard deviation will be 1. (Rows of the matrix are data samples, columns represent features.)

b) Write function `myPCA(x,k)` that for a given normalized input matrix x returns first k eigenvectors. Use the same algorithm that was presented in class, your program should contain the following easily identifiable sections: creation of a covariance matrix, computation of eigenvalues and eigenvectors, sorting according to the eigenvalues, returning first k eigenvectors.

c) Apply your implementation to the data set of average prices of houses in Boston. The data set is available in StatLib archive at `http://lib.stat.cmu.edu/datasets/boston`

Use the first 13 attributes from the data set to perform PCA, **omit the 14th target attribute (average house price) from the analysis.** Using a 2-dimensional graph, display first vs. second principal component and highlight (e.g. using color) all houses that have 14th attribute at most 15 (cheap houses) and also highlight (e.g. using a different color) houses where 14th attribute is at least 30 (expensive houses).

What conclusions can you draw from the resulting graph and from the composition of the first and second principal component?

Hand in all of the code (with appropriate comments), resulting graph, and your conclusions.