Machine Learning II: Assignments #1
14 performance points (max),
email: PDF+code to jan.nagler@gmail.com
due: Tue, October 8, 2019

1. Data creation and sample size realities
   (a) Generate a random artificial data set X, Y (surrogates) that shows no linear correlations but nonlinear correlations: Ideal (target) covariance $cov(X, Y) = 0$, yet X and Y are not independent and feature dependencies of your choice.
   Can the ideal target covariance of zero be reached? If not, why not?

   (b) Create surrogates exhibiting a given correlation coefficient $r = r_{xy}$ (parameter of the function). Create target examples for $r = 1$, $r = 0$, $r = 0.5$, $r = -0.5$ and $r = -1$. Decide yourself which plots you want to present and are meaningful.

   (c) Implement a causal relationship of the common effect case. Compute the correlations (in terms of r) between X and Y, and X and Z.

   (d) optional
   Study numerically how the sample variance of the sample mean of n samples of a random variable with target μ and target σ^2 depends on the sample size n. Target refers to the mean and variance of the ideal random variable (not the realized sample).
   Is it $var(\overline{X}) = \sigma^2/n$?

2. PCA
   Create a surrogate data set for the cases (a, 4 blobs) and (b, 2 touching parabola spreads) as shown in the lecture, but in a higher-dimensional space (not 2d). Perform a PCA/Class prediction with ovr logistic regression analysis as developed in the lecture. Study prediction boundaries.

3. K-Means
   (a) Create surrogate data in 2 dimension. Create 4 blobs (clusters), labeled. Perform k-means analysis as shown in lecture. Design data such that the 4 blobs are not overlapping.

   (b) Design data such that the 4 blobs are partially overlapping. Compare the elbow plots of (a) and (b).
   Details are given in class.

   (c) optional
   Study more complicated cases, find or develop quantitative measures.