Data Mining Exercise 5: 08.12.2022

Remember to enroll yourself on the examination of the current course no later than 8 days before an examination date!

1. Iris.txt file contains variables that are highly correlated. In this case principal component analysis (PCA) can be used in reducing data dimension. Run pca() using Matlab on Iris data. How many percent the two first principal components explain from the whole data variance?

2. Order the importance of variables in the Iris data using the Relief algorithm. Use nearest neighbors from 3 to 10 and test whether the number of nearest neighbors affects the result or not.

3. Select 40 first cases from each class using original Iris data set. Train a Naïve Bayesian classifier using these 120 cases (fitcnb). Use your model and classify the remaining 10 cases from each class (predict). Give your classification result in the form of a confusion matrix.

4. Repeat the third task using two principal components that you calculated in the first task.

5. If merely nominal (excluding binary) variables are included, the encoding of data can be prepared for some data mining algorithms as neural networks as follows. To simplify, let all variables have the same dimensionality of m (relatively small) for their categories. A value can only be one of the m alternatives. We encode each value with a bit sequence of length m, in which the other bit values are equal to 0 except one that is equal to 1. The latter corresponds to the nominal value, which a case includes for the pertinent variable. This encoding was used for amino acids, which forms a set of {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, when m=20. For instance, if a value occurs to be C, the second bit is 1 and all the others are 0's. Encode polyproline type II (PP II) of amino acid sequence chunk GGKAPAMM. This encoding is valid as neural network input. What is its disadvantage regarding the number of input variables (nodes) required by a neural network? How many input nodes (number of bits) are needed? See all distance measures considered in the lecture material. Which one of them would be suitable for binary data considered here? How many different symbol sequences are there if all possible amino acids are scanned for as long sequences as the preceding sample GGKAPAMM? Give also its approximation in the form of $x \cdot 10y$. (Actually, PP II structures are not searched for with mere sequences, but other information is also required, which results in more complex search problems.)