Deliverable-Thyroid Disease Classification

Alyssa Hajj Assaf

March 3, 2021

Description

The model used for the Thyroid Disease Classifier is a Parzen window with Gaussian kernel, which is a Kernel density estimation algorithm. Given the as inputs the T3-resin uptake test (percentage), the total Serum thyroxin (measured by the isotopic displacement method), the total serum triiodothyronine as measured by radioimmuno assay, the basal thyroid-stimulating hormone (TSH) (measured by radioimmuno assay) and the maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value, it returns the estimated pathology for the given patient's laboratory result, mainly (1)-Normal Thyroid function, (2)- Hyperty-roid and (3)-Hypothyroid.

When a new data point x is given to the classifier, it calculates the following sum:

$$\hat{f}_x(x) = \frac{1}{N\sqrt{2\sigma^2\pi}} \sum_{i=1}^{N} \exp -\frac{1}{2}(\frac{d(x_i,x)}{\sigma})^2$$

Where x_i belongs to the training data set, N is the size of the training data set and σ is a parameter set to 1.5 during the training phase of the algorithm over 3-fold cross validation. The function $d(x_i, x)$ is the Euclidean distance

$$d(x_i, x) = \sqrt{\sum_{j=1}^{P}(x_{ij} - x_j)^2}$$

Where $j \in P$ is the parameter space. The classifier computes the $\hat{f}_x(x)$ for each label class and assign to x the label for which $\hat{f}_x(x)$ is maximum. Hyperpamaters of this classifier are the Euclidean distance function (chosen for $d(x_i, x)$) and the value of k for the k-fold cross validation. procedure

Time and Space Complexity

In the learning procedure, we have $\frac{N}{4}$ for both validation and the test set in each fold, for a total of 3 fold. Thus, for each data point in the validation set ($\frac{N}{4}$), we calculate an Euclidean distance matrix, parse the matrix to calculate the weighted sum for each label and take the maximum over the label space($\frac{N}{4} + \frac{N}{4} + 3$). This procedure is done over 3-fold The training process of

this classifier: $O[3 * \frac{N}{4}(\frac{N}{4} + \frac{N}{4} + 3)]$. The learning procedure complexity of this classifier is $O(N^2)$. Evaluation process time complexity is $O(N^2)$ as well, since the reduced size of the data set acts as a constant.
The space complexity is dominated by the Euclidean distances matrix in the evaluation procedure, which is $O(N * P)$. Thus, a matrix of size $\frac{3}{4}N * P$ is created and deleted at each iteration over the testing set and represents the highest amount of memory allocation for this classifier.

Specification on the 3-fold cross validation
The thyroid data set used was shuffle and evenly divided into 4 parts. Three of these for parts were kept for the 3-fold cross validation procedure ans one was kept for the evaluation procedure.