Problem 1: Radiation Therapy (15 pts)

Let us assume that the dataset we are making predictions of is about breast cancer. In this case, the dataset could be divided into two types: With and without tumors. The dataset with tumors can be divided into benign tumors and malignant tumors.

Generally, the classification of pathological images will go through three steps: 1) image preprocessing; 2) feature extraction; 3) predictive modeling. Traditional feature extraction mainly relies on manually extracted features and is combined with machine learning to learn the manually extracted features, to judge new medical records images. Although this method can obtain more representative feature information, it is time-consuming and labor-intensive and requires professional knowledge of pathology. In addition, it is difficult to cover the entire feature space by manually extracting features, which leads to the weak generalization ability of such methods and cannot be applied to clinical diagnosis.

Algorithm:

(1) Based on a preset window, the full-section pathological image is divided into a plurality of divided regions of the same size.

(2) Based on the soft attention mechanism on each divided area, each pixel is given a weight value, and each divided area is cropped into multiple image blocks of the same size to obtain areas containing important information.

(3) The region containing important information is input into the classification network, to calculate the cancer probability value of each image block through the classification network.

(4) The image blocks of the full slice pathological image are sorted according to the cancer probability value, and outliers are removed based on a grid screening mechanism to obtain features of a maximum number of image blocks.

(5) The acquired features are input into the LSTM network model, and the image-level classification result of the full-slice force image is calculated.

Problem 2: Prioritizing patients (15 pts)

For blood tests, many blood metrics reflect a person's physical health. Data such as white blood cells, red blood cells, platelets, or hemoglobin. There are dozens of different indicators for different inspection directions and methods. If all these indicators are used for the input features of the model, there is no doubt that the model is likely to over-fit, because these indicators may not contain the information we need, which will confuse the machine and make wrong judgments.

Thus, after normalization, what we want to do is filter on features. The main method I would like to use is the t-test. This method is a simple method often used to filter features, and its main process is to calculate the degree of difference between the means of the population represented by two samples.

Assuming dataset M has n vectors: $x_i = (x_{i1}, x_{i2}, \ldots, x_{it}), 1 \le i \le n$. n is the number of the samples, t is the number of features in each swampland e, x_i represents the feature vector of the i[th] sample instance. In each sample, $Y \in \{+1, -1\}$. All of the samples with $y_i = +1$ have a sample mean μ^+, sample variance σ^+. The samples with $y_i = -1$ have a sample mean μ^-, sample variance σ^-. The statistic calculated by the method is:

$$T(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{\sigma_i^{+2}}{n^+} + \frac{\sigma_i^{-2}}{n^-}}}$$

n^+ is the total number of samples with x_i label $y = +1$ in the total samples, n^- is the total number of samples with x_i label $y = -1$ in the total samples.

The larger the value of the statistic calculated by the t-test method, the greater the difference between the means of the two samples; the smaller the value of the statistic calculated by the t-test method, the smaller the difference between the sample means. Finally, the calculated values of the t-test statistic of each feature are sorted in descending order, and the ability of the first feature to distinguish two types of samples is stronger.

After features reduction, I want to implement a support vector machine model. A support vector machine is a classifier that can maximize the interval between categories. It can solve the problem of linear inseparability in low-dimensional space. The specific solution is to achieve linearly separable data after mapping to high-dimensional space. Support vector machines are suitable for data classification problems with features such as small samples, nonlinearity, and high dimensions.

The commonly used kernel function forms of support vector machines mainly include polynomial kernel function and Gaussian kernel function.

1. Polynomial kernel function. The specific form of the polynomial kernel function is:

$$K(x, z) = (x \cdot z + 1)^p$$

2. Gaussian kernel function. The specific form of the Gaussian kernel function is:

$$K(x, z) = \exp\left(-\frac{\| x - z \|^2}{2\sigma^2}\right)$$

How to choose the kernel function, the parameters of the corresponding kernel function, and the penalty factor C, there is no unified and authoritative standard. At present, the selection of the kernel function, the parameters of the kernel function, and the penalty factor is mainly judged based on experience, which sometimes causes the problem of overfitting or under-learning.

The specific process of feature screening using a t-test is to divide the training samples into cancer groups and normal individual groups. Each of the same indicators corresponding to the cancer group and the normal individual group were subjected to a t-test respectively, and then the indicators were sorted from small to large according to the p-value corresponding to each indicator. Because the null hypothesis of the t-test method is that the mean values of the data corresponding to the same indicators in the cancer group and the normal individual group are equal, the smaller the p-value calculated by the t-test method, the more confident it is to reject the null hypothesis, that is, the null hypothesis that the mean values of this blood index data in cancer patients and normal individuals are rejected. Therefore, the smaller the P-value, the greater the data difference in the blood index between the cancer patient group and the normal individual group. Thus, we can select a specific number of features, like 20 features, to train the model.

Then, classify the data by SVM. In using the SVM classifier to discriminate, the result of SVM is

very sensitive to the selection of parameters and is mainly affected by two parameters. On the one hand, is the penalty factor C and on the other hand the form of the kernel function and the corresponding parameters. Different kernel function forms have different parameters. Different kernel function forms have different parameters. According to common usage, I choose the RBF kernel function. Its mathematical form is $\exp(-\gamma|U-V|^2)$, the only parameter is γ. Its advantages are good generalization ability and very few parameters. After the RBF kernel function is determined, only the parameters C and γ affect the SVM. Because the sample size may be large, the use of cross-validation is computationally expensive and time-consuming. So, I choose the grid search method. This method refers to searching for the optimal parameters according to a given step size within the range limited by the parameters.

Finally, preprocessing, fitting, and analysis are performed according to the provided blood index data of cancer patients and normal people so that a preliminary diagnosis of patients can be made in advance.