DAT410 Module 5 Assignment 5 – Group 26

Yahui Wu (MPMOB) (15 hrs)
yahuiw@chalmers.se
Personal number: 000617-3918
Tianshuo Xiao (MPMOB) (15 hrs)
tianshuo@chalmers.se
Personal number: 000922-7950

February 21, 2023

We hereby declare that we have both actively participated in solving every exercise. All solutions are entirely our own work, without having taken part of other solutions

Reading and reflection

This article discusses the accuracy of fine-needle aspiration biopsy in the diagnosis of breast cancer. It reviews 56 cases from 1976-1984 and compares the results of digital cell image analysis, nuclear morphometry, and other methods. The study found that digital image analysis and machine learning techniques can improve the accuracy of breast fine needle aspirates. The best single-plane diagnostic classifier based on mean texture, the worst area, and the worst smoothness separated 97.3% of the cases successfully. The projected prospective accuracy was 97%.

From this article, we get some important features that can be applied to model building. The best single-plane classifier separated benign from malignant points based on three nuclear feature values for each case: mean texture, the worst area, and the worst smoothness. This also means that we have to focus on the data of _2, which represents "worst" value over a set of samples.

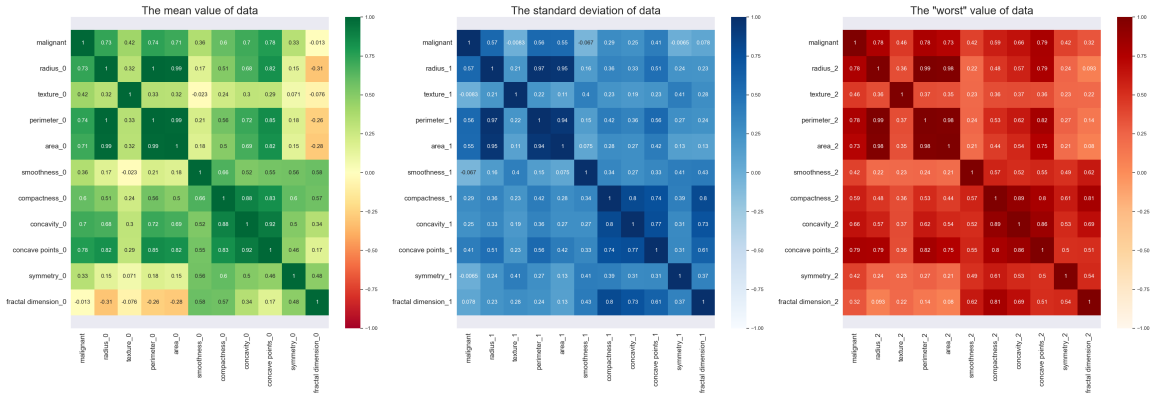
To deploy the machine learning model we built and make sure it has practical value in reality, we would like to make the model intepretable. The interpretability of a machine learning model is important and it is not suffice to only obtain a well predicted output from the model, we also want a good explanation of the outcome. Although the definition of interpretability is lack somehow, there is a growing body of literature proposes purportedly interpretable algorithms and some aspects to depict the interpretability of a model. Firstly, interpretability could be regarded as a prerequisite of trust. Does the ML model seem to be reliable for users can reflect the interpretability to some extent. However, it is hard to define trust and the definition could be subjective. Secondly, researchers would like to find the causality of a given outcome by a model rather than only finding relationships between the input and output superficially. A ML model with good interpretability could let researchers infer causal relationships from observational data. Thirdly, a well-designed ML learning model should perform transferability when the environment where it was deployed is unstable. Humans will not make a completely different choice when there is a subtle change of the factor that affects their behaviour but ML models will do. While the machine-learning objective might be to reduce error, the real-world purpose is to provide useful information and such an ability is called informativeness. The last point is that the ML model should make fair and ethical decisions and such a demand requires good interpretability of a model.

Interpretability has two techniques and model properties - transparency and post hoc interpretability. Transparency is considered at the level of the entire model (simulatability), at the level of individual components such as parameters (decomposability), and at the level of the training algorithm (algorithmic transparency). A interpretable model should be an simple model which is easy to be fully understood by users. Simulatability may admit two subtypes: one based on the size of the model and another based on the computation required to perform inference. To make a model more interpretable, each part of the model should be explained and understood by users and it's called decomposability if a model can achieve that. The last notion of transaprency is algorithmic transparency. No matter what data is given as an input, the algorithms should converge to a unique solution.

Unlike transparency requires a precise and clear explanation of a ML model, post hoc interpretability represents a distinct approach to extracting information from learned models that does not elucidate precisely how a model works. There are several methods to make a model more interpretable in terms of hoc interpretability. Besides a model to make predictions, text explanations method could use another model to generate an explanation of the prediction. Another common approach to generating post hoc interpretations is to render visualizations in the hope of determining qualitatively what a model has learned. Local explanations focus on explaining what a neural network depends on locally instead the whole mapping. Explanation by example is to find which other examples are most similar with respect to the model.

Implementation

1. A rule-based classifier
First we processed the data and drew the correlation hotspots for the three cases.



Correlogram of mean data Correlogram of standard data Correlogram of worst data

Figure 1: Correlogram of different data

By comparing the mean, standard deviation and "worst" value over a set of samples, we found that the correlation between different features and malignant was highest in "worst" value, so we plotted the correlation between 10 features in this set of data. We found that the highest correlation was found in "worst" value.

radius texture perimeter

Figure 2: 10 features' simple binary test

The ideal test should be such that there exists a threshold that splits FN and FP better in two regions.From the above ten figures, we can find that radius, concave points, and perimete thresholds are more obvious, so we are going to choose these three parameters as the criteria for the classifier.

radius vs perimeter | radius vs concave points | concave points vs perimeter

Figure 3: 3 features' Scatter plots

In the assignment, we are asked to consider: cell size, cell shape, cell texture and cell homogeneity. The radius and perimeter of the cell we choose can reflect the shape and size of the cell.For the homogeneity of the cells, we decided to choose concave points instead by reading the paper.
By calibration, we determine its threshold value as follows:
radius: 18
concave points: 0.16
perimeter: 120
Finally, we validate our rule-based classifier.
The accuracy score is: **0.9420**
The F_1 score is: **0.9177**
2.Random forest classifier
We randomly split all the data into a train and a test set, We randomly divide all the data into training and test sets and apply them to random forest classifier.

	precision	recall	f1-score	support
0	0.96	0.99	0.97	71
1	0.98	0.93	0.95	43
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

Table 1: Random forest classifier classification report

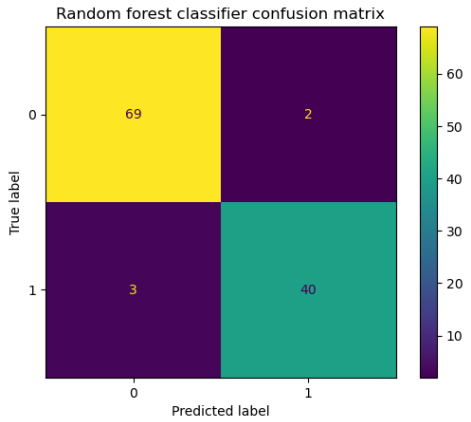


Figure 4: Confusion matrix

We find the predictor variables that the random forest considers most important, as represented in the following table:

Feature	Importance
concave points_2	0.153929
area_2	0.142351
perimeter_2	0.112579
concave points_0	0.100162
radius_2	0.095216

Table 2: Random forest classifier important features

From the above table, we can know that:
The accuracy score is: **0.96**
The 1 F_1 score is: **0.95**
The 0 F_1 score is: **0.97**

3.Trade off
The transparency of a model can be a property which is proposed to confer interpretability according to Zachary Lipton's paper. A model with good interpretability should be transparent enough for users to simply understand its working mechanism by giving users visual or textual artifacts. Decision tree could be our first choice since each nodes and the threshold could be visualized by using "plot_tree" function of DecisionTreeClassifier from scikit-learn. From the visualization of how the diagnosis was made by the decision tree classifier in this task, we can easily trace each step of the classification and the threshold selected in each node. The decision tree classifier gives us the clue of how the decision was made for each instance which means we can not only make diagnosis for each patient but also give the reason of making such a diagnosis. It seems that decision trees have interpretability to some content. However, in our case, there are many features taken into account which could lead to a bigger model with longer computational time. The massive features could also make the diagnosis more difficult to explain and reduce the interpretability. Therefore, we would like to have less but more representative features to make sure we can explain our diagnoses without jeopardizing the accuracy a lot at the same time. We used the parameter called "max_features" in the DecisionTreeClassifier from scikit-learn which could set the number of features to consider when looking for the best split.

We did the classification several times with different number of features for each time and recorded the accuracy score and F1 score. We gave the plot showing the relationship between the scores and number of features in the model.

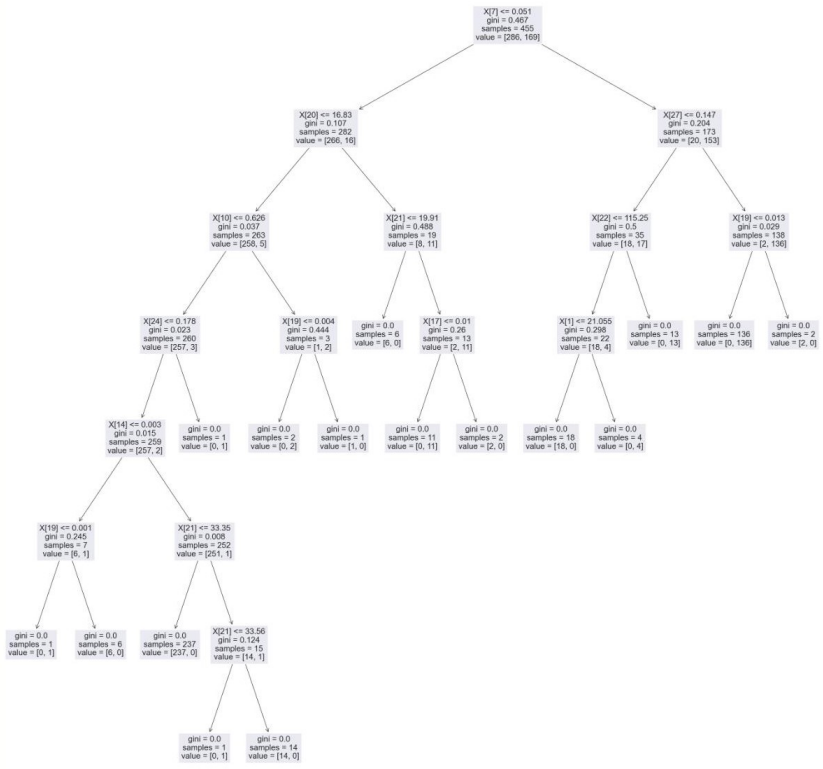


Figure 5: Decision tree

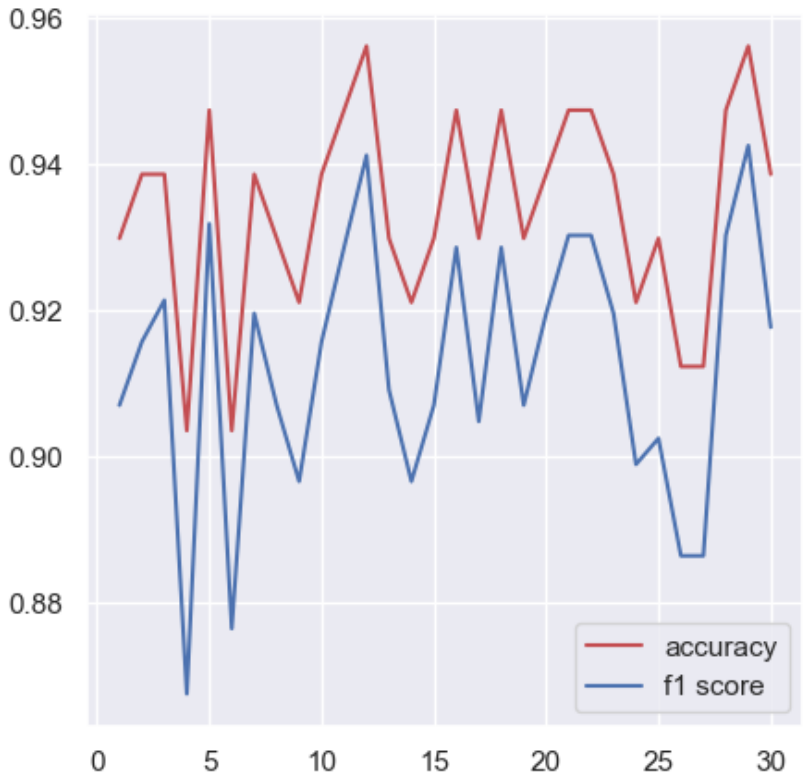


Figure 6: Accuracy and F1 score with different features selected

From the figure above, we found that when there were 12 features selected, the model had a good performance with highest accuracy and F1 score and the number of features selected could make the model interpretable.

Answers
Compare the results of these three classifiers:

	Rule-based classifier	Random forest classifier	Decision tree classifier
Accuracy	0.942	0.96	0.938
F1 score	0.918	0.96	0.917

Table 3: Three different classifier comparison

By comparing these three models, we can come up with a ranking:1.Random forest classifier 2.Rule-based classifier 3.Decision tree
1.Rule-based classifiers have the highest interpretability. It is the most interpretable because when we building the classifier, we explicitly point out the features and thresholds for the classification and are clear about its working process.When we choose the classification features, we are judged by the high or low relevance, so the classification results will be more accurate. But there may also be misclassification. We shift the threshold to the right in order to avoid generating misclassification.
2.Random forest classifier. The random forest classifier is less interpretable due to the fact that the random forest algorithm combines multiple decision trees, making it challenging to interpret individual

decision trees and the importance of each feature. However, the random forest classifier performs better.
3.Decision Tree Classifier generates a tree-like model of decisions, making it easier to understand the decision-making process.
From Figure 1 we can get that, the correlation between perimeter and radius is 1. There are some significant interactions between features. Besides, when we select some features to plot the accuracy and f1 score curves(Figure 6), we found that the performance of the model is constantly changing, so there is an interaction between different features.

Discussion

Yahui Wu

Interpretability of a ML system is important since it relates to whether people using such a system can rely on the output it gave or not. Moreover, it limits the application of ML systems to many areas where rigorousness and ethically made decisions are needed. Although we still lack of a widely agreed definition of interpretability today since it is a subjective topic, we have some properties that could reflect and confer interpretability of a ML system. Unlike a ML model with "black-boxness" that users can't learn any working mechanism from the model, transparency requires a model to be simply understood by users. In the case of medical diagnoses, we would like to say the model has good interpretability if it can tell us based on what (clinical symptoms, FNA results.) it made the diagnosis. In other words, the model needs to display each step of its decision made sequence thus we can explain the diagnosis to the users or patients. In our last implement of the diagnostic system, we used decision tree classifier and draw the decision tree to show all the nodes and the threshold of each node. We consider the decision tree to be a model with interpretability based on its transparency on the premise that the features of the data are understandable and not abstract so we can straightforward explain the diagnosis based on the features. The advantage of defining interpretability by transparency is that we have several options to build a model which has transparency itself such as simple rule-based classifier, decision tree and single-feature linear classification model. However, when we seek higher transparency of a model, the accuracy of the diagnosis could be affected and the decision model could have transferability issues despite it gives us sufficient explanations.
Another property to confer interpretability of a model is post hoc interpretability. Post interpretability provides users with explanation on the model which has already been trained rather than the working mechanism of the model. For those models that have "black-boxness", post hoc interpretability could be a choice. The common approaches to post hoc interpretations include natural language explanations, visualizations of learned representations or models, and explanations by example. In the case of tumor diagnosis, we can compare the new cases introduced to the diagnostic system with diagnosed previous cases. We can find the distribution of the features in test set given a specific diagnosis $q(X|Y)$ and the distribution of the features in training set $p(X|Y)$ to see if there are similarities between new cases and previous cases. The advantage of this concept of interpretability is that opaque models can be interpreted after the fact, without sacrificing predictive performance.

Tianshuo Xiao

Definition:
Interpretability can be summarized by the following equation, $Argmax_E Q(E|Model, Human, Data, Task)$. Q is an explanatory evaluation equation and E is a specific method to achieve interpretability. The whole process is for us to seek an interpretative method that allows a specific group of people with a specific human experience to have a maximum understanding of a specific model given a specific data and for a specific task.
In this assignment, we can see that there are ten different cell features in the dataset and three different sets of measurements. Our goal is to make a prediction about whether it is a malignant breast cancer or not. We set up classifiers based on the cellular status of different malignant tumors and make predictions.
The goal of pre-training interpretability is to visualize the data with as best understanding of the data as possible. The interpretability goal of the training process can be expressed as
$Argmax_{E,Model} Q(E|Model, Human, Data, Task)$. Training the model while creating interpretability in the training process is actually equivalent to the interpretable model that we create. We aim for the best model performance. For example, the decision tree that we used in the third question. The decision tree is interpretable, but we also know that as more and more decision rules are added, the

whole tree becomes very difficult to maintain, and according to the curve in the figure we find that the correctness rate oscillates.

The interpretable goal after training is to explain the decision basis of these black box models. Sensitivity analysis is to see for which data instances your model is very sensitive. For example a classifier, if we remove a data point and there is a drastic change in the decision boundary of the model, then we say that this data point is very sensitive and it is also an important basis for the discrimination of this classifier. In this assignment, we take the sensitivity analysis and shift the threshold right to ensure that we do not miss the case of cancer.

Potential benefits and drawbacks

When we evaluate models, most of the time we consider the performance of the model, that is, metrics such as accuracy. But accuracy is not the whole point in realistic and complex problems.

Interpretability has several potential benefits in the healthcare field. First, interpretability can increase trust in models and predictions because clinicians can understand how the models arrive at their decisions, rather than a black box state. This trust can lead to greater adoption of predictive models and other decision support systems in healthcare organizations. What is more, interpretability can help clinicians identify errors and biases in the models and correct them. Finally, interpretability can provide insight into the characteristics and factors that are most important in predicting patient health outcomes, which can help clinicians develop more effective interventions and treatments.

However, there are potential drawbacks to interpretability in healthcare. Interpretability can be time-consuming and resource-intensive, as clinicians may need to invest significant effort to understand and validate the model's output. Besides, highly interpretable oversimplified models may sacrifice predictive accuracy for the sake of interpretability, such as the difficulty of balancing sensitivity and specificity in the setting of thresholds in this assignment. In adddition, interpretability may be less important in some clinical settings where speed and accuracy are paramount and the ability to interpret the model output is less critical. Some domains of medicine deal with interpretability not as true interpretability but as interpretability for the specific medical problem, an exercise that examines whether deep network algorithms for medical diagnosis problems are consistent with existing human knowledge of the corresponding medical problem.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. (2016), 11351144. DOI:doi.org/10.1145/2939672

[2] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning."arXiv preprint arXiv:1702.08608(2017).

Reflection on the previous module

Yahui Wu

In the previous module, we implemented a machine translation model based on IBM model one which contained two parts - language model and translation model. The language model is based on the Hidden Markov Model (HMM) and the translation model is based on the EM algorithm to find most likely translated word from a word alignment. In the more specific decoding part of our assignment, we used word alignment to find the most likely translated word given the source sentence and used language model to find the best sentence among permutations of translated words. In reality, the word alignment technique can be applied to DAN test and image processing. The language models can be used for problems with time series like signal processing. In the translation model, the word alignment could be regarded as a latent variable which is unobservable. Similarly, the recommendation system we designed in the previous assignment finds the preference of different users and evaluate it as a latent variable using some algorithms. However, if we try to compute the probability of a sentence that contains a word that did not appear in the training text, the translation system will crash.
In current neural networks translation, there are some "stereotype" issues caused by biased training data. It could be a good behaviour since the training data we used comes from our real world and its bias could be suitable for daily contexts.

Tianshuo Xiao

In the last lesson, we focused on natural language processing and reflected on it while we were doing the assignment. In the last assignment, the main point of natural language processing was word-to-word correspondence. Rule-based translation systems and neural network translation systems have something in common in that they both contain an intermediate expression meaning. Neural networks mainly rely on the encoder to encode the source text in terms of meaning. This allows for training on many languages simultaneously. In rule-based translation systems, local dialects, for example, can be difficult to translate because of the relatively small amount of available training data.
In the process of translation, we use maximum likelihood estimation to ensure that each word in the sentence is the highest probability. By using a word-by-word translation order, the top k words are then selected from the target topic and then they are ranked according to the probability and selected the first ranked word. Another translation method is direct translation, where a number of possible translations are obtained for each word. We also need to pay attention to the grammar and order of the sentences.
When we have a large amount of text to translate, for example 10 TB, we can take a random sampling approach to translation, which means dividing a large task into subtasks, for example using a random forest algorithm. When the probability is small, we can use the logarithm of the probability to make the values work with more stability.
The translator sometimes automatically selects the gender of the subject. When evaluating the performance of the translator, you need to consider the grammar of the sentence, fluency, etc. For example, a KNN algorithm can be used to calculate the relative distance. Another approach is to replace the sentences in some way and analyze the similarity of the strings in the sentences.

Summary of lectures

Yahui Wu

Diagnostic Systems
February 14, 2023
In the last class, we discussed about diagnostic systems. The early expert systems use logistic for diagnosis inference based on a knowledge base. In the 1970-80s, Miller et al. developed INTERNIST-1 which was a rule based ranking system for automated diagnosis. In the 1990s, neural networks were used for breast cancer diagnosis with limited data. To improve the early knowledge-based diagnostic systems, we need to figure out the definition of diagnostics. Basically, a diagnosis task is to find out the cause of some phenomenon and in the case of medical practice, it is to infer the disease causing a symptom. Therefore, diagnoses are based on data of the patient that could be collected from different medical records. There are international diagnostic criteria specifying combinations of signs, symptoms and tests to determine diagnosis. The core of automated diagnostics are tests. In a test, for example, a binary test, it is important to determine a properly chosen threshold, the selection of threshold has significance in sensitivity and specificity tradeoff. However, when it comes to multiple tests/symptoms, the threshold measurement could perform poorly. Naive Bayes could be helpful in this case since it assumes all symptoms are conditionally independent given the disease. We can build the Bayes model based on: $p(D|X_1, ..., X_2) = p(D|X_1, ..., X_j - 1)\frac{p(X_j|D)}{\sum_d p(X_j|d)p^{j-1}(d)}$ where D is the disease j is the jth symptoms. By using Naive Bayes model, we can readily estimate $p(X_j|D)$ from data and it can be used for differential diagnosis or multi-disease diagnosis.

Tianshuo Xiao

Diagnostic Systems
February 14, 2023
In the last lecture, we learned mainly about the knowledge related to diagnostic systems. This included the history of diagnostic systems, the applications of diagnostic systems and how to build a diagnostic system.
Early diagnostic systems were mainly used in medical applications and were based on the logic of knowledge base for reasoning. In the 1970's, Mycin was developed and used to diagnose serious infections. By the 1990's, neural networks were used in medicine, for example for the diagnosis of breast cancer. However, these early systems had many drawbacks, such as expert systems that relied on manually managed knowledge bases, lacked training data and had poor generalization capabilities. Diagnosis refers to the task of finding out the cause of some phenomenon and pick among possible causes, which widely used in medical systems. In this system, it specifies combinations of signs, symptoms and tests that are used to determine diagnosis. For example, include clinical diagnosis, laboratory diagnosis and radiology diagnosis.
The core of diagnosis is testing and selecting the right features. The test is generally based on the originally collected data and a simple binary test is performed, from which a suitable threshold is selected. Theoretically, the threshold is optimal when it lies between FN and FP. Sometimes there is a trade-off between sensitivity and specificity. If the value more sensitive, it would be less specific. Besides, we can use ROC curve to view the tradeoff. We can keep adjusting the threshold to get superb. What is more, we can use model based tests, to estimate the probability that the assigned label is right. We take Nave Bayes as an example, test the accuracy of the model by calculating conditional probabilities and predict missing values. The most important is that we must trade off specificity and sensitivity and balance them.
In conclusion, the accuracy of machine learning for disease prediction is gradually improving at this stage. They can classify diseases according to different conditions of cells and can analyze pathological

images.