Predicting Breast Cancer Tumor Behavior through Machine Learning

Ben Goldberg

Northwestern University

CIS435 Practical Data Science Using Machine Learning

Assignment 2: Supervised Learning with Classification and Unsupervised Learning with Clustering

February 6, 2022

**Business Problem**

The field of modern medicine continues to grow and evolve. Recently, due to a worldwide pandemic, medical science has been put to the test to contain the coronavirus and return the world to normal. Coronavirus is an extreme example, but it represents how far the medical field has come in triaging problems. AI/ML has come a far way since its creation and can now be used to assist the medical field even further. Using the power of AI/ML, we can help doctors diagnose and treat patients at records speeds. One of the hardest steps in treating a patient, is correctly diagnosing the cause. With AI/ML, we can use the power of historical data and clustering to predict potential medical issues. Armed with machine learning models, medical professionals can have a quick and accurate second opinion before diagnosing a patient. Having an automatic second opinion can be the difference of life and death to a patient. In this assignment, we will apply machine learning to assist medical professionals in accurately diagnosing breast cancer tumors. The ultimate output of the machine learning model will predict if a tumor is malignant or benign based on information provided.

**Machine Learning Applications**

Before we start looking into breast cancer tumors, I want to walk through some other examples in the medical field where machine learning can be used. One example is reading chest X-Ray images to determine if a patient has COVID. A classification algorithm like Naïve Bayes, can be used to inform a medical professional if a patient had COVID, pneumonia, or normal scans. Coronavirus is a deadly virus and quick feedback from a Naïve Bayes model will enable doctors to apply preventative measures and reduce the chance of death or significant damage.

Another example in the medical field is to better understand and potentially find a cure for Alzheimer’s disease. Using a KMeans unsupervised model, we could group patient details to see if there are any underlying similarities that a doctor might not see. We can see if there are any potential indicators of Alzheimer’s disease earlier to provide preventative treatment. If we can get ahead of the disease, it can greatly increase a patient’s quality of life.

A final example is using logistic regression to determine whether a patient has type 1 or type 2 diabetes. Both types of diabetes are major life altering diagnoses, but treatment of each differs. Type 2 diabetes with proper treatment and lifestyle changes, it can be put into remission. However, type 1 diabetes must be controlled with insulin and there is no chance of remission. Both type 1 and type 2 diabetes present the same symptoms but approaches to managing it is very different. Using logistic regression, you could assist in diagnosis of which type of diabetes much earlier. A patient could be treated for that specific type of diabetes and greatly increase their quality of life.

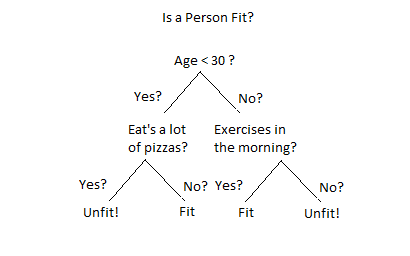
**Machine Learning Algorithms**

In this assignment, we compare 4 different machine learning algorithms using both supervised and unsupervised learning. Supervised learning machine learning models take historical data with labels to train a model and help it “learn” from previous examples. With what the model has learned, it tries to predict outcomes on new data. However, unsupervised learning does not take historical data into account. Unsupervised learning models read in the available data to cluster or group data into labels. For the analysis below, we use 3 supervised learning classification algorithms: Naïve Bayes, Decision Trees, and Logistic Regression. On the other hand, we use 1 unsupervised machine learning algorithm: K-Means.

Text

Description automatically generated with medium confidenceNaive Bayes algorithm is a supervised learning algorithm based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. (scikit) The below formula is used to read in features and predict a particular class.

Naïve Bayes is a great and simple algorithm to implement, but unfortunately the naïve portion can limit the model. In the real world, there are not many situations where all variables are equivalent. However, if that situation is relevant, this model is highly recommended.

With a Decision Tree model, the goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Below is an example of a simple decision tree.

From the example above, we can see how feature decisions can determine a particular classification. In our breast cancer tumor classification, we ask different questions to ultimately land us in either a malignant or benign classification.

The last supervised learning algorithm is Logistic Regression. Logistic Regression is a linear model used in classification. It uses a logistic function to determine whether a value falls into a binary, multinomial, or ordinal classes. (scikit) Binary classes will only have two possible outcomes, while multinomial and ordinal have 3 or more. The difference between multinominal or ordinal is whether the classes are ordered or not. In the assignment, I use binary logistic regression to determine whether a tumor is malignant (1) or benign (0). The following function is used to determine logistic regression.

Logo

Description automatically generated with medium confidence

After the supervised learning models above are used and compared, we switch gears into unsupervised learning using K-Means algorithm. The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields. (scikit) In the assignment, I will focus on 2 clusters of data where we are looking for malignant or benign groupings and what features are driving that conclusion.

**Data Preprocessing Discussion**

To ensure maximum accuracy from our machine learning models, we need prepare our data for analysis. The first step I took to understand the data is seeing its shape. Our dataset contains 569 entries with 23 features. Each of the features are floating values describing the tumor except for ID and Diagnosis which were integers. Next, I looked across the 569 entries to ensure we did not have any null values. Null values would have a large impact on the models and removal of those entries would have been critical. Lastly, I gathered summary metrics on each feature to check for any irregular values. Based on my review, none of the datapoints were needed to be removed.

**Explaining Metrics**

A screenshot of a computer

Description automatically generated with low confidenceWith the data ready and our machine learning algorithms selected, I first applied logistic regression to our dataset. There are 24 features in our dataset and we need to gather the importance of each feature. Including features which the model deemed unimportant could potentially skew the results. After running feature importance, I received the following results.

The negative values represent which features are less important to the model. We can see that some features including the *perimeter\_mean* and *concavity\_mean* are great indicators of whether a tumor was malignant or benign. However, when I removed the negative features from the model, I found that it was less performant than leaving them in. As a result, I left them in for the purpose of our analysis.

Next, I applied the Naïve Bayes algorithm to our dataset. With Naïve Bayes, it treats each feature as an independent factor and for this study, I will include all features for consistency. Because of the “naïve” treatment of features, I would encourage trimming of features to further tune the model.

**A screenshot of a computer

Description automatically generated with low confidence**For the third model, I applied a decision tree algorithm to the dataset. In the graphic below, we can see the feature importance.

When given the importance for a decision tree model, all importance must sum to 1 and less important features are pushed closer to 0. We can see in the table above that the model heavily favored *concave points\_mean* and *texture\_mean*.

Lastly, I implemented the unsupervised learning K-Means model on the dataset. When applying this algorithm, a number of clusters must be specified. For this assignment, I created two clusters to match the malignant or benign results produced from the above models.

**Interpreting Results**

**Supervised Learning:**

To compare the results of the supervised learning models above, we can use a few methods. I chose to use a confusion matrix, accuracy score, and ROC curve with corresponding area under curve to rate the results. The accuracy score is computed by counting each of the correct predictions and dividing it by the total number of possible predictions. For an accuracy score, we are looking for a value as close to 100% as possible. The accuracy score should be computed on both the training set and test set to see if the dataset properly trained the model. In the below graphic I have each of the test and training set accuracies listed for all supervised learning algorithms.

Table

Description automatically generated

The highest scoring algorithm would be the Decision Tree with a 93.71% accuracy score on the test set. Although accuracy seems promising, it does not tell the entire story. Accuracy is simply a measure of correctness but not a complete picture into the validity of your model. In order to ensure we are using the correct model; we must also look at the Receiver Operating Characteristic (ROC) curve. An ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier. In this case, it will illustrate the model’s ability to produce either malignant or benign outcomes. The ROC curve plots the true positive rate (y axis) against the false positive rate (x axis). The area under the line plotted with the curve is called the area under curve or AUC. The AUC is a value between 0 and 1 where a 1 is a perfect score. To check the accuracy of our model, we need to understand the AUC values. Below is a side-by-side comparison of the 3 Chart, line chart

Description automatically generatedROC curves and respective AUC values.

Decision Tree algorithm may have had the highest accuracy but ranked lowest in terms of AUC. This is a clear indicator that accuracy is not a full representation to determine the best model for use.

For the last comparison, we will use the confusion matrix to review our model performance. A confusion matrix is a 2 x 2 plot that details the true positive, true negative, false positive, and false negative results. True positive and true negative are results where the model accurately predicted the result. However, false positive and false negative results indicate that the model predicted the wrong result. When working within the healthcare industry, it is important to understand the implications of a false positive or false negative. A single false negative could mean the difference between life and death for a patient. In the case of a false positive, there is a bit more tolerance for this error because a patient would still get treatment, even if they didn’t need it. I have included the confusion matrix for each supervised learning algorithm below.



I highlighted the false negative results in orange based on my explanation above. I believe that one false negative could cost a patient their life and ensure that number is as close to zero as possible is a great indicator for the best model to use. Based on the AUC and false negative values, I will use Naïve Bayes algorithm for the next portion of the assignment.

**Unsupervised Learning:**

When applying the K-Means unsupervised learning model to our dataset, we can view clusters of our data. I reviewed many combinations of the dataset, but an interesting finding is focused on *radius\_mean* of a tumor. The K-Means model makes a clear delineation of malignant (red) and benign (green) tumors. Based on the graph, if the *radius\_mean* of the tumor grows greater than around 16, the model determines the tumor to be malignant. I have placed the graph below.

Chart, scatter chart

Description automatically generated

**Recommended Steps**

Comparing the 3 supervised learning models and 1 unsupervised learning model, I determined that K-Means is great for analysis, but the Naïve Bayes model should be used for prediction. Naïve Bayes model had the lowest false positive count, highest AUC and still performed with greater than 92% accuracy in prediction. Naïve Bayes model would be great to assist medical professionals in predicting whether a tumor is malignant or benign. However, I believe the model should only be used in an assisting capacity because it is still subject to error. Human review is necessary to weigh in external factors outside of what was input into the model to ensure the best patient diagnosis. In regards to next steps, with additional data points, I would further tune the model by reducing the number of features while focusing on reducing false negatives to 0.

**References**

*Decision trees*. scikit. (n.d.). Retrieved February 7, 2022, from https://scikit-learn.org/stable/modules/tree.html

*Linear Models*. scikit. (n.d.). Retrieved February 7, 2022, from https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression

*Naive Bayes*. scikit. (n.d.). Retrieved February 7, 2022, from https://scikit-learn.org/stable/modules/naive\_bayes.html

Sehra, C. (2020, November 30). *Decision trees explained easily*. Medium. Retrieved February 7, 2022, from https://chirag-sehra.medium.com/decision-trees-explained-easily-28f23241248

Wikimedia Foundation. (2021, December 3). *Confusion matrix*. Wikipedia. Retrieved February 7, 2022, from https://en.wikipedia.org/wiki/Confusion\_matrix