Breast cancer diagnosis through machine learning

Jonah Northwood

University of Minnesota Morris, Computer Science senior seminar

Fall, 2020

The Big Picture

- Survival rates of breast cancer are closely related to how early it can be effectively detected and treated
- Machine learning can be used to improve effectiveness of detecting whether tumors are cancerous or not

Table of Contents

Table of Contents

Typical breast cancer diagnosis

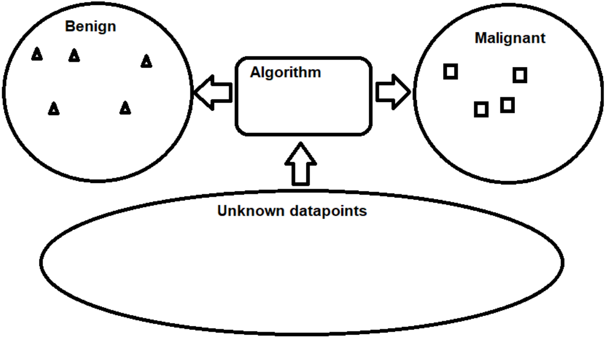
- Detecting a tumor
- Analyze the tumor
- **Classify the tumor**
- Treat the tumor as necessary

What is machine learning?

Machine learning is an application of artificial intelligence (AI) that lets systems automatically learn and improve.

Classification

- Type of machine learning
- Sorts data points into classes



Training

- Give algorithms data that has already been classified
- Learns what in the data makes it more likely to belong to one class over another
- Uses that information to classify future unknown data

Dimensionality

- Number of variables being looked at, e.g. tumor radius and texture
- Algorithms will 'plot' these points on a graph to compare them
- Can have very high numbers of dimensions which can affect algorithms in different ways
- There are ways to reduce dimensions, but some information is lost

Algorithms

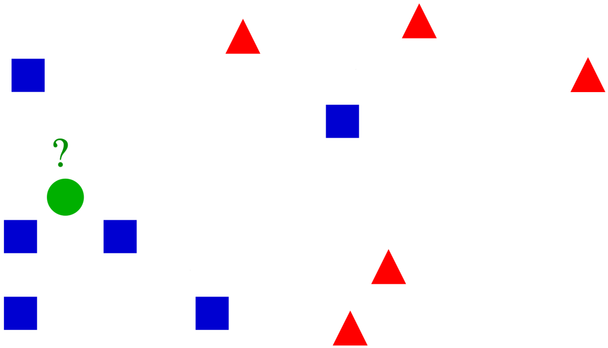
Many different machine learning algorithms
They all have different advantages and disadvantages depending on that data being looked at

- k-Nearest neighbors
- Support vector machine

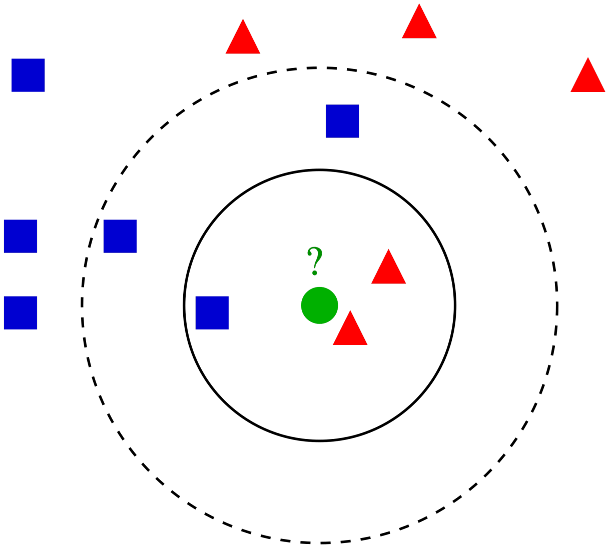
k-Nearest Neighbors (KNN)

Machine learning algorithm that classifies data points by looking at a number (k) of that points closest neighboring points

k-Nearest Neighbors (KNN)



k-Nearest Neighbors (KNN)



Support Vector Machine (SVM)

Machine learning algorithm that classifies data points by separating classes with a hyperplane

- A hyperplane is a space that is one dimension less then the one being dealt with
- The hyperplane that best separates the classes is furthest hyperplane from any given point

Support Vector Machine (SVM)

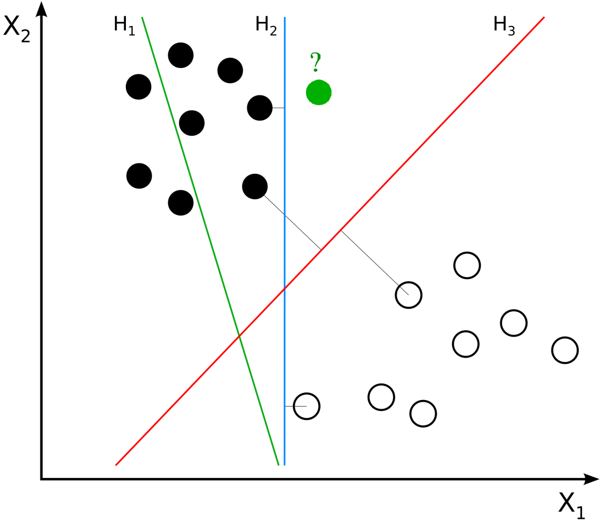


Table of Contents

Data Preprocessing

Improves performance of algorithm on a dataset

- Feature importance
- Reducing dimensionality

Feature importance

Different variables vary in scale and units
If this isn't taken into account results will be skewed
Examples include:

- Standard scaling
- Min-max normalization

Standard scaling

Scales all feature so they have a mean value of 0 and a standard deviation of 1 which makes them easier to compare

$$y = \frac{x - \mathrm{mean}(x)}{\mathrm{Stdev}(x)}$$

Min-max normalization

Rescales all features so they range between 0 and 1

$$y = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Reducing dimensionality

Principle component analysis (PCA) can be used to reduce dimensionality

- Variables that are highly correlated, like a person's height and weight, can be combined into one variable
- Concentrates most 'unique' data in a few variables
- This allows other variables to be ignored, reducing dimensions while losing the least amount of 'important' data

Algorithm Evaluation

There are many different ways algorithms can be evaluated to understand how effective they were

Different evaluations tell you different things about the performance of the algorithm

- Confusion Matrix
- Sensitivity
- Specificity
- Accuracy
- Area under curve
- Cohen's kappa

Confusion Matrix

Visually compares true positive and negative vs false positive and negative

		Prediction	
		Benign	Malignant
True	Benign	105(TP)	6(FN)
	Malignant	2(FP)	58(TN)

Table: Kaklamanis et al. SVM Confusion matrix [2]

Sensitivity

Measure of true positive rate, ranges from 0 to 1

- Sensitivity $= 0$ means no positives were predicted as positives
- Sensitivity $= 1$ means all positives were predicted as positives

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of total positives in data set}}$$

Specificity

Measure of true negative rate, ranges from 0 to 1

- Specificity $= 0$ means no negatives were predicted as negatives
- Specificity $= 1$ means all negatives were predicted as negatives

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of total negatives in data set}}$$

Accuracy

Measure of rate of data points identified correctly, ranges from 0 to 1

- Accuracy $= 0$ means nothing was predicted correctly
- Accuracy $= 1$ means everything was predicted correctly

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total number of data points}}$$

Area under curve

Area under curve (AUC) more directly represents the probability of classifying a true positive vs a false positive, ranges from 0 to 1

Developed by radar engineers

- AUC $= 0.5$ means its as good as random guessing
- AUC $= 1$ means predictions are 100% correct

Cohen's kappa

Calculation on values from confusion matrix, ranges from -1 to 1

- Cohen's kappa takes into account random chance into its evaluation
- Kappa = 0 means the algorithm is performing as well as randomly guessing
- Kappa = 1 means it is perfectly categorizing the data
- Kappa = -1 means it is categorizing everything incorrectly

Cohen's kappa

		Prediction	
		Benign	Malignant
True	Benign	105(TP)	6(FN)
	Malignant	2(FP)	58(TN)

Table: Kaklamanis et al. SVM Confusion matrix [2]

TD = total data points
TB = number of data points that are actually benign
PB = number of data points that are predicted to be benign
TM = number of data points that are actually malignant
PM = number of data points that are predicted to be malignant

Cohen's kappa

$$\text{expected accuracy} = \frac{\frac{(\text{TB}*\text{PB})}{\text{TD}} + \frac{(\text{TM}*\text{PM})}{\text{TD}}}{\text{TD}}$$

$$\kappa = \frac{\text{accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

Cohen's kappa

		Prediction	
		Benign	Malignant
True	Benign	105(TP)	6(FN)
	Malignant	2(FP)	58(TN)

Table: Kaklamanis et al. SVM Confusion matrix [2]

$$\kappa = \frac{0.9532 - 0.5375}{1 - 0.5375}$$

$$\kappa = 0.8988$$

Cohen's kappa

		Prediction	
		Benign	Malignant
True	Benign	85(TP)	30(FN)
	Malignant	22(FP)	34(TN)

Table: SVM Confusion matrix

$$\kappa = \frac{0.6959 - 0.5434}{1 - 0.5434}$$

$$\kappa = 0.3340$$

Table of Contents

Papers

Wisconsin breast cancer dataset, looks at 9 features of the tumor itself, including radius, area, perimeter, texture

- **Sharma et al.**, 2017 [4]
- **Kaklamanis et al.**, 2019 [2]
- Chakradeo et al., 2019 [1]
- Saoud et al., 2019 [3]

k-Nearest Neighbors (KNN)

- KNN performs well on large datasets
- KNN benefits strongly from reducing dimensionality

Metric	k-nearest neighbor	support vector machine
Specificity	94.7%	84.9%
Sensitivity	90.09%	88.2%
Accuracy	93.06%	89.55%
AUC	92.39%	86.55%

Table: Results on Sharma et al. diagnostic dataset [4], 699 entries

Support Vector Machine (SVM)

- SVM is comparatively better then KNN on smaller datasets
- SVM deals with higher dimensions better than KNN

Metric	k-nearest neighbor	support vector machine
Specificity	61.2%	79.7%
Sensitivity	40.89%	41.2%
Accuracy	82.56%	89.73%
AUC	51.045%	60.45%

Table: Results on Sharma et al. prognostic dataset [4], 199 entries

Kaklamanis et al. results

Metric	k-nearest neighbor	support vector machine
Accuracy	96.49%	95.32%
Kappa	0.8145	0.8988

Table: Results from Kaklamanis et al. data [2]

Conclusion

Machine learning techniques can be used to improve current breast cancer diagnosis so patients can begin treatment as soon as possible

Questions

References I

K. Chakradeo, S. Vyawahare, and P. Pawar.
Breast cancer recurrence prediction using machine learning.
In *2019 IEEE Conference on Information and Communication Technology*. Institute of Electrical and Electronics Engineers, 2019.

M. M. Kaklamanis and M. E. Filippakis.
A comparative survey of machine learning classification algorithms for breast cancer detection.
In *Proceedings of the 23rd Pan-Hellenic Conference on Informatics*. Association for Computing Machinery, 2019.

H. Saoud, A. Ghadi, and M. Ghailani.
Proposed approach for breast cancer diagnosis using machine learning.
In *Proceedings of the 4th International Conference on Smart City Applications*. Association for Computing Machinery, 2019.

References II

A. Sharma, S. Kulshrestha, and S. Daniel.
Machine learning approaches for breast cancer diagnosis and prognosis.
In *2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp)*. Institute of Electrical and Electronics Engineers, 2017.