CSC384 Project: Breast Cancer Prediction based on Malignancy of Tumors

Armand Silviu Gurgu, CDF Account Name: c5gurgua
Pavel Litvinovich, CDF Account Name: c5litvin
Zeev Suprun, CDF Account Name: c5suprun

Type of Project: Bayes Net

Roles Performed by each member:
Zeev:

- Writing code to predict the class of test examples using Bayes net (major)
- Testing the accuracy of various Bayes net topologies (minor)
- Creating the correlation matrix (minor)

Pavel:

- Programming functions that construct joint and conditional probability factors (major)
- Determining how to use naive Bayes network topology to predict tumor class (major)

Armand:

- Determining how to use naive Bayes network topology to predict tumor class (major)
- Calculating information gain of different variables (minor)
- Graphing conditional probabilities (minor)

Everyone:

- Report writing (major)

Project Motivation

The correct prediction of breast cancer is very important for many people who have tumors, since failure to detect cancer at early stages can significantly complicate things later, maybe even leading to patient's death. Thus, it is very important to use all possible information to assist in making the correct prediction of tumor malignancy. Previous statistics about the malignancy of tumors given their cell features is an invaluable source of information for that purpose. This was the goal of the project: to use statistics to construct a model, that generates the probability of having cancer given the features of the tumor. These probability distributions were then used in constructing a Bayesian network.

Methods

The task of correctly predicting breast cancer given a patient's diagnosis was formulated as a Bayesian network problem where each factor in the network represents a particular cell feature from the dataset.

For this project we were given a dataset of 699 data entries (effectively 683 entries, since 16 data entries didn't have some attribute values, so they were not used in the analysis), each one with the patient's id, 9 tumor features (valued from 1 to 10) and the class (2 standing for benign, 4 for malignant tumor). The typical data entry looked like this:

[1041801,5,3,3,3,2,3,4,4,1,4]

where the first number is the sample id, the last is tumor class, and 9 values in between are tumor's characteristics in the following order(from left to right):

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

Using this data, we constructed a Bayes Net to predict the class of an example. For the prediction we used the following Bayes Net topology:

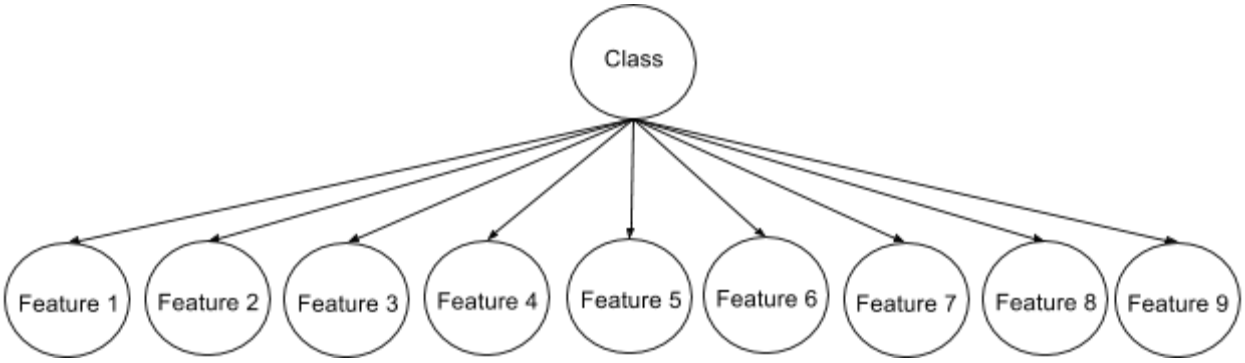

Figure 1: Naive Bayes Net Structure

The equation corresponding to this net is as follows:

$$Pr(Class = c|\ Feat_1\ =\ input_1,\ ...,\ Feat_9 = input_9) = Pr(Class = c) \prod_{i=1}^{9} Pr(Feat_i = input_i|Class\ =\ c)$$

Where $input_i$ is the input to the ith feature.

This is a network for Naive Bayes Classifier[1]. In this network there is at most one conditional dependence. The proposed network showed good results, as discussed in Evaluation and Results section.

The values of $Pr(Class)$ and $Pr(Feat_i|Class)$ for corresponding Conditional Probability Tables(CPTs) were calculated using the following formulas:

$$\Pr(Class = class) = \frac{\#\ of\ instances\ of\ Class = class}{Total\ \#\ of\ instances}$$

$$\Pr(Feat_i = val|Class = class) = \frac{\#\ of\ instances\ of\ Feat_i = val \wedge Class = class}{Total\ \#\ of\ instances\ of\ Class = class}$$

In order to distinguish the class of the tumor under consideration, the probability of it being malignant using the above net is calculated and then the probability of it being benign is calculated. The larger probability among them is taken as a distinguisher between the two classes. This can be summarized with the following formula:

$$Class^{prediction} = \underset{class}{\operatorname{argmax}}\, Pr(Class = class) \prod_{i=1}^{9} Pr\,(Feat_i =\ input_i|Class = class)$$

In addition to the Bayesian network described above other networks were constructed, which had the same structure, but include fewer features. These networks are: 9 networks, each with only a single feature used for classification; the original network with the feature least correlated with the class excluded (based on the results obtained from the 9 simple networks), and the network with only the 3 features most

correlated with the class included (again, based on the results of the testing of 9 simple Bayes nets). The performance of these nets was evaluated and compared to each other.

Evaluation and Results

The probability distributions for malignant tumors given each of the 9 cell features are shown in the figures in Appendix A.3. The graphs show that as the number of a particular cell feature goes up, the likelihood of getting cancer also increases sharply for all features. The graph for mitosis also suggests that mitosis is a bad indicator for cancer because when the number of mitosis is greater than or equal to 2, the likelihood of getting cancer is at least 0.77.

To assess the performance of the Bayesian network, some training examples were used to construct the Bayesian network and the remaining data examples were used to assess the performance of the Bayesian network. To assess the performance of the Bayesian network constructed, the % correctness classification metric was used:

$$\%\ Correctness = \frac{\#\ of\ examples\ classified\ correctly\ by\ the\ network}{total\ \#\ of\ examples\ used\ in\ the\ test\ set}$$

Another, more important metric was also used to evaluate the performance of the net: The F_1 score.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Where

$$Precision = \frac{tp}{tp + fp}$$
$$Recall = \frac{tp}{tp + fn}$$

Where *tp* is the number of true positives in the data, *fp* is the number of false positives, and *fn* is the number of false negatives.

The reason F1 score allows for a better metric than % correctness is that the majority of examples are benign tumours, so a classifier that always predicted benign would achieve a % correctness of 65%, even though it would be a bad classifier. However, the F_1 score of this classifier would be 0, which better reflects how well the classifier performs.

Evaluating the Effectiveness of Different Predictor Naive Bayes Nets:

The most simple naive Bayes net structure uses only a single variable to predict the class of the tumor.

For the following Bayes nets, approximately 70% of examples (478 examples) were used for training, and the remaining examples were used for the test set. This gives enough training examples to achieve good accuracy, while leaving enough test examples to verify the accuracy well.

Predictor Variable	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
fraction Correct	0.863	0.961	0.922	0.932	0.951	0.922	0.980	0.937	0.849
F1 Score	0.622	0.918	0.846	0.860	0.900	0.814	0.957	0.843	0.537

As can be seen from the above chart, even using a single variable to determine whether tumor is malignant or benign can provide quite good results. The best results come from using Bland Chromatin as a predictor variable, which yields an F_1 score of 0.957. Bland Chromatin also provided the highest information gain among all the features (see Appendix A.2). The variable that is the least useful for predicting the class is Mitosis, with an F_1 score of only 0.537 . This is unsurprising, since class has the smallest correlation with Mitoses than any other variable (see Appendix A.1), and Mitoses has the lowest information gain among all the features (see Appendix A.2).

Another Bayes net structure is one that uses all nine features in order to predict the class, as depicted in Figure 1. Using the same training and test set as above, using this Bayes net results in an F_1 score of 0.947, and an accuracy rate of 0.976. While this is quite accurate, it is less accurate than using only Bland Chromatin as an indicator variable.

One way to attempt to improve on the above Bayes net is to attempt excluding the Mitoses feature, since it has the lowest F_1 score. The resulting Bayes net had an F_1 score of 0.957, and an accuracy of 0.980. These results are now as good as using only Bland Chromatin as a predictor variable.

The final Bayes net used only the three variables that, individually, achieved an F1 score of at least 0.9: Uniformity of Cell Size, Single Epithelial Cell Size, and Bland Chromatin. This Bayes net yields an F_1 score of 0.968, and an accuracy of 0.985. This is the most accurate of all of the Bayes nets used.

Limitations and Obstacles

One of the major obstacles that we encountered was the relative lack of training data. For example, there were no data points that had a value of 9 for the mitosis feature, which limited the domain of values that the mitosis feature could take. This obstacle also meant that the factors for class given more than one cell feature, as follows, were impossible to construct:

$$Pr(Class = class \mid Feat_1 = input_1, Feat_2 = input_2)$$

This is because, for many combinations of feature values, there were no examples with that combination of values. This limited our choice in Bayes net topology, and was one of the factors that led to our choice to use naive Bayes nets.

Conclusions

This investigation allowed the team to explore the naive Bayes network topology and assess its performance on a real-world application. Several of the naive Bayes nets used performed very well, despite the assumption that the feature variables are independent of each other. One interesting finding was that some of the simple Bayes nets, which used only one feature variable, showed very high performance in some cases. Most notable of these was the Bayes net constructed using only the “bland chromatin” feature, which outperformed the “full” Bayes net which used all nine feature variables, and performed no worse than the Bayes net that excluded only the least relevant feature, Mitosis. The Bayes net that performed best used only three features to predict the class of the tumor: Uniformity of Cell Size, Single Epithelial Cell Size, and Bland Chromatin.

Even though the Bayes nets used performed well, getting more examples in the dataset would likely increase the accuracy. It is also possible that using Bayes network topologies that are more complex than the naive Bayes network would give better results.

References

[1] “Naive Bayes Classifiers” by Prof. Andrew W. Moore, Carnegie Mellon University (2004): http://www.autonlab.org/tutorials/naive02.pdf

Acknowledgements

Spoke to Sonya Allin about the naive Bayes classifier and information gain, and Erin Delisle about our implementation of the naive Bayes classifier and information gain

Appendix A

A.1: The correlation matrix

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
Clump Thickness	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Uniformity of Cell Size	0.64	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Uniformity of Cell Shape	0.65	0.91	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Marginal Adhesion	0.49	0.71	0.69	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Single Epithelial Cell Size	0.52	0.75	0.72	0.59	1.0	0.0	0.0	0.0	0.0	0.0
Bare Nuclei	0.59	0.69	0.71	0.67	0.59	1.0	0.0	0.0	0.0	0.0
Bland Chromatin	0.55	0.76	0.74	0.67	0.62	0.68	1.0	0.0	0.0	0.0
Normal Nucleoli	0.53	0.72	0.72	0.6	0.63	0.58	0.67	1.0	0.0	0.0
Mitoses	0.35	0.46	0.44	0.42	0.48	0.34	0.35	0.43	1.0	0.0
Class	0.71	0.82	0.82	0.71	0.69	0.82	0.76	0.72	0.42	1.0

A.2: Information Gain

Information Gain for Different Cell Features								
Bare Nuclei	Bland Chromatin	Clump Thickness	Marginal Adhesion	Mitoses	Normal Nucleoli	Single Epithelial Cell	Cell Shape	Cell Size
0.0179	0.5652	0.3908	0.3266	-0.5734	0.4793	0.3379	0.1978	-0.5480

A.3: Probability distributions of cancer given features: