CSE 469 Final Project: Identifying Cancerous Masses with Data Mining

Peter M. VanNostrand
University at Buffalo
pmvannos@buffalo.edu

1 Introduction

In medical science one increasingly common task is attempting to determine if a mass found within a patient is cancerous or benign. This can have large impacts on patient health as missing a cancer during screen may cause a patient to go untreated. Equally dangerously if a non-cancerous mass is identified as cancerous a patient may receive unnecessary chemotherapy or radiation treatment which could precipitate a decline in their health. Unfortunately, reviewing patient biopsies to determine if they are cancerous is a difficult task which requires a great deal of time from an expert. This makes manually reviewing samples a slow and costly process, particularly as many non-cancerous samples must be screened to find the less frequent malignant ones. Ideally, doctors would be able to filter out samples which are clearly not cancerous, and then review only the samples which could plausibly be cancer. This would save doctor's time which would reduce the cost of healthcare and allow doctors to treat more patients.

2 Formulation

The problem of cancer screen can be viewed as a data mining issue. Given many input samples from patient biopsies we would like to determine which are likely to be cancerous and which are not. This is essentially a two-class classification problem. Given the input features we need to find a mapping that generates a binary yes or no inference where yes indicates that a sample is cancerous and no indicates that a sample is benign. This can be accomplished using a logistic regression. Using a set of inputs labeled by experts we can train a logistic regression to map the input features to the correct binary label, then we use the trained regression on novel samples to predict whether they are likely to be cancerous.

3 Dataset

3.1 Source

For this exercise we will be using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. This dataset was obtained from the University of California Irvine Machine Learning Repository. The WDBC Dataset consists of 569 sample, each sample is generated from digitized image of a fine needle aspirate (FNA) of a breast mass.

3.2 Details

A FNA is a method for extracting an examining cells in which a small hollow tube is used to extract a sample of cells which are then dyed and viewed under a microscope. Each sample is composed of 32 attributes. The first two attribute is a unique image ID corresponding to the image that sample represents. The second attribute is a label, either "M" or "B" for malignant or benign. The remaining 30 attributes represent a set of 10 cellular characteristics calculated from automatic measurements of the cell. These characteristics listed below

Table 1: Cellular Characteristics

Radius
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Fractal Dimension

For each characteristic the mean, standard deviation, and maximum value are reported for that cell yielding the 30 features.

3.3 Statistics

Of the 569 total samples 357 are benign and 212 are malignant. As this dataset is relatively balanced between the possible labels we will not consider methods such as anomaly detection or sample weighting.

Examining the dataset further we take the average of the mean of each feature to determine their typical values. These values are shown below.

Table 2: Typical Characteristic Values

Characteristic	Mean
Radius	14.1273
Texture	19.2896
Perimeter	91.9690

Area	654.8891
Smoothness	0.0964
Compactness	0.1043
Concavity	0.0888
Concave Points	0.0489
Symmetry	0.1812
Fractal Dimension	0.0628

3.4 Preprocessing

As we can see in Table 2 the range of typical value is very broad, from very close to zero for the fractal dimension to several hundred for the total area. To speed up training and prevent overflow errors we will normalize all 30 features to the range $[0, 1]$

For convenience we will place the labels in a separate array and use 1.0 and 0.0 instead of "M" and "B" to represent the expert labeled result. We will also drop the unique image ID as it does not contain meaningful information.

From the total 569 samples we will use 80% (459) for training the logistic regression classifier, the remaining 20% (114) will be used to test the performance of the final classifier.

4 Algorithm

To perform the two-class classification of this dataset we will be using logistic regression. In logistic regression we compute the logit of the input features, that is we take a weighted sum of the input feature values and add a bias term. This generates a number which we pass through the sigmoid function to generate the probability that the cell in cancerous. Cells which are more like than not to be cancerous are identified as malignant, adjusting the cutoff for likelihood would allow for a controllable false positive to false negative ratio.

In our application the logistic classifier works as follows. Given the thirty input features $x_1 + x_2 + \cdots + x_{30}$ we first determine the log-odds as

$$\ell = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{30} x_{30}$$

Then the sigmoid function is used to convert this log-odds into the probability that the cell in cancerous p

$$p = sigmoid(\ell) = \frac{1}{1 + e^{-\ell}}$$

This probability ranges from [0,1] with 0 representing a certain non-cancerous prediction and a 1 representing a certain cancerous prediction. During inference to convert this to a classification we round to the nearest integer to produce the predicted class.

To train the logistic classifier we then compute the loss, that is the difference between the probability of being cancerous and the known class label of 0 or 1. We then compute the partial derivate of the loss with respect to the array of values β. This partial derivative is used to update the values of β to make the classifier more accurate. This repeats for many iterations as the classifier's accuracy converges.

5 Experiments

The logistic classifier described above was trained on 80% of the available dataset. Then using the remaining 20% we performed the two-class classification and compared the predicted class labels with the known labels.

5.1 Metrics

To measure the performance of our classifier we measure the number of true positives, true negatives, false positives, and false negatives and then compute the following metrics

Accuracy: The accuracy of the classification is essentially the percentage of samples which were correctly identified

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The precision represents the percentage of samples identified as cancerous that are in fact cancer, i.e. what percentage of all diagnosed cancer cases are correct

$$precision = \frac{TP}{TP + FP}$$

Recall: The recall represents the percentage of all cancerous cases were identified, i.e. what percentage of patients with cancer are being identified as having cancer

$$recall = \frac{TP}{TP + FN}$$

Ideally all three of these metrics would be the maximum possible value of one. In that case all cancer patients would be correctly identified, and no health individuals would be unnecessarily treated for cancer.

5.2 Results

Using a 0.5 likelihood cutoff (i.e. classifying all samples with a 50% or greater likelihood of being cancerous as malignant) I ran the logistic classifier and obtained the following results.

Table 3: Classifier Results Cutoff=0.5

Metric	Value
Accuracy	0.9646
Precision	0.9778
Recall	0.9362

We can see that overall the performance is decent, but probably not sufficient for such a high stakes application area. From the precision value see that almost all patients identified as having cancer do in fact have cancer, but unfortunately the lower recall value indicates that some cancer patients have been given a clean bill of health.

As the cancer positive results of this system are likely to be reviewed by an expert while the cancer negative results are not, we should prioritize the recall of our system over its precision. In other words, we should ensure that we do not clear any cancer patients as healthy, even if that means some healthy individuals are accidentally flagged as likely to have cancer. This will allow doctors to review the positive cancer classifications and apply their expert knowledge to cases which are uncertain.

To do this we can lower the cutoff likelihood. Using a cutoff of 0.4 (i.e. classifying all samples with a 40% or greater likelihood of being cancerous as malignant) obtained the following results.

Table 4: Classifier Results Cutoff=0.4

Metric	Value
Accuracy	0.9381
Precision	0.8600
Recall	1.0000

From these values we can see that the overall accuracy and precision have decreased, but the recall is significantly better. This allows the system to act as a filtering system to remove samples with low likelihood of being cancer. While this is an imperfect solution it is still capable of saving doctor's time and therefore reducing healthcare costs. With a larger dataset a more refined model may be possible.