Gaussian Discriminant Analysis (GDA) to Diagnose Malignant Breast Tumors

Matthew P. Burruss

May 2019

1 Introduction

Gaussian Discriminant Analysis (GDA) is a technique to build models to classify certain inputs. In this example, we will use inputs describing cell nuclei in a possibly cancerous tumor taken from breast tissue to categorize it as either malignant (cancerous) or benign. In a more general sense, GDA can be used when the input is real-valued, continuous and approximately normally distributed; although techniques can be used to apply GDA to real-valued discrete variables (i.e. Naive Bayes). By making strong assumptions about the inputs, GDA can be a powerful predictive tool because it is 'data efficient' meaning fewer samples are needed to train more parameters.

The following report covers: 1) the underlying mechanics of GDA with an introduction to the statistics involved, 2) a description of the tumor classification problem and the underlying dataset[1], 3) methods of model creation, 4) a discussion of the results. Furthermore, a Github link is provided that contains a program to execute and train the model.

2 Background of Problem

GDA is misnomer because it is actually a generative learning algorithm (not a discriminant). If you are wondering what the difference is then that's good because the names don't give away that much. The following paragraphs will introduce GDA, provide its mathematical definition, and then show how it works in the applied example.

A discriminant learning algorithm can be thought of as a way to "discriminate" between some input, let's say x and a finite number of outputs, say y, by passing the input to a function that returns the output. On the other hand, a generative algorithm "generates" models for each of the outputs. It then sees which model the input most likely fits by comparing the probability that this new input could exist in the various models that have been designed.

[1]Dataset: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

In mathematical terms, a discriminant algorithm will discover $p(y|x;\theta)$ where θ is the model parameters. Common examples of a discriminate algorithm are linear and logistic regressions. For example, a linear regression can be modelled as $\hat{y}^{(i)} = \theta_0 + x_1\theta_1 + ... + x_n\theta_n$. By calculating an estimate of θ_i that minimizes error on the training set, we can then predict possible outputs for a new input.

A generative algorithm on the other hand will find $\underset{y}{\text{argmax}}(p(x|y))$ which means the model y for $p(x|y)$ in which the input x most likely exists. You can think of y as a distribution spanning n dimensions where n is the size of the input vector or feature vector x. Using a set of training data, we create the model and then find the best fit for the data. Using Baye's Theorem, we can show that

$$\underset{y}{\text{argmax}}\, p(x|y) = \underset{y}{\text{argmax}}\, \frac{p(y|x)p(y)}{p(x)} = \underset{y}{\text{argmax}}\, p(y|x)p(y) \tag{1}$$

We are able to ignore the denominator because it is independent of the specific model y and therefore does not affect our maximum calculation.

To explain the concept in applied manner, the number of models can be thought of as the number of categories. In this case, we have two categories to represent the possible diagnosis of the patient's breast tissue: a benign tumor or a malignant tumor. Therefore we can represent our categories, $y^{(i)} \in (0, 1)$. The dataset includes 357 benign samples and 212 malignant samples. Our input $x_j^{(i)}$ represents the i'th patient with $j \in (0...29)$ features or parameters listed below. The features were measured from the sample tissue's cell nuclei.

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter squared / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension (coastline approximation - 1)

For each of the inputs described above, the first instance represents the mean, the second represents the standard deviation, and the third represents the "worst" occurrence of the feature. For example, $x_0^{(i)}$ is the i'th sample's mean cell nucleus radius, $x_{10}^{(i)}$ is the i'th sample's standard deviation cell nucleus radius, and $x_{20}^{(i)}$ is the i'th sample's largest cell nucleus radius.

3 Statistical Methods Used

GDA is based on two primary distributions: multivariate normal (MVN) distributions and bernoulli distributions. The multivariate distribution is the multidimensional form of the normal distribution. In the model, we make the strong assumption that $P(X|Y=k) \backsim MVN(\mu_k, \Sigma_k)$ and $P(Y=k) = Bernoulli(\phi_k)$ where μ_k and Σ_k are the population parameters of the k'th multivariate normal distribution estimated by $\hat{\mu_k}$ and $\hat{\Sigma_k}$ and ϕ_k is the population parameter of the Bernoulli trial of the k'th model estimated by $\hat{\phi_k}$.

In order to calculate our parameter estimates ϕ_i, μ_i, Σ_i which in our case we have 6 parameters to estimate (one for each category malignant or benign) we perform loglikelihood estimation.[2]

3.1 MVN Distribution

The probability density function (pdf) of the mutlivariate distribution is given below:

$$p(x|y=i) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{(1/2)}} e^{-\frac{1}{2}(x-\mu_i)^T(\Sigma^{-1})(x-\mu_i)} \tag{2}$$

The distance in the exponent represents the Mahalanobis distance which means the distance from the vector x to the model's distribution (which in our case is estimated from the training input). This distance can be thought of as a rotated and scaled Euclidean distance extended to multiple dimensions and therefore gives us an indication of how many standard deviations our input is from the model (e.g how similar a new patient's tissue sample compares to a model of benign tissue).

For the MVN, the parameters include the vector of the means μ_i and the covariance matrix Σ_i which is positive, semi-definite. The maximum likelihood estimators (MLE) for MVN is shown below:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \{Y^{(i)} == k\} x^{(i)}}{\sum_{i=1}^{N} \{Y^{(i)} == k\}} \tag{3}$$

$$\hat{\Sigma}_k = \frac{1}{\sum_{i=1}^{N} \{Y^{(i)} == k\} - 1} \sum_{i=1}^{N} \{Y^{(i)} == k\}(x^{(i)} - \hat{mu}_i)(x^{(i)} - \hat{mu}_i)^T \tag{4}$$

The equations above can be interpreted as followed. The estimate of the mean vector is the sum of all the input vectors $x^{(i)}$ whose output was equal to k divided by the total number of occurrences of output i. The covariance matrix for each model is calculated in a similar manner and represents the shape and rotation of the distribution. Figure 1 is shows the estimated models for benign and malignant tumor cells given two parameters as well as well the training sample falls along the 2D contour. The MLEs described above provide unbiased estimators of the population parameters.

[2]The covariance Σ can be calculated as a shared matrix so that each distribution has the same shape.

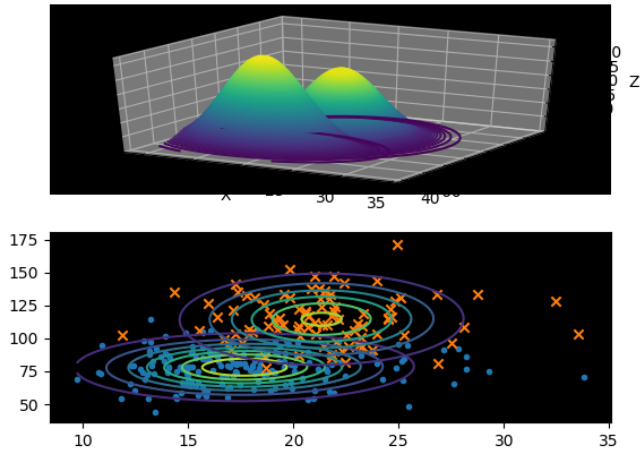

Figure 1: The two fitted model MVN distributions showing the effects of the mean radius of the nuclei and the mean texture of the nuclei in the sample tissue. The image on top shows the pdf of the MVN created by the sample input and the image on the bottom shows the contours of the distribution as well as the category of the tumor (x = malignant and circle = benign)

3.2 Bernoulli Distribution

The Bernoulli distribution represents an independent trial where all we are concerned with is the probability of success or failure. It is parameterized by ϕ_i where ϕ_i is the probability of a success and $1 - \phi_i$ is the probability of failure. The MLE provides an unbiased estimator of ϕ_i and is shown below

$$\hat{\phi}_k = \frac{1}{N} \sum_{i=1}^{N} \{Y^{(i)} == k\} \tag{5}$$

In the context of the problem, we use the Bernoulli trial to give us an indication of the prevalence of a category occurring compared to all the other categories in our training sample ($p(y = k)$).

4 Methods of Model Training

Using the MLEs described above, we can generate models for benign and malignant tissue samples based on the training data. We then hope that given a new input x we can solve $\underset{y}{\text{argmax}}(p(x|y))$ and categorize the input in, in our case, as cancerous or not cancerous in an accurate fashion.

To ensure the reliability of the model–that is that its performance on a test set can mimic its performance on a real patient–I have taken precaution to split the dataset into three parts: training, testing, and validation which correspond to respectively 60%, 20%, and 20% of the entire dataset ($N = 569$, $N_{train} = 341$, $N_{test} = 114$, $N_{validation} = 114$).

The method was as followed. Randomly shuffle the training and testing dataset for 100 iterations and calculate a model MLEs based on the training dataset. Calculate the overall accuracy of the predictions by summing the number of correct predictions in testing set over the size of the testing set. Store the model that performed the best during the iterations and then evaluate the performance of the model on the validation dataset. In doing so, the model avoids a common error in ML which is shuffling a single dataset and iterating to find the best model. This error makes the user aware of the validation dataset and therefore the model seems to be better than it actually is.

5 Results

Using the methods described above, the model was then evaluated on a validation data set ($N = 114$, $Benign\ (B) = 70$, $Malignant\ (M) = 44$). After predicting the classification of the tissue samples, sensitivity (identifying cancerous tissue as malignant) and specificity (identifying non-cancerous tissue as benign) analysis on the results was conducted to have an indication of the expected prevalence of Type I and Type II errors.

	Disease				
Test	**Present**	**n**	**Absent**	**n**	**Total**
Positive	True Positive	a=42	False Positive	c=1	a + c = **43**
Negative	False Negative	b=2	True Negative	d=69	b + d = **71**
Total		a + b = **44**		c + d = **70**	

Figure 2: The sensitivity and specificity analysis table.

Figure 2 shows the table used to calculate the sensitivity and specificity of the test and figure 3 shows the derived probabilities. A 95% confidence interval for each of the values listed below were calculated from the sensitivity/specificity test. Below is also a description of what these values indicate [3].

- Sensitivity: probability that a test result will be positive when the disease is present (true positive rate).
- Specificity: probability that a test result will be negative when the disease is not present (true negative rate).

[3] Calculations and descriptions from: https://www.medcalc.org/calc/diagnostic$_test.php$

Statistic	Formula	Value	95% CI
Sensitivity	$\frac{a}{a+b}$	95.45%	84.53% to 99.44%
Specificity	$\frac{d}{c+d}$	98.57 %	92.30% to 99.96%
Positive Likelihood Ratio	$\frac{Sensitivity}{1-Specificity}$	66.82	9.53 to 468.25
Negative Likelihood Ratio	$\frac{1-Sensitivity}{Specificity}$	0.05	0.01 to 0.18
Disease prevalence	$\frac{a+b}{a+b+c+d}$	38.60% (*)	29.63% to 48.17%
Positive Predictive Value	$\frac{a}{a+c}$	97.67% (*)	85.70% to 99.66%
Negative Predictive Value	$\frac{d}{b+d}$	97.18 % (*)	89.90% to 99.26%
Accuracy	$\frac{a+d}{a+b+c+d}$	97.37% (*)	92.50% to 99.45%

Figure 3: The sensitivity and specificity analysis table.

- Positive likelihood ratio: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease.
- Negative likelihood ratio: ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease.
- Positive predictive value: probability that the disease is present when the test is positive.
- Negative predictive value: probability that the disease is not present when the test is negative.
- Accuracy: overall probability that a patient will be correctly classified.