Lecture 8: Classification I - Naive Bayes
Modeling Social Data, Spring 2017
Columbia University

Keerti Agrawal

March 10, 2017

1 Introduction

1.1 What is Classification?

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

1.1.1 Mathematical Definition

Input: As with regression, in a classification problem we start with measurements x_1, x_2, . . , x_n in an input space X.
Output: The discrete output space Y is composed of K possible classes:
1. $Y = \{1, +1\}$ or $\{0, 1\}$ is called binary classification.
2. $Y = \{1, . . . , K\}$ is called multiclass classification.
Instead of a real-valued response, classification assigns x to a category. For pair (x, y), y is the class of x.

1.1.2 Defining a Classifier

Classification uses a function f (called a classifier) to map input x to class y.

$$y = f(x)$$

2 Naive Bayes Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features.

2.1 Assumption

All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

2.2 Bayes Theorem

A theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause.

Formula:

$$P(\theta|D) = P(\theta)\frac{P(D|\theta)}{P(D)}, \tag{1}$$

2.2.1 Disease Example

Consider a hypothetical population of 10,000 people.

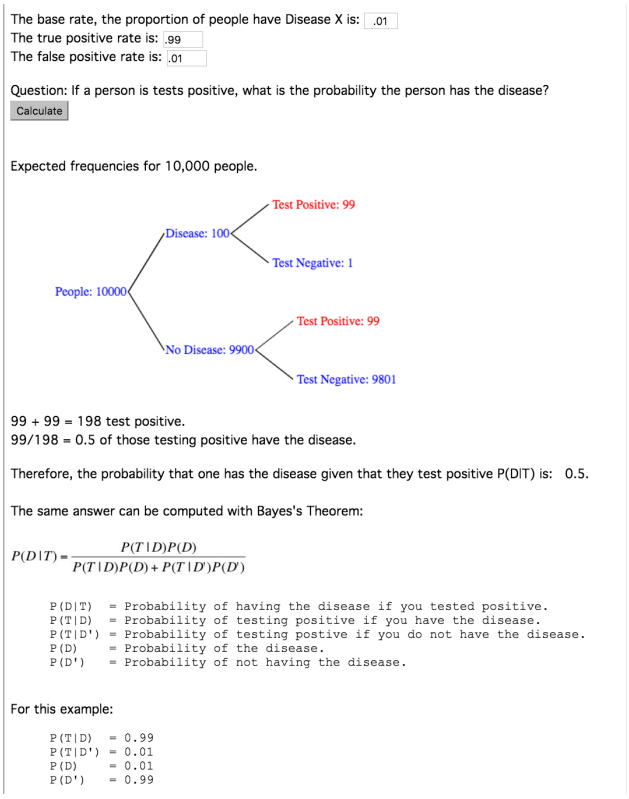

Figure 1: Bayes Theorem Example (Source: http://onlinestatbook.com/2/probability/bayes_demo.html)

2.3 Maximum Likelihood Estimate

The probability of observing the data set in a class C, given parameters θ: (iid assumption)

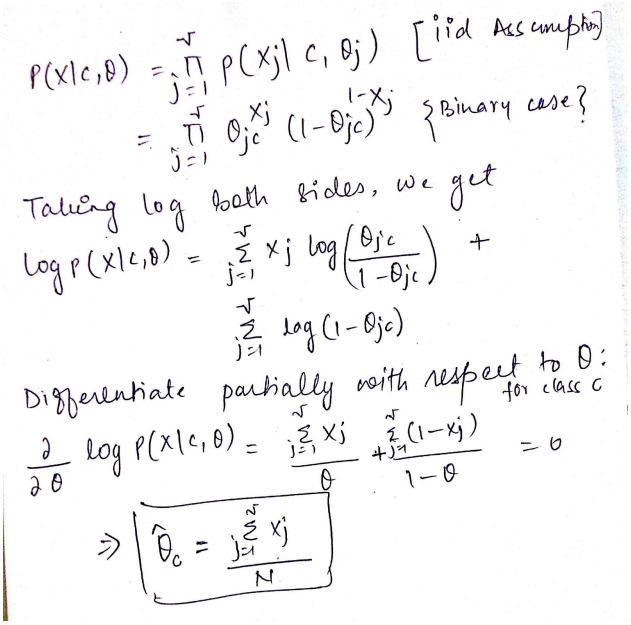

Figure 2: MLE for Naive Bayes

2.4 Advantages & Disadvantages of Naive Bayes

2.4.1 Advantages

- Easy to implement
- Requires a small amount of training data to estimate the parameters
- Good results obtained in most of the cases

2.4.2 Disadvantages

- Assumptions: class conditional independance, therefore loss of accuracy
- Practically, dependencies exist among variables
- Zero conditional probability problem

2.5 Zero conditional probability problem explained

- If a given class and feature value never occur together in the training set then the frequency based probability estimate will be zero.
- This is problematic since it will wipe out all information in the other probabilities when they are multiplied.
- It is therefore often desirable to incorporate a small sample correction in all probability estimates such that no probability is ever set to be exactly zero.
- Laplace smoothing could be the one solution to eliminate this problem.

3 Logistic Regression

Logistic Regression removes the over-fitting by Naive Bayes for zero prior case. It doesn't assume the feature vectors to be uncorrelated.

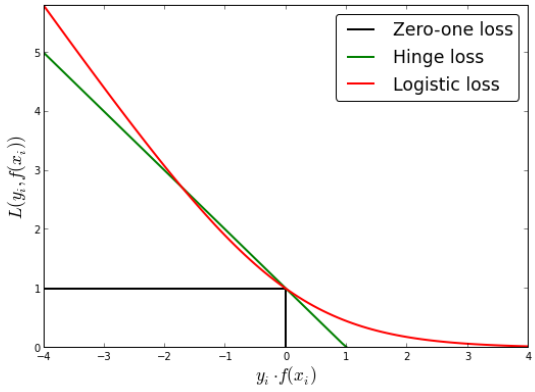

Figure 3: Loss functions

Binary Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables, which can be discrete and/or continuous. The important point here to note is that in linear regression, the expected values of the response variable are modeled based on combination of values taken by the predictors. In logistic regression Probability or Odds of the response taking a particular value is modeled based on combination of values taken by the predictors. Like regression (and unlike log-linear models that we will see later), we make an explicit distinction between a response variable and one or more predictor (explanatory) variables.

3.0.1 Log Odds Ratio

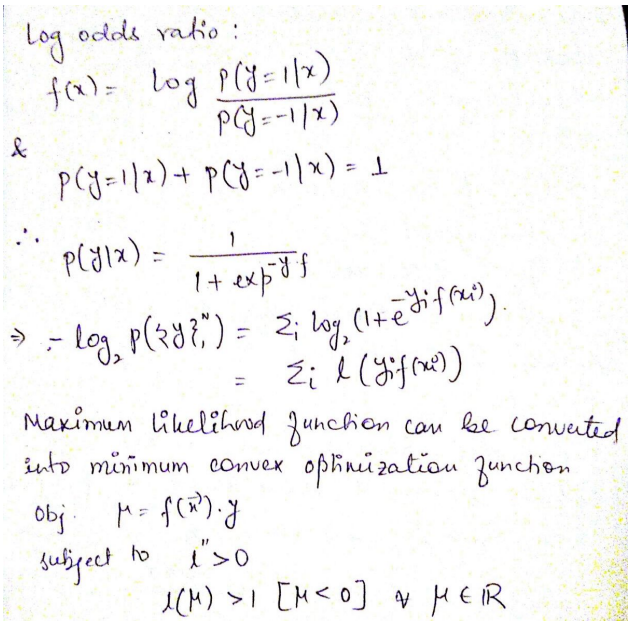

Figure 4: Log Odds Ratio: Logit Function

Despite the probabilistic framework of logistic regression, all that logistic regression assumes is that there is one smooth linear decision boundary. It finds that linear decision boundary by making assumptions that the P(Y—X) of some form, like the inverse logit function applied to a weighted sum of our features. Then it finds the weights by a maximum likelihood approach. The decision boundary it creates is a linear decision boundary that can be of any direction.

3.1 Advantages & Disadvantages of Logistic Regression

3.1.1 Advantages

- Convenient probability scores for observations.
- Multi-collinearity is not really an issue and can be countered with L2 regularization to an extent.

3.1.2 Disadvantages

- Doesn't perform well when feature space is too large.
- Doesn't handle large number of categorical features/variables well.
- Using MLE for parameter might not give closed form solution, therefore use iterative algorithms like Gradient descent (Boosting).

3.2 Boosting

While boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight.