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

Vedant Dharnidharka

March 10, 2017

1 Classification

Classification problem deals with categorization of given examples into a set of categories.

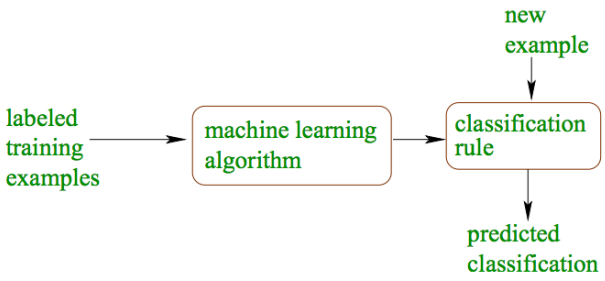

Figure 1: Classification

Input: As with regression, in a classification problem we start with measurements x_1, x_2, . . , x_n in an input space X.
Output:
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.

2 Naive Bayes Classifier

The Naive Bayes classifier is based on the Bayes rule of conditional probability. It makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other. In some cases it is also seen that Naive Bayes outperforms many other comparatively complex algorithms.

2.1 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.

Let (Ω,P) be a probability space. (Ω is the sample space; P is the probability distribution.)
For any event A,B in Ω
Formula:

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}, \tag{1}$$

Let E, H_0, H_1 in Ω.
Conditioned on E, of H_0 and H_1, we find which one is more probable.
Compare $P(H_0)P(E|H_0)$ to $P(H_1)P(E|H_1)$

2.1.1 Example

Suppose result of test for genetic disease is correct with probability 95%, and suppose the disease is rare: any given person has disease with probability 1%.

Question: If test comes back positive for disease, is it more likely that you have disease or do not?

$$\begin{aligned} E &= \text{test comes back positive for disease} \\ H_0 &= \text{do not have disease} \\ H_1 &= \text{have disease} \\ \mathbb{P}(E \mid H_0) &= 0.05 \\ \mathbb{P}(E \mid H_1) &= 0.95 \\ \mathbb{P}(H_0) &= 0.99 \\ \mathbb{P}(H_1) &= 0.01 \end{aligned}$$

Want to compare $\mathbb{P}(H_0 \mid E)$ to $\mathbb{P}(H_1 \mid E)$, so compare

$$\mathbb{P}(H_0) \cdot \mathbb{P}(E \mid H_0) = 0.99 \cdot \underline{0.05} \quad \text{and} \quad \mathbb{P}(H_1) \cdot \mathbb{P}(E \mid H_1) = \underline{0.01} \cdot 0.95 .$$

Figure 2: Bayes Theorem Example(Source: http://www.cs.columbia.edu/ djhsu/coms4771-f16/lectures/slides-generative.4up.pdf)

2.2 Maximum Likelihood Estimate

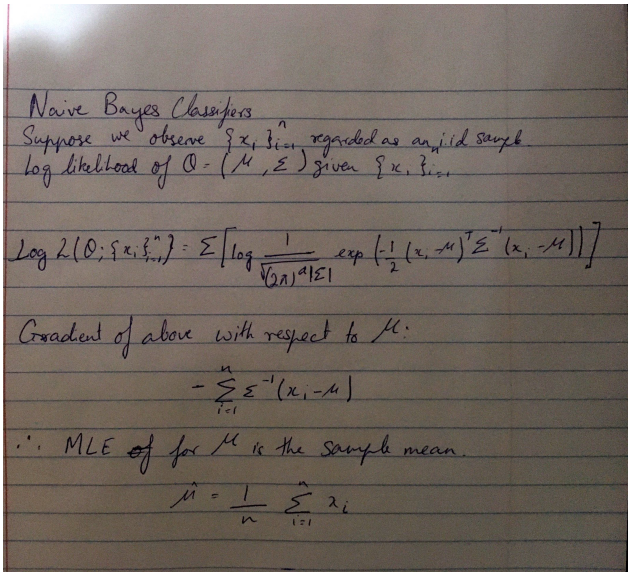

Figure 3: MLE for Naive Bayes

2.2.1 Advantages

- Simple, many variations
- Can leverage domain knowledge of class conditonals
- Can be very efficient when K is large.

2.2.2 Disadvantages

- Classifier relies on formula (via Bayesŕule) that assumes data comes from estimated distribution, which is generally not true.
- Modeling P away from decision boundary between classes is a wasted effort.

3 Logistic Regression

In logistic regression probability of the response taking a particular value is modeled based on combination of values taken by the predictors.

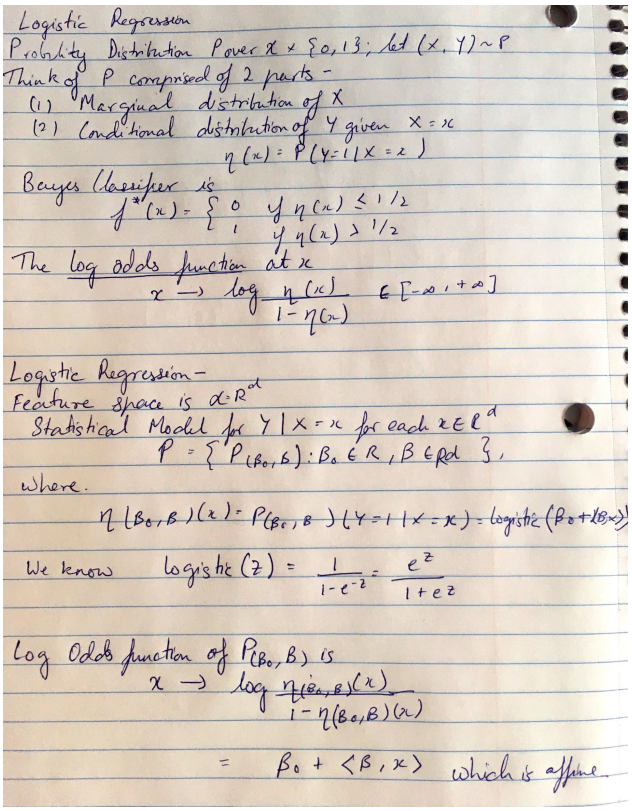

Figure 4: Logistic Regression-Linear Classifier

3.1 Boosting

Boosting is nothing but using a learning algorithm that provides rough rules-of-thumb to construct a very accurate predictor.

Motivation: Easy to construct classification rules that are correct more-often-than-not (e.g., If over 5 percent of the

e-mail characters are dollar signs, then its spam.), but seems hard to find a single rule that is almost always correct.

Assumption: Availability of a base or weak learning algorithm which produces a weak classifier. Boosting improves the performance of the weak learning algorithm while treating it as a black box . Weak classifiers are not entirely trivial which means that the error rates are at least a better than a classifier whose every prediction is a random guess. The weak classifiers can be moderately inaccurate, but not as bad as random guessing.

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 assigned weights again. Therefore when an entry is wrongly classified, the classifier which correctly classifies has its weight increased.