Rudimentary Naive Bayes Classifier In Haskell

Dhruv Rajan

May 25, 2017

1 Implementation

- The implementation had two main parts: transformations on data representations, and probabilistic inference
- **Data Transformations**
  - Involved developing "summary" statistics representations
  - Input: a list of n-vectors `[[Float]]`
  - Final representation: summary statistics of features (μ, σ) per-feature, per-class `[(Int, [(Float, Float)])]`
  - Creating these "summary" statistics comprises "training" the model
  - This could be done much more cleanly, in an "update" manner for individual entries, as opposed the current pipeline of inefficient transformations, and with more emphasis on training
  - Additionally, can explore exploiting datatypes and laziness
- **Probabilistic Inference**
  - Predict classification of new vectors using the Naive Bayes Algorithm.
  - Must calculate conditional probabilities for every feature - would be better to utilize the Distribution abstraction
  - Available libraries seem to only support manipulation of discrete distributions, and do not allow the representation of continuous distributions (gaussian, exponential, etc.)
  - Areas to look at: continuous distribution abstraction support, utilization of distribution abstraction, other inference algorithms, modular sub-algorithms or standard distribution transformations that may be common between other machine learning classification algorithms

2 Naive Bayes Algorithm

Naive Bayes is a classification algorithm which utilizes Baye's Rule:

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

Each entry is a vector, $\langle 1, 3, 7, ... \rangle$, classified by an integer category $0, 1, 2, 3....$ The classifier is a function which maps a vector to a category: $clf :: \chi \rightarrow C$. The insight is that learning this is equivalent to learning the conditional probability $P(C|X)$ for arbitrary C, X.

What can be approximated:

$$P(C_k) \tag{2}$$
$$P(X_i|C_k) \tag{3}$$
$$P(X_1, X_2, ..X_n|C_k) = \Pi_{i=1}^{n} P(X_i|C_k) \tag{4}$$

$$\chi = X_1, X_2, X_3...X_n \tag{5}$$
$$P(C_k|\chi) = P(C_k|X_1, X_2, ...X_n) \tag{6}$$
$$= \frac{P(X_1, X_2, ...X_n|C_k) \cdot P(C)}{P(X_1, X_2, ...X_n)} \tag{7}$$