A Note on an Elementary Formulation of the Naive Bayesian Model

jiyucho9145

February 9, 2019

Abstract

Some mail server programs and some mail client programs classify messages into several categories by the naive Bayesian classifier automatically. The algorithm of the naive Bayesian classifier is based on the naive Bayesian model, which satisfies the two specific assumptions.

In this note, we define a probabilistic model (message receiving model) without the two assumptions and show that the message receiving model coincides with the naive Bayesian model under the two assumptions.

Contents

1 Introduction

2 Message Receiving Model

2.1 Message Receiving Model

Definition 2.1. Let W be an nonempty finite set, $E = \mathrm{Map}(W, \{0,1\}), C = \{c_1, c_2), c_1 \neq c_2$, then we call (W, E, C) a message receiving model in this note.

Definition 2.2. Let $g : F \to C, F \subset E, F \neq \emptyset$, then we call $g : F \to C$ a training data for a message receiving model (W, E, C) in this note.

Definition 2.3. Let $e \in F$, then we define a support $\mathrm{Supp}_g(e)$ of e as follows:

$$\mathrm{Supp}_g(e) = \{w \in W; e(w) \neq 0\}. \tag{1}$$

Definition 2.4. Let $Z \subset W$, then we define a set $M_g(Z)$ as follows:

$$M_g(Z) = \{e \in F; Z \subset \mathrm{Supp}_g(e)\}. \tag{2}$$

Definition 2.5. Let $w \in W$, then we define a set $M_g(w)$ as follows:

$$M_g(w) = M_g(\{w\}). \tag{3}$$

2.2 Message Receiving Measures

Definition 2.6. We call the map

$$m_g : \mathrm{Pow}(F) \to \mathbb{R}; G \mapsto \mathrm{card}(G) \tag{4}$$

a message receiving measure of $g : F \to C$ in this note. Where $\mathrm{Pow}(S)$ denots the power set of S for arbitrary set S.

Proposition 2.1. $m_g : \mathrm{Pow}(F) \to \mathbb{R}$ is a measure on measurable space $(F, \mathrm{Pow}(F))$.

Proposition 2.2. Let $C_{g,i} = g^{-1}(c_i)$, then

$$\mathrm{card}(C_{g,1}) + \mathrm{card}(C_{g,2}) = \mathrm{card}(F). \tag{5}$$

2.3 Message Receiving Probabilities

Definition 2.7. We call the map

$$P_g : \mathrm{Pow}(F) \to \mathbb{R}; G \mapsto m(G)/m(F) \tag{6}$$

a message receiving probability of $g : F \to C$ in this note.

Proposition 2.3. $P_g : \mathrm{Pow}(F) \to \mathbb{R}$ is a probability (measure) on measurable space $(F, \mathrm{Pow}(F))$.

Proposition 2.4.

$$P_g(C_{g,1}) + P_g(C_{g,2}) = 1. \tag{7}$$

2.4 Message Receiving Conditional Probabilities

Definition 2.8. We call the conditional probability

$$P_g(-|-) : \mathrm{Pow}(F) \times \mathrm{Pow}(F) \to \mathbb{R} \tag{8}$$

a message receiving conditional probability of $g : F \to C$ in this note.

Proposition 2.5. For arbitrary $M \subset F$, following equation is satisfied:

$$P_g(C_{g,1}|M) + P_g(C_{g,2}|M) = 1. \tag{9}$$

Proposition 2.6. For arbitrary $M \subset F$, following equation is satisfied:

$$P_g(C_{g,1}|M) = \frac{P_g(M|C_{g,1})P(C_{g,1})}{P_g(M|C_{g,1})P(C_{g,1}) + P_g(M|C_{g,2})P(C_{g,2})}. \tag{10}$$

3 Comparation with Naive Bayesian Model

3.1 Calculationg Conditional Probabilities under the Two Specific Assumptions

Theorem 3.1. Assume that $g : F \to C$ satisfies following two conditions: (i) $P_g(C_{g,1}) = P_g(C_{g,2}) = 1/2$; (ii) for arbitrary $w_1, w_2, \ldots, w_r \in W$,

$$P_g(M_g(\{w_1, w_2, \ldots, w_r\})) = P_g(M_g(w_1))P_g(M_g(w_2)) \cdots P_g(M_g(w_r)). \tag{11}$$

i.e, $M_g(w_1), M_g(w_2), \ldots, M_g(w_r)$ are independent. And, let $Q_g(w_1, w_2, \ldots, w_r), R_g(w_1, w_2, \ldots, w_r)$ be following functions:

$$Q_g(w_1, w_2, \ldots, w_r) = \prod_i^r (P_g(M(w_i)|C_{g,1})P_g(C_{g,1})), \tag{12}$$

$$R_g(w_1, w_2, \ldots, w_r) = \prod_i^r (\frac{P_g(M(w_i)) - P_g(M(w_i)|C_{g,1})P_g(C_{g,1})}{1 - P_g(C_{g,1})}) \tag{13}$$

Then, the following equation is satisfied

$$P_g(C_{g,1}|M_g(\{w_1, w_2, \ldots, w_r\})) = \frac{Q_g(w_1, w_2, \ldots, w_r)}{Q_g(w_1, w_2, \ldots, w_r) + R_g(w_1, w_2, \ldots, w_r)} \tag{14}$$