Naïve Bayes Exercise

CSCI 374 Oberlin College Fall 2017
October 25, 2017

Introduction

For this exercise, we will be using a small baking recipe data set posted to the class webpage:
http://cs.oberlin.edu/~aeck/Fall2017/CSCI374/Handouts/CSCI374_BakingExample_Instances.xlsx

Our goal is to predict the type of baked good represented by a recipe based on its ingredient list.

The attributes in the data set include: $A = \{Main\ Grain, Secondary\ Grain, Rising\ Agent, Shortening, Shortening\ Amount, Sweetener, Milk, Eggs, Additional\}$

and the possible labels include: $C = \{Bread, Pastry, Desert\}$

Calculating $P(C_i)$

To calculate the prior probability of each label, please fill in this table.

As a reminder:

$$P(C_i) = \frac{n_{C_i}}{|X|}$$

Label C_i	Count	$P(C_i)$
Bread		
Pastry		
Desert		
Total		

Calculating $P(A_j = x_j \mid C_i)$

To calculate the probability that an attribute has a given value if the instance has a given label, fill in the following tables. Please use pseudocounts of $a = 1$ and $b =$ the number of values that the attribute can have.

As a reminder:

$$P(A_j = x_j \mid C_i) = \frac{n_{x_j, C_i} + a}{n_{C_i} + b}$$

Tables for $C_i = Bread$

Attribute = Main Grain

Attribute Value x_i	Count	With Pseudocount	$P(MG = x_i \mid Bread)$
Bread Flour			
All Purpose Flour			
Cake Flour			
Total			

Attribute = Secondary Grain

Attribute Value x_i	Count	With Pseudocount	$P(SG = x_i \mid Bread)$
Cocoa Powder			
Whole Wheat Flour			
None			
Total			

Attribute = Rising Agent

Attribute Value x_i	Count	With Pseudocount	$P(RA = x_i \mid Bread)$
Yeast			
Baking Soda			
Total			

Attribute = Shortening

Attribute Value x_i	Count	With Pseudocount	$P(S = x_i \mid Bread)$
Butter			
Oil			
Crisco			
None			
Total			

Attribute = Shortening Amount

Attribute Value x_i	Count	With Pseudocount	$P(SA = x_i \mid Bread)$
Large			
Medium			
Small			
None			
Total			

Attribute = Sweetener

Attribute Value x_i	Count	With Pseudocount	$P(Sw = x_i \mid Bread)$
Sugar			
Honey			
None			
Total			

Attribute = Milk

Attribute Value x_i	Count	With Pseudocount	$P(M = x_i \mid Bread)$
Yes			
No			
Total			

Attribute = Eggs

Attribute Value x_i	Count	With Pseudocount	$P(E = x_i \mid Bread)$
Yes			
No			
Total			

Attribute = Additional Ingredients

Attribute Value x_i	Count	With Pseudocount	$P(A = x_i \mid Bread)$
Chocolate			
Cinnamon			
Fruit			
Malt			
None			
Total			

Tables for $C_i = Pastry$

Attribute = Main Grain

Attribute Value x_i	Count	With Pseudocount	$P(MG = x_i \mid Pastry)$
Bread Flour			
All Purpose Flour			
Cake Flour			
Total			

Attribute = Secondary Grain

Attribute Value x_i	Count	With Pseudocount	$P(SG = x_i \mid Pastry)$
Cocoa Powder			
Whole Wheat Flour			
None			
Total			

Attribute = Rising Agent

Attribute Value x_i	Count	With Pseudocount	$P(RA = x_i \mid Pastry)$
Yeast			
Baking Soda			
Total			

Attribute = Shortening

Attribute Value x_i	Count	With Pseudocount	$P(S = x_i \mid Pastry)$
Butter			
Oil			
Crisco			
None			
Total			

Attribute = Shortening Amount

Attribute Value x_i	Count	With Pseudocount	$P(SA = x_i \mid Pastry)$
Large			
Medium			
Small			
None			
Total			

Attribute = Sweetener

Attribute Value x_i	Count	With Pseudocount	$P(Sw = x_i \mid Pastry)$
Sugar			
Honey			
None			
Total			

Attribute = Milk

Attribute Value x_i	Count	With Pseudocount	$P(M = x_i \mid Pastry)$
Yes			
No			
Total			

Attribute = Eggs

Attribute Value x_i	Count	With Pseudocount	$P(E = x_i \mid Pastry)$
Yes			
No			
Total			

Attribute = Additional Ingredients

Attribute Value x_i	Count	With Pseudocount	$P(A = x_i \mid Pastry)$
Chocolate			
Cinnamon			
Fruit			
Malt			
None			
Total			

Tables for $C_i = Desert$

Attribute = Main Grain

Attribute Value x_i	Count	With Pseudocount	$P(MG = x_i \mid Desert)$
Bread Flour			
All Purpose Flour			
Cake Flour			
Total			

Attribute = Secondary Grain

Attribute Value x_i	Count	With Pseudocount	$P(SG = x_i \mid Desert)$
Cocoa Powder			
Whole Wheat Flour			
None			
Total			

Attribute = Rising Agent

Attribute Value x_i	Count	With Pseudocount	$P(RA = x_i \mid Desert)$
Yeast			
Baking Soda			
Total			

Attribute = Shortening

Attribute Value x_i	Count	With Pseudocount	$P(S = x_i \mid Desert)$
Butter			
Oil			
Crisco			
None			
Total			

Attribute = Shortening Amount

Attribute Value x_i	Count	With Pseudocount	$P(SA = x_i \mid Desert)$
Large			
Medium			
Small			
None			
Total			

Attribute = Sweetener

Attribute Value x_i	Count	With Pseudocount	$P(Sw = x_i \mid Desert)$
Sugar			
Honey			
None			
Total			

Attribute = Milk

Attribute Value x_i	Count	With Pseudocount	$P(M = x_i \mid Desert)$
Yes			
No			
Total			

Attribute = Eggs

Attribute Value x_i	Count	With Pseudocount	$P(E = x_i \mid Desert)$
Yes			
No			
Total			

Attribute = Additional Ingredients

Attribute Value x_i	Count	With Pseudocount	$P(A = x_i \mid Desert)$
Chocolate			
Cinnamon			
Fruit			
Malt			
None			
Total			

Predicting Baked Good Type

Calculate the most probable label for each of the following instances.

As a reminder:

$$P(C_i|\boldsymbol{x}) \propto P(C_i) \prod_{j=1}^{|A|} P\left(A_j = x_j \,\middle|\, C_i\right)$$

1. $x^1 = \{Bread\ Flour, None, Yeast, Butter, Large, Sugar, Yes, Yes, None\}$

2. $x^2 = \{All\ Purpose\ Flour, Whole\ Wheat\ Flour, Yeast, None, None, None, No, No, None\}$

3. $x^3 = \{Bread\ Flour, None, Yeast, Butter, Large, Sugar, Yes, Yes, Chocolate\}$

4. $x^4 = \{All\ Purpose\ Flour, None, Baking\ Soda, Oil, Medium, Sugar, No, Yes, Chocolate\}$

5. $x^5 = \{All\ Purpose\ Flour, None, Baking\ Soda, Butter, Large, Sugar, No, Yes, None\}$

Assume that the correct labels were:

$$y^1 = Bread, \quad y^2 = Bread, \ y^3 = Pastry, \ y^4 = Pastry, \ y^5 = Desert$$

What was your predictive accuracy?