Algoritmo de Naïve Bayes para la predicción de Depósitos a Plazo de un cliente en un Banco

German López Rodrigo

I. Resumen

El presente artículo pretende mostrar al lector de cualquier área, las bondades de utilizar el algoritmo de Naïve Bayes para la predicción de datos. Para ello se recurre al desarrollo de un caso aplicado al desempeño de marketing de una institución bancaria la cual desea pronosticar si un cliente suscribirá un depósito a plazo con la meta de diseñar una campaña de marketing más particular.

II. Palabras clave

Naïve Bayes, Aprendizaje Automático, Marketing, Algoritmo predictivo.

III. Descripción del problema

Hoy en día, dirigirse a la audiencia adecuada para una campaña de marketing puede ahorrarle a una empresa miles de dólares, si se lleva a cabo en la dirección correcta. Para ello, las empresas deben de optar por implementar algoritmos y técnicas que sean capaces extraer reglas que puedan ayudar en el marketing objetivo, con la finalidad de obtener una predicción de los clientes interesados en el producto que ofrece la empresa.

- **Pregunta de investigación**

  ¿Qué información registrada de las personas en el sector bancario se puede utilizar para predecir si suscribirá un depósito a plazos?

- **Objetivo**

  Determinar un modelo probabilístico que permita predecir si un cliente de un banco suscribirá un depósito a plazo mediante la implementación del algoritmo de Naïve Bayes.

- **Fuente de Datos**

  Bank [1] es una fuente de datos que consta de 41188 observaciones y 20 características, de las cuales 10 son numéricas y 10 son características nominales. Esta fuente de datos es la que será utilizada para cumplir con el objetivo principal del presente artículo de investigación. Cabe aclarar que la fuente de datos fue dividida en dos conjuntos: un conjunto de 40000 observaciones para construir el modelo de Naïve Bayes y otro conjuntó de 1188 observaciones para realizar las pruebas del modelo de Naïve Bayes.

IV. Introducción

Conforme pasa el tiempo, la tecnología avanza cada vez más y junto con esta, el ser humano busca satisfacer sus diferentes necesidades aplicándola en diferentes áreas de interés. Una de estas grandes aplicaciones, sin duda alguna, es la de poder hacer predicciones o pronósticos de eventos futuros de acuerdo al comportamiento observable de un conjunto de mediciones, observaciones o experimentos pasados. Por ejemplo, en el ámbito financiero se utiliza mucho este tipo de técnicas y algoritmos para predecir o pronosticar el comportamiento de sus ingresos con el fin de ayudar en la toma de decisiones.

Actualmente las empresas en al área de marketing se han percatado que poseen una gran cantidad de datos sobre sus clientes que no están aprovechando, es por este motivo que se ven en la necesidad de implementar este tipo de técnicas y algoritmos para predecir o pronosticar que clientes son más susceptibles a comprar un producto o servicio con la finalidad de comprender en qué debe centrarse la campaña de marketing y en qué no.

Naïve Bayes se trata del modelo más simple de clasificación con redes bayesianas, su estructura de la red es fija y solo necesita aprender los parámetros (probabilidades). El fundamento principal del clasificador Naïve Bayes es la suposición de que todos los atributos son independientes conocido el valor de la variable clase.

El modelo de clasificación de Naïve Bayes se define como:

$$y = y_k = max\{p(y_j) \prod_{i=1}^{n} p(x_i|y_k)\} \tag{1}$$

Donde:

y_k**:** Es el conjunto de variables dependientes.
x_i**:** Es el conjunto de variables independientes.

Debido a la hipótesis de independencia usada en el Naïve Bayes, la expresión para obtener la hipótesis MAP queda como sigue:

$$C_{MAP} = arg\_max_{c\varepsilon\Omega_C} P(c) \prod_{i=1} P(A_i|c) \tag{2}$$

Los parámetros que tenemos que estimar son $P(A_i|c)$ para cada atributo y la probabilidad a priori de la variable clase $P(c)$. Veamos cómo hacerlo dependiendo de que el atributo A_i sea discreto o continuo.

Atributos Discretos:

La estimación de la probabilidad condicional se basa en las frecuencias de aparición que obtendremos en la base de datos. La probabilidad condicional de un evento B dado un evento A, denotado como P (B | A), es:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3}$$

Si $n(x_i, Pa(x_i))$ es el número de registros de la base de datos en que la variable X_i toma el valor x_i y los padres de $X_i(Pa(X_i))$ toman la configuración denotada por $Pa(x_i)$, entonces las formas de estimar $P(x_i|Pa(x_i))$ son:

Estimación por máxima verosimilitud: El número de casos favorables dividido por el número de casos totales.

$$P(x_i|PA(x_i)) = \frac{n(x_i, Pa(x_i))}{n(Pa(x_i))} \tag{4}$$

Estimación por la ley de la sucesión de Laplace: El número de casos favorables más uno dividido por el número de casos totales más el número de valores posibles.

$$P(x_i|PA(x_i)) = \frac{n(x_i, Pa(x_i)) + 1}{n(Pa(x_i)) + |\Omega_{X_i}|} \tag{5}$$

Atributos Continuos:

Debido a que Naïve Bayes supone que el atributo en cuestión sigue una distribución normal; la estimación de la probabilidad condicional se basa en calcular es la media μ y la desviación típica σ condicionadas a cada valor de la variable clase.

$$P(A_i|c)N(\mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} exp(-\frac{(X-\mu)^2}{2\sigma^2}) \tag{6}$$

La probabilidad condicional de una variable continua x = v dado un evento A, denotado como P (x = v | A), es:

$$P(x = v|A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(v-\mu_A)^2}{2\sigma_A^2}} \tag{7}$$

Donde μ_A es la media de A y σ_A^2 es la varianza de A.

V. DESARROLLO DE CONTENIDO

Para poder pronosticar si un cliente suscribirá un depósito a plazo, se utilizó el modelo de clasificación de Naïve Bayes. El modelo de clasificación propuesto es construido a partir de un algoritmo desarrollado en lenguaje Python, a continuación se explica el proceso que sigue el algoritmo para la construcción de dicho modelo. El algoritmo consta de tres etapas principales:

- **Etapa de Preprocesamiento de Datos**
  En esta etapa el algoritmo se encarga de obtener que variables independientes son categóricas y cuales son numéricas con el fin de poder darles el tratamiento adecuado.

- **Etapa de Procesamiento de Datos**
  En esta etapa el algoritmo se encarga de realizar la construcción del modelo de Naïve Bayes, para esto el primer paso que realiza el algoritmo es realizar el cálculo de las probabilidades a priori del conjunto de variables dependientes.

$$P(C = c_x) = \frac{N_{cx}}{N} \tag{8}$$

Donde:

c_x: Es el conjunto de variables dependientes.
N_{cx}: Es el número de casos exitosos de una variable dependiente.
N: Es el tamaño del conjunto de variables dependientes.

Una vez calculada las probabilidades a priori del conjunto de variables dependientes, el algoritmo se encarga de calcular las probabilidades a posteriori de las variables independientes; utilizando la ecuación número 4 para los atributos discretos y la ecuación número 7 para los atributos continuos. Cabe aclarar que en el caso de los atributos continuos no se calcula la probabilidad en la etapa de entrenamiento lo que se calcula es la media y la varianza de una variable aleatoria.

Media de una Variable Aleatoria: La media o valor esperado de una variable aleatoria discreta $X\epsilon\{x_i|i = n_1, \ldots, n_k\}$ con función de masa de probabilidad f, se define de la siguiente manera:

$$\mu = E(X) = \sum_{i=n_i}^{n_k} f(x_i)i \tag{9}$$

Varianza de una Variable Aleatoria: La varianza de la variable aleatoria discreta $X\epsilon\{x_i|i = 1, \ldots, n\}$ con función de masa de probabilidad f, se define de la siguiente manera:

$$\sigma^2 = V(X) = \sum_{i=1}^{n} f(x_i)(x_i - E(X))^2 \tag{10}$$

- **Etapa de Resultados**

  En esta última etapa el algoritmo se encarga de evaluar el modelo de Naïve Bayes recién construido, para esto requiere de otra fuente de datos que solo contenga las variables independientes. Lo que realiza el algoritmo es pasar la nueva fuente de datos por la etapa de preprocesamiento, una vez finalizada la etapa de preprocesamiento se procede por ir obteniendo las probabilidades a posteriori de cada variable respecto a la variable a clasificar; debido a que las probabilidades ya fueron calculadas en la etapa de entrenamiento lo único que tiene que hacer el algoritmo es ir multiplicando cada una de estas probabilidades conforme lo indica la ecuación número 1.

  Tras haber realizado las multiplicaciones según indica la ecuación número 1, el algoritmo obtiene una lista de probabilidades de la cual el algoritmo obtiene la que tenga el valor más grande y lo transforma en el valor correspondiente de la variable dependiente.

VI. Resultados

A continuación se presentan los resultados obtenidos al ejecutar el algoritmo anteriormente descrito sobre la fuente de datos Bank [1]. Debido a que en este tipo de algoritmos el único resultado que arrojan es la clasificación ya realizada, a continuación se muestra las probabilidades a priori de las variables dependientes (y) y las probabilidades a posteriori de cada una de las variables independientes obtenidas en la etapa de entrenamiento. Cabe aclarar que en el caso de los atributos continuos lo que se muestra son los valores de la media y de la varianza.

Probabilidades a Priori:

- P(y=no): 0.9
- P(y=yes): 0.1

Probabilidades a Posteriori:

- N(age | y=no): [39.8608, 9.9056]
- N(age | y=yes): [40.6367, 13.3387]
- P(job=blueCollar | y=no): 0.2355
- P(job=blueCollar | y=yes): 0.1502
- P(job=entrepreneur | y=no): 0.0365
- P(job=entrepreneur | y=yes): 0.029
- P(job=admin | y=no): 0.2487
- P(job=admin | y=yes): 0.288
- P(job=management | y=no): 0.0707
- P(job=management | y=yes): 0.0725
- P(job=technician | y=no): 0.165
- P(job=technician | y=yes): 0.1547
- P(job=housemaid | y=no): 0.026
- P(job=housemaid | y=yes): 0.023
- P(job=services | y=no): 0.0994
- P(job=services | y=yes): 0.0727
- P(job=unemployed | y=no): 0.0238
- P(job=unemployed | y=yes): 0.0295
- P(job=selfemployed | y=no): 0.0349
- P(job=selfemployed | y=yes): 0.035
- P(job=retired | y=no): 0.0352
- P(job=retired | y=yes): 0.0845
- P(job=unknown | y=no): 0.0078
- P(job=unknown | y=yes): 0.0075
- P(job=student | y=no): 0.0165
- P(job=student | y=yes): 0.0532
- P(marital=married | y=no): 0.6114
- P(marital=married | y=yes): 0.5523
- P(marital=divorced | y=no): 0.1133
- P(marital=divorced | y=yes): 0.1013
- P(marital=single | y=no): 0.2735
- P(marital=single | y=yes): 0.3435
- P(marital=unknown | y=no): 0.0018
- P(marital=unknown | y=yes): 0.003
- P(education=basic9y | y=no): 0.1526
- P(education=basic9y | y=yes): 0.1095
- P(education=universitydegree | y=no): 0.2884
- P(education=universitydegree | y=yes): 0.3568
- P(education=basic4y | y=no): 0.102
- P(education=basic4y | y=yes): 0.0915
- P(education=highSchool | y=no): 0.2322
- P(education=highSchool | y=yes): 0.2223
- P(education=professionalCourse | y=no): 0.1277
- P(education=professionalCourse | y=yes): 0.123
- P(education=unknown | y=no): 0.0395
- P(education=unknown | y=yes): 0.0522
- P(education=basic6y | y=no): 0.0572
- P(education=basic6y | y=yes): 0.0437
- P(education=illiterate | y=no): 0.0004
- P(education=illiterate | y=yes): 0.001
- P(default=no | y=no): 0.7789
- P(default=no | y=yes): 0.8928
- P(default=unknown | y=no): 0.221
- P(default=unknown | y=yes): 0.1072
- P(default=yes | y=no): 0.0001
- P(default=yes | y=yes): 0.0
- P(housing=no | y=no): 0.4529
- P(housing=no | y=yes): 0.436
- P(housing=yes | y=no): 0.5229
- P(housing=yes | y=yes): 0.5427
- P(housing=unknown | y=no): 0.0242
- P(housing=unknown | y=yes): 0.0213
- P(loan=no | y=no): 0.8234
- P(loan=no | y=yes): 0.8315
- P(loan=yes | y=no): 0.1524
- P(loan=yes | y=yes): 0.1472
- P(loan=unknown | y=no): 0.0242
- P(loan=unknown | y=yes): 0.0213
- P(contact=telephone | y=no): 0.3808
- P(contact=telephone | y=yes): 0.1835
- P(contact=cellular | y=no): 0.6192
- P(contact=cellular | y=yes): 0.8165
- P(month=may | y=no): 0.3426

- P(month=may | y=yes): 0.2215
- P(month=apr | y=no): 0.0581
- P(month=apr | y=yes): 0.1348
- P(month=jul | y=no): 0.1812
- P(month=jul | y=yes): 0.1182
- P(month=nov | y=no): 0.1024
- P(month=nov | y=yes): 0.0895
- P(month=sep | y=no): 0.0087
- P(month=sep | y=yes): 0.0265
- P(month=aug | y=no): 0.1534
- P(month=aug | y=yes): 0.1338
- P(month=jun | y=no): 0.1322
- P(month=jun | y=yes): 0.129
- P(month=oct | y=no): 0.0112
- P(month=oct | y=yes): 0.0555
- P(month=dec | y=no): 0.0026
- P(month=dec | y=yes): 0.0222
- P(month=mar | y=no): 0.0075
- P(month=mar | y=yes): 0.069
- P(dayofweek=wed | y=no): 0.1996
- P(dayofweek=wed | y=yes): 0.2005
- P(dayofweek=thu | y=no): 0.2105
- P(dayofweek=thu | y=yes): 0.223
- P(dayofweek=fri | y=no): 0.1939
- P(dayofweek=fri | y=yes): 0.184
- P(dayofweek=tue | y=no): 0.1928
- P(dayofweek=tue | y=yes): 0.207
- P(dayofweek=mon | y=no): 0.2032
- P(dayofweek=mon | y=yes): 0.1855
- N(duration | y=no): [220.3793, 207.3155]
- N(duration | y=yes): [576.1475, 414.6459]
- N(campaign | y=no): [2.6508, 2.8902]
- N(campaign | y=yes): [2.0817, 1.7282]
- N(pdays | y=no): [983.8873, 121.556]
- N(pdays | y=yes): [841.7803, 362.6016]
- N(previous | y=no): [0.1344, 0.412]
- N(previous | y=yes): [0.3678, 0.7082]
- P(poutcome=nonexistent | y=no): 0.8854
- P(poutcome=nonexistent | y=yes): 0.7312
- P(poutcome=success | y=no): 0.0133
- P(poutcome=success | y=yes): 0.1475
- P(poutcome=failure | y=no): 0.1013
- P(poutcome=failure | y=yes): 0.1212
- N(emp.var.rate | y=no): [0.2359, 1.4904]
- N(emp.var.rate | y=yes): [-1.204, 1.7426]
- N(cons.price.idx | y=no): [93.5978, 0.5611]
- N(cons.price.idx | y=yes): [93.2077, 0.6052]
- N(cons.conf.idx | y=no): [-40.6569, 4.3936]
- N(cons.conf.idx | y=yes): [-39.5124, 6.2782]
- N(euribor3m | y=no): [3.7956, 1.6455]
- N(euribor3m | y=yes): [2.3188, 1.8009]
- N(nr.employed | y=no): [5175.9408, 65.0346]
- N(nr.employed | y=yes): [5113.7855, 79.5979]

VII. CONCLUSIÓN

De acuerdo a los resultados obtenidos se puede concluir que es posible construir con precisión modelo de Naïve Bayes para pronosticar si un cliente suscribirá un depósito a plazo, ya que los porcentajes de clasificación, es decir el número de casos que clasificó correctamente, tienen un margen de error mínimo del 6% y es posible que pueda mejorar su eficiencia con la ayuda del experto, ajustando los datos mismos, esto es, agregando variables o cambiando sus parámetros.

REFERENCES

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 Link: https://www.kaggle.com/henriqueyamahata/bank-marketing