PROGRAMA EDUCATIVO
Maestría en ciencias de la computación

UNIDAD DE APRENDIZAJE
MINERIA DE DATOS

Procesos de análisis supervisado

ELABORACION
ADRIAN TRUEBA ESPINOSA

PRESENTACIÓN DEL CURSO

La unidad de aprendizaje “Minería de Datos”, se imparte en la Maestría en Ciencias de la Computación. Tiene la finalidad de desarrollar las competencias en los alumnos para aplicar las metodologías para la predicción de datos a partir de algoritmos supervisados y no supervisados

Objetivo del curso

En este curso el estudiante conocerá y aplicara las metodologías para la predicción de datos que permitan pronosticar salidas de datos y revelar sus relaciones a partir de algoritmos empleados en la minería de datos supervisados y no supervisados

Justificación académica

Con el uso de las tecnologías de la información el almacenaje de datos se ha incrementado en las ultimas dos décadas de manera sustancial, estos datos pueden contener conocimiento que a simple vista no se pueden visualizar y entender. Para atacar esta situación se utiliza la minería de datos y con ello poder el conocimiento que pueda ser explotado y ser usado por las empresas o entidades publicas a fin de mejorar procesos o incrementar la entrada de recursos económicos

CONTENIDO DEL CURSO

Unidad I. Métodos para el tratamiento y análisis de datos
Unidad II. **Procesos de análisis supervisados**
Unida III. Procesos de análisis no supervisados
Unidad IV. Métodos estimadores de error
Unidad VI. Métodos para análisis del índice de aciertos

METAS A ALCANZAR

Que el alumno desarrolle las competencias técnicas y de especialidad para el tratamiento de datos con la metodología de análisis supervisado con el algoritmo de bayes

OBJETIVO DEL MATERIAL DIDÁCTICO

Que el alumno conozca el método para predicción y análisis de datos con algoritmos supervisado de Bayes

METODOLOGÍA DEL CURSO

El curso se desarrollará bajo el siguiente proceso de estudio:

1. Exposición de parte del profesor mediante la utilización de este material en diapositivas.
2. Control de lecturas selectas que el profesor asignará para complementar la clase.
3. Trabajos donde se investigarán temas, conceptos, procesos y métodos de los temas por ver.
4. Participación en clases
5. Desarrollo de trabajo individual en casos de estudio

UTILIZACIÓN DEL MATERIAL DE DIAPOSITIVAS

El material didáctico visual es una herramienta de estudio que sirve como guía para que el alumno repase los temas más significativos de “el análisis supervisado”, cabe aclarar que será un tutor el cual proporcionará las ideas generales del tema, asiendo ejercicios en el salón de clase.

UNIDAD DE COMPETENCIA

Proceso de análisis supervisado

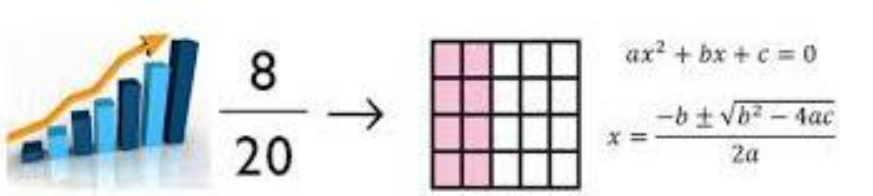
Las técnicas de Minería de Datos se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas

En esta unidad de aprendizaje se abordara las técnicas supervisada con el clasificador **Bayesiano**

Clasificador Bayesiano

Este clasificador está basado en el **Teorema de Bayes,** también conocido como teorema de la **probabilidad condicionada**, pues el porcentaje obtenido se verá condicionado o afectado por otro dato más.

- Tiene un enfoque probabilístico de inferencia.
- Asume que las incógnitas siguen distribuciones probabilísticas.
- Se consigue una solución óptima por medio de distribuciones y datos observados.
- Da la posibilidad de realizar una ponderación de la posibilidad de ocurrencia de una hipótesis de manera cuantitativa.



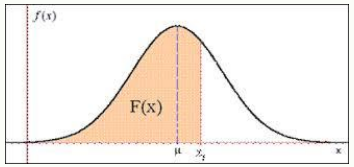
Es una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados. Utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones.

Esta basada en modelos de probabilidades que incorporan fuertes suposiciones de independencia (SI). Que no tienen ningún efecto sobre la realidad

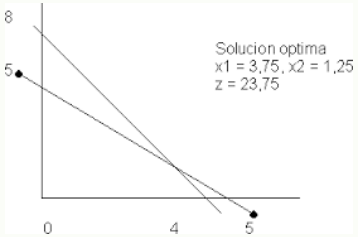
Es posible derivar modelos de probabilidades utilizando el teorema de Bayes (atribuido a Thomas Bayes). En función de la naturaleza del modelo de probabilidades, el algoritmo puede prepararse en un entorno de aprendizaje supervisado.

$$P(A|B) = \frac{P(A) \times P(B|A)}{\sum P(A) \times P(B|A)}$$

Un bayesiano



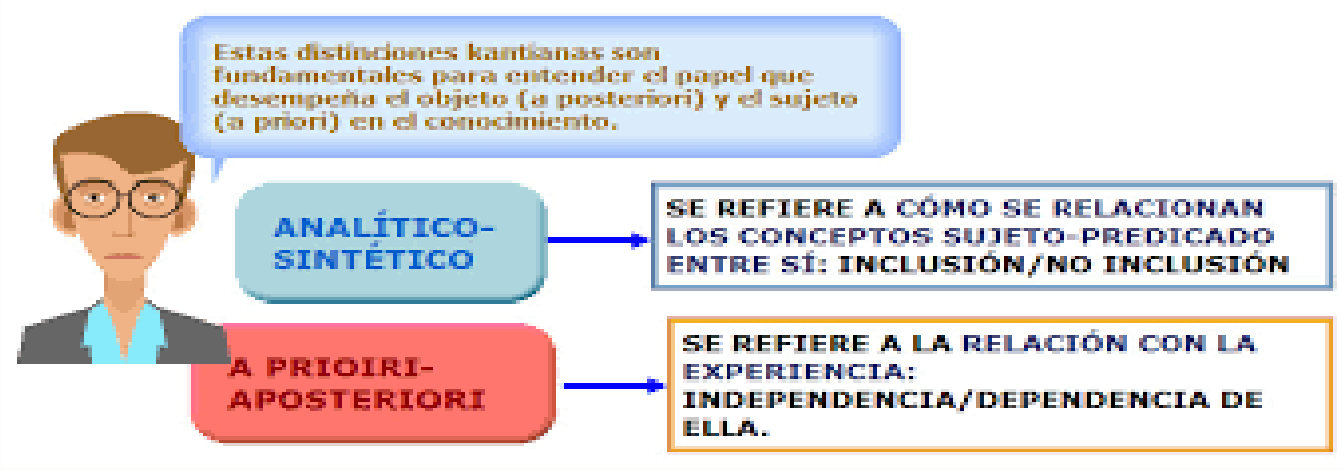
- Tiene un enfoque probabilístico de la inferencia
- Asume que las incógnitas siguen distribuciones probabilísticas



- Conseguir una solución óptima por medio de estas distribuciones y datos observados
- Da la posibilidad de realizar una ponderación de la posibilidad de ocurrencia de una hipótesis de manera cuantitativa.

Aprendizaje bayesiano

El aprendizaje se puede ver como el proceso de encontrar la hipótesis más probable, dado un conjunto de ejemplos de entrenamiento D y un conocimiento a priori sobre la probabilidad de cada hipótesis.

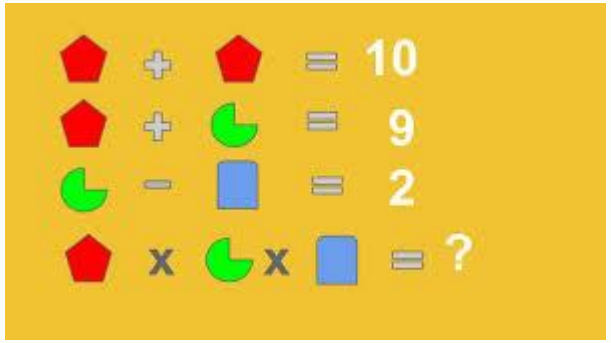


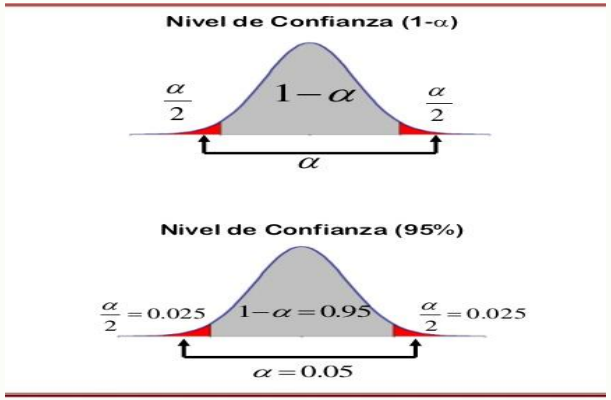
- Cada conjunto de datos de entrenamiento afecta a la probabilidad de las hipótesis. Esto es más efectivo que descartar directamente las hipótesis incompatibles.

- Se incluye conocimiento a priori: probabilidad de aserción de una hipótesis; y la distribución de probabilidades del conjunto de datos.

- Se asocia a un porcentaje de confianza cada predicción y combina predicciones en base a un intervalo de confianza.

- Una instancia es clasificada en función de la predicción de múltiples hipótesis, ponderadas por sus probabilidades.





Dificultades

• Necesita un conocimiento a priori. Si no se tiene este conocimiento estas probabilidades deben ser estimadas.

• El costo de procesamiento computacional es alto. En general es lineal con el número de hipótesis candidatas.

• Asume una independencia de las características

Ventajas

- Su implementación es muy fácil y obtiene buenos resultados de clasificación en la mayoría de los casos.
- Es rápido de entrenar
- Rápido de clasificar
- No es sensitivo a características poco relevantes
- Maneja información discreta y subjetiva
- No tiene problema manejando flujos de datos continuos

Objetivo del Clasificador Bayesiano

- Calcular estadísticamente las similitudes de las instancias de prueba contra las de referencia.

- Asumir que las características que conforman las instancias son variables aleatorias y que son independientes entre si.

Clasificador Bayesiano

- Esta basado en el teorema de Bayes que asume que un vector de características es una multivariable Gaussiana.

- La probabilidad de clasificar mal una muestra $\mathbf{x}$ es minimizada por la regla de decisión de Bayes, al asignarlo a la clase con la probabilidad *a priori* mas grande.

- El tipo de aprendizaje de este clasificador es supervisado.

Aplicaciones

- El clasificador Bayesiano ha sido aplicado para:
  - Reconocimiento de patrones
  - Procesamiento de imágenes
  - Reconocimiento de objetos con visión artificial
  - Reconocimiento de voz

Clasificador Bayesiano

- Sea $\{c_1, \ldots, c_m\}$ el conjunto de etiquetas de m clases y $\mathbf{x}$ es un vector a clasificar. La probabilidad de que el vector $\mathbf{x}$ sea de la clase c_i se denota como $p(c_i|\mathbf{x})$ del teorema de Bayes:

$$p(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)p(c_i)}{p(\mathbf{x})}$$

- Se reescribe como:

$$p(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)p(c_i)}{\sum_{j=1}^{m} p(\mathbf{x}|c_j)p(c_j)}$$

Clasificador Bayesiano

- Asumiendo *a priori* probabilidades $p(c_i)$ iguales.

- La regla de decisión con función uniforme de costo es escoger la clase para la cual $p(\mathbf{x}|c_i)$ es el mayor. Esto es:

$$\arg\max_i p(\mathbf{x}|c_i)$$

Clasificador Bayesiano

- Dada una instancia $\mathbf{x} = [x_1, \dots, x_n]$. Se asignan a esta instancia probabilidades

$$p(c_k | x_1, \dots, x_n)$$

- Para cada una de las k clases posibles.

- En la practica solo hay interés en el numerador porque el denominador es constante.

Clasificador Bayesiano

- El numerador es equivalente al modelo de probabilidad:

$$p(c_k|x_1, \dots, x_n)$$

- Se reescribe con:

$$p(c_k|x_1, \dots, x_n) = p(c_k)p(x_1, \dots, x_n|c_k)$$
$$p(c_k)p(x_1|c_k)p(x_2, \dots, x_n|c_k, x_1)$$
$$p(c_k)p(x_1|c_k)p(x_2|c_k, x_1)p(x_3, \dots, x_n|c_k, x_1, x_2)$$

Clasificador Bayesiano

- Se asume que las variables son independientes, entonces:

$$p(x_i|c_k, x_j) = p(x_i|c_k)$$
$$p(x_i|c_k, x_j, x_k) = p(x_i|c_k)$$

- Así, sucesivamente, entonces, el modelo de probabilidad se escribe como:

$$p(c_k|\mathbf{x}) = p(c_k)p(x_1|c_k)p(x_2|c_k)\cdots p(x_n|c_k)$$

Clasificador Bayesiano

- En otras palabras:

$$p(c_k|\mathbf{x}) = p(c_k)\prod_{i=1}^{n} p(x_i|c_k)$$

- Una instancia $\mathbf{x}$ es de la clase c_k al encontrar el valor de k que maximice la probabilidad:

$$\arg\max_{k\in\{1,\ldots,m\}} p(c_k)\prod_{i=1}^{n} p(x_i|c_k)$$

Clasificador Bayesiano

- Una típica suposición es que la probabilidad esta distribuida normalmente.

$$p(x_i|c_k) = \frac{1}{\sigma_{i,k}\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right)$$

- Donde:
  - x_i es una *i*-esima componente de la instancia.
  - $\mu_{i,k}$ es la media de la *i*-esima componente de x de la clase k.
  - $\sigma_{i,k}^2$ es la varianza de la *i*-esima componente de x de la clase k.

Clasificador Bayesiano con distribución Gaussiana

- Sea el vector $\mathbf{x}$ de clase c_j con distribución Gaussiana, media $\boldsymbol{\mu}_j$ y con matriz de covarianza $\boldsymbol{\Omega}_j$, entonces:

$$p(\mathbf{x}|c_j) = \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^T \boldsymbol{\Omega}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right)}{\sqrt{(2\pi)^n\left|\boldsymbol{\Omega}_j\right|}}$$

- Donde n es la dimensión de $\mathbf{x}$ y $\left|\boldsymbol{\Omega}_j\right|$ es la determinante de la matriz $\boldsymbol{\Omega}_j$.

Ejemplo

- Sean los vectores:

$$x_1 = \begin{bmatrix} 10 \\ 0 \\ 1 \end{bmatrix}, x_2 = \begin{bmatrix} 9 \\ 10 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 9 \\ 7 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 3 \\ 8 \\ 5 \end{bmatrix}$$

- Se desea construir un clasificador Bayesiano para dos clases. Los vectores 1 y 2 son muestras de la clase 1; análogamente, los vectores 3 y 4 son muestras de la clase 2.
- Se calculan sus respectivos vectores de media y desviación estándar:

$$\mu_1 = \begin{bmatrix} 9.5 \\ 5 \\ 0.5 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 0.5 \\ 7.1 \\ 0.5 \end{bmatrix}; \mu_2 = \begin{bmatrix} 6 \\ 7.5 \\ 3 \end{bmatrix}, \sigma_1 = \begin{bmatrix} 4.3 \\ 0.7 \\ 2.9 \end{bmatrix}$$

Ejemplo

- Supóngase que se desea clasificar el vector:

$$\vec{z} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

- Se calcula la probabilidad de que z sea de clase 1, es decir $p(\vec{z}|c_1)$:

$$p(z_1|c_1) = \frac{\exp\left(\frac{(2-9.5)^2}{-2(0.5^2)}\right)}{0.5\sqrt{2\pi}} = 1.1 \times 10^{-49}$$

$$p(z_2|c_1) = \frac{\exp\left(\frac{(1-5)^2}{-2(7.1^2)}\right)}{7.1\sqrt{2\pi}} = 0.047$$

$$p(z_3|c_1) = \frac{\exp\left(\frac{(3-0.5)^2}{-2(0.5^2)}\right)}{0.5\sqrt{2\pi}} = 2.9 \times 10^{-6}$$

Ejemplo

- Se calcula la probabilidad de que z sea de clase 2, es decir $p(\vec{z}|c_2)$:

$$p(z_1|c_2) = \frac{\exp\left(\frac{(2-6)^2}{-2(4.3^2)}\right)}{4.3\sqrt{2\pi}} = 0.06$$

$$p(z_2|c_2) = \frac{\exp\left(\frac{(1-7.5)^2}{-2(0.7^2)}\right)}{0.7\sqrt{2\pi}} = 1.07 \times 10^{-19}$$

$$p(z_3|c_2) = \frac{\exp\left(\frac{(3-3)^2}{-2(2.9^2)}\right)}{2.9\sqrt{2\pi}} = 0.13$$

Ejemplo

$$p(\vec{z}|c_1) = p(z_1|c_1)p(z_2|c_1)p(z_3|c_1) = 1.5 \times 10^{-56}$$
$$p(\vec{z}|c_2) = p(z_1|c_2)p(z_2|c_2)p(z_3|c_2) = 8.4 \times 10^{-22}$$

- Como $p(\vec{z}|c_1) < p(\vec{z}|c_2)$ entonces el vector $\vec{z}$ pertenece a la clase 2.

Ejemplo

Un ejemplo tomado de Pang-Ning,2014 .

Con la siguiente tabla de datos:

Id	Casa	Estado Civil	Salario trimestral(miles)	Evasor
1	Si	Soltero	125	No
2	No	Casado	100	No
3	No	Soltero	70	No
4	Si	Casado	120	No
5	No	Divorciado	95	Si
6	No	Casado	60	No
7	Si	Divorciado	220	No
8	No	Soltero	85	Si
9	No	Casado	75	No
10	No	Soltero	90	Si

Ejemplo

Bien ahora supongamos que nos piden clasificar la siguiente información:

Id	Casa	Estado Civil	Salario trimestral(miles)	Evasor
11	No	Casado	120	¿?

Primero establecer el modelo de probabilidades del clasificador, después se determinar la probabilidad de que sea un evasor contra que no lo sea, es decir , P(X=Si) y P(X=No), en donde X representa si es evasor o no.

Ejemplo

Para obtener el modelo, se debe buscar las ocurrencias de cada valor de atributo con respecto a su clase, por ejemplo , para P(Casa = SI | Es evasor = No) tenemos que es 3/7, esto debido a que tenemos 7 filas en donde evasor = No y tenemos 3 filas en donde Casa = Si y Evasor = No.

Obtenemos la siguiente tabla:

Evento	Ocurrencia
P(Casa = Si | Es evasor = No)	3/7
P(Casa = No | Es evasor = No)	4/7
P(Casa=Si | Es evasor = Si)	0
P(Casa = No | Es evasor = Si)	1
P(Estado civil = Soltero| Es evasor = No)	2/7
P(Estado civil = Divorciado| Es evasor = No)	1/7
P(Estado civil = Casado| Es evasor = No)	4/7
P(Estado civil = Soltero| Es evasor = Si)	2/3
P(Estado civil = Divorciado| Es evasor = Si)	1/3
P(Estado civil = Casado| Es evasor = Si)	0

Ejemplo

Para el atributo salario trimestral (que es continuo), se tienen dos opciones, una es discretizarlo (CAIM) y la otra es asumir una distribución de probabilidad (usualmente una distribución Gaussiana).
Usaremos la segunda opción, para ello se debe de obtener la media y la desviación estándar, para aquellos salarios anuales que son evasores y las mismas variables para los que no lo son.

Entonces:
Media No Evasores = 110. ((125 + 100 + 70 + 120 + 60 + 220 + 75) / 7)
Desviación Estándar No Evasores = 54.54.
Media Evasores = 90.
Desviación Estándar Evasores = 5

Ejemplo

Ahora aplicaremos la fórmula de la distribución

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

Esta fórmula se aplica para el valor de nuestro caso, que es de 120 (es el valor del salario trimestral que tenemos que clasificar)

P(Salario Trimestral = 120 | Evasor = No) = $\frac{1}{\sqrt{2(3.1416)(54.54)}} \exp - \frac{(120-110)^2}{2 \cdot 54.54^2}$ = 0.0072

P(Salario Trimestral = 120 | Evasor = Si) = $\frac{1}{\sqrt{2(3.1416)(5)}} \exp - \frac{(120-90)^2}{2 \cdot 5^2}$ =0.0000000012

que da como resultado : 0.0072 y para

Ejemplo

Ya con esto ahora, resta utilizar la fórmula del clasificador Bayesiano, que es la siguiente

$$P(Y|X) = P(Y) \prod_{i=1}^{d} P(Xi|Y)$$

Aquí P(Y) es la probabilidad de que ocurra la clase, para el ejemplo tenemos 10 filas en donde 7 corresponden a la clase evasor = No y tres a Evasor = Si, entonces, para evasor = NO, P(Y) = 7/10 y para Evasor = Si, P(Y) = 3/10.
d = a todas las probabilidades implicadas para los atributos en cuestión (P(Casa | Evasor), P(Estado Civil | evasor) , P(Salario Trimestral | Evasor))

Ejemplo

Para clasificar el ejemplo :

Id	Casa	Estado Civil	Salario trimestral(miles)	Evasor
11	No	Casado	120	¿?

Tendremos que aplicar la fórmula para Evasor = Si y Evasor = No,
y tomar el valor superior:

$P(X \mid No) = 7/10 \times P(\text{Casa} = No \mid \text{Evasor} = No) \times P(\text{Estado civil} = \text{Casado} \mid \text{Evasor} = No) \times P(\text{Salario trimestral} = 120 \mid \text{Evasor} = No)$

$P(X \mid No) = 7/10 \times 4/7 \times 4/7 \times 0.0072$

$P(X \mid No) = 0.0024.$

Ejemplo

P(X | Si) = 3/10 x P(Casa = No | Evasor = Si) x P(Estado civil = Casado | Evasor = Si) x P(Salario trimestral = 120 | Evasor = Si)

P(X | Si) = 3/10 x 1 x 0 x 0.0000000012

P(X | Si) = 0.

Entonces tenemos que la persona pertenece a la clase Evasor = No.
Como verán este clasificador es sencillo y eficaz, no toma muchos recursos computacionales y mientras los datos no estén correlacionados tendrá un alto grado de confiabilidad.

Trabajo en casa

Leer el articulo:

Diversidad y complejidad de la resistencia a medicamentos del HIV-1: Clasificación de mutaciones para predecir susceptibilidad o resistencia
AUTORES: Alma Ríos, Jesús González, Rigoberto Fonseca Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla México {jagonzalez, rfonseca}@inaoep.mx

El articulo lo obtiene de :

https://ccc.inaoep.mx/~esucar/Clases-mgp/Proyectos/Diversidad_y_complejidad_de_la_resistencia_a_medicamentos_del_HIV.pdf

Par entregar: Discuta las tres metodologías usadas dando énfasis a los dos de BAYES

Referencias

- Janos Abonyi, Balazs Feil, “Cluster analysis for data mining and system identification”. Birkhauser Verlag, 2007.
- Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, “Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence”. Prentice Hall, 1997.
- Shoichiro Nakamura, “Métodos numéricos aplicados a software”. Prentice Hall, 1992
- Introduction to data mining, Pang-Ning Tang, Adisson Wesley, 2014