# Diplomado en Inteligencia de Negocios
## Módulo

Minería de Datos

UNIVERSIDAD NACIONAL DE COLOMBIA

# Análisis Supervisado III
# Modelos Probabilísticos

Diplomado en Inteligencia de Negocios
Módulo 3

UNIVERSIDAD NACIONAL DE COLOMBIA

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificador Bayesiano
- Naive Bayes
- Red de creencias
- Clasificación sensible al costo

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificador Bayesiano
- Naive Bayes
- Red de creencias
- Clasificación sensible al costo

# Probabilidad

- Formalización de la noción intuitiva de la posibilidad de que un evento ocurra

$$P(E) = \frac{número\, de\, veces\, que\, sucede\, E}{posibles\, eventos}$$

- Cuál es la probabilidad de obtener el número 6 si lanzo un dado?
- Cuál es la probabilidad de obtener 10 o más si lanzamos dos dados?
- Variable aleatoria: una variable que puede tomar diferentes valores de acuerdo con una distribución de probabilidad

# Probabilidad Conjunta

- Es la probabilidad de que dos eventos sucedan a la vez:

  $P(X=x,Y=y)$

  probabilidad de que $X$ y $Y$ tomen los valores $x$ y $y$ a la vez

- $P(dado\_1=4,dado\_2=6) = ?$

# Probabilidad Condicional

- Probabilidad de que una variable aleatoria pueda tomar un valor particular dado el valor de otra variable aleatoria

$$P(Y=y \mid X=x)$$

- se refiere a la probabilidad que la variable $Y$ puede tomar el valor de $y$ dado que la variable $X$ toma el valor de $x$

# Probabilidad Condicional

- Cuál es la probabilidad de obtener 10 al lanzar un par de dados si sé que uno de los dados cayo en 4?

  *suma = dado_1 + dado_2*

  *P(suma=10 | dado_1=4) = ?*

# Teorema de Bayes

- Las probabilidades condicionales de X y Y están relacionadas:

$$P(X,Y) = P(Y|X)\ P(X) = P(X|Y)\ P(Y)$$

- Teorema de Bayes

$$P(Y|X) = P(X|Y) \cdot P(Y) \,/\, P(X)$$

- Ejercicio

  2 equipos. Equipo 0 gana el 65% de las veces, equipo 1 gana 35% de las veces. De los juegos ganados por el equipo 0, el 30% son jugados en la cancha del equipo 1. El 75%, de las victorias del equipo 1 son ganados cuando juegan en casa. Si el equipo 1 juega de local, cuál equipo es el favorito a ganar?

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificadores Bayesiano
- Naive Bayes
- Red de creencias
- Clasificación sensible al costo

# Clasificador Bayesiano

- Considere que cada atributo y la etiqueta de clase son variables aleatorias
- Dado un registro con atributos (A1, A2,…, An)
- El objetivo es predecir la clase C
- Específicamente, nosotros deseamos encontrar el valor de C que maximice P(C| A1, A2,…,An )
- Podemos estimar P(C| A1, A2,…,An ) directamente a partir de los datos?

# Solución

- calcule la probabilidad *a posteriori* $P(C \mid A_1, A_2, \ldots, A_n)$ para todos los valores de C usando el teorema de Bayes:

$$P(C|A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n|C)P(C)}{P(A_1 A_2 \ldots A_n)}$$

- Escoja el valor de C que maximice
  $P(C \mid A_1, A_2, \ldots, A_n)$
- Equivalente a escoger el valor de C que maximice
  $P(A_1, A_2, \ldots, A_n|C)\, P(C)$
- Cómo se estima $P(A_1, A_2, \ldots, A_n \mid C)$?

# Problema de tomar una decisión

Dada las condiciones del clima, es posible jugar tenis?

| Outlook | Temperature | Humidity | Windy | Class |
|---|---|---|---|---|
| sunny | hot | high | false | N |
| sunny | hot | high | true | N |
| overcast | hot | high | false | P |
| rain | mild | high | false | P |
| rain | cool | normal | false | P |
| rain | cool | normal | true | N |
| overcast | cool | normal | true | P |
| sunny | mild | high | false | N |
| sunny | cool | normal | false | P |
| rain | mild | normal | false | P |
| sunny | mild | normal | true | P |
| overcast | mild | high | true | P |
| overcast | hot | normal | false | P |
| rain | mild | high | true | N |

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificadores Bayesiano
- Naïve Bayes
- Red de creencias
- Clasificación sensible al costo

# Clasificador Naïve Bayes

- Asume independencia entre los atributos $A_i$ cuando la clase es dada:
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Se debe estimar $P(A_i \mid C_j)$ para todo $A_i$ y $C_j$.
  - Un nuevo ejemplo esclasificado como $C_j$ si $P(C_j)\, \Pi\, P(A_i \mid C_j)$ es máximo.

# Cómo Estimar las Probab. a Partir de los Datos?

UNIVERSIDAD NACIONAL DE COLOMBIA

| *Tid* | **Refund** | **Marital Status** | **Taxable Income** | **Evade** |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 95K | **Yes** |
| 6 | No | Married | 60K | **No** |
| 7 | Yes | Divorced | 220K | **No** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 75K | **No** |
| 10 | No | Single | 90K | **Yes** |

- Clase: $P(C) = N_c/N$
  - ej., $P(No) = 7/10$, $P(Yes) = 3/10$
- Para atributos discretos:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

$|A_{ik}|$ es el número de instancias con atributo $A_i$ y pertenecientes a $C_k$

Ejemplos:

$P(Status=Married|No) = 4/7$

$P(Refund=Yes|Yes)=0$

# Cómo Estimar las Probab. a Partir de los Datos?

- Para atributos continuos:
  - Discretizar: el rango en bins
    - un atributo ordinal por bin
    - viola la suposición de independencia
  - Separación: (A < v) o (A > v)
    - Escoger solo uno de los dos intervalos como nuevo atributo
  - Estimación de la distribución de probabilidad:
    - Asuma que el atributo tiene una distribución normal
    - Use los datos para estimar los parámetros de la distribución (ej.,media y desviación estándar)
    - Una vez que la distribución de probabilidad se conoce, se puede usar para estimar $P(A_i|c)$

# Cómo Estimar las Probab. a Partir de los Datos?

| Tid | Refund | Marital Status | Taxable Income | Evade |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

- Distribución normal:

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

  - Uno por cada par $(A_i,c_i)$
- Para (Income, Class=No):
  - Si Class=No
    - media muestral = 110
    - varianza muestral = 2975

$$P(Income=120|No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

# Ejemplo del Clasificador Naïve Bayes

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ $10^{-9}$ = 0

Puesto que P(X|No)P(No) > P(X|Yes)P(Yes)

entonces P(No|X) > P(Yes|X)
=> Clase = No

# Clasificador Naïve Bayes

- Si una de las probabilidades condicionales es 0, entonces toda la expresión se vuelve 0
- Estimación de la probabilidad:

$$\text{Original: } P(A_i|C)=\frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i|C)=\frac{N_{ic}+1}{N_c+c}$$

$$\text{m-estimate: } P(A_i|C)=\frac{N_{ic}+mp}{N_c+m}$$

c: número de clases

p: probabilidad a priori

m: parámetro

# Naïve Bayes (Recapitulación)

- Robusto a ejemplos ruidosos
- Maneja valores faltantes simplemente ignorando la instancia durante los cálculos de la estimación de probabilidad
- Robusto a atributos irrelevantes
- La suposición de independencia puede no cumplirse para algunos atributos:
  - Se deben usar otras técnicas tales como redes de creencias Bayesianas

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificadores Bayesiano
- Naive Bayes
- Red de creencias
- Clasificación sensible al costo

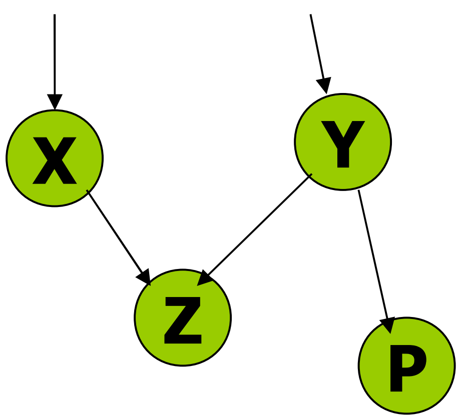
# Bayesian Belief Networks Redes de Creencias Bayesianas



- Modelar la probabilidad condicional de clases P(X|Y) sin el supuesto de independencia
- Permite especificar qué par de atributos son condicionalmente dependientes
- Pasos:
  - Representación y construcción del modelo
  - Estimación de probabilidades condicionales
  - Inferencia sobre el modelo

# Red de Creencias

- Un modelo gráfico de relaciones causales:
  - Representan dependencia condicional entre las variables
  - Variables no explícitamente relacionadas se consideran condicionalmente independientes




- Nodos: variables aleatorias
- Enlaces: dependencias
- X y Y son los padres de Z, y Y es el padre de P
- No hay dependencia entre Z y P
- No tiene bucles o ciclos

# Ejemplo Red de Creencia

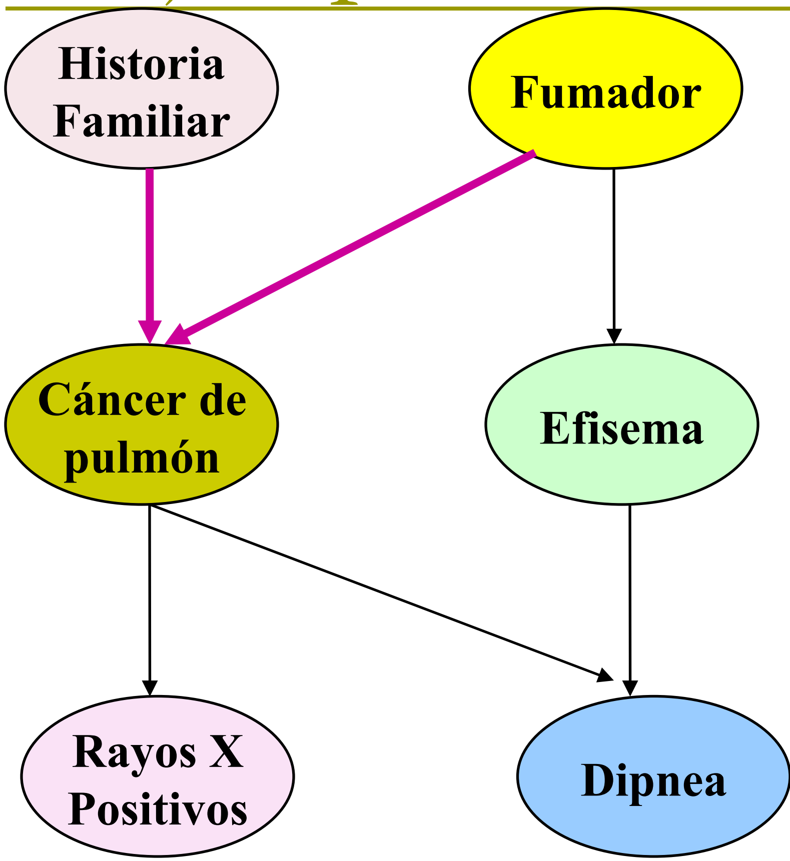



Tabla de probabilidad condicional (TPC) para la variable cáncer de pulmón:

| | (HF, F) | (HF, ~F) | (~HF, F) | (~HF, ~F) |
|---|---|---|---|---|
| CP | 0.8 | 0.5 | 0.7 | 0.1 |
| ~CP | 0.2 | 0.5 | 0.3 | 0.9 |

La derivación de la probabilidad de una combinación particular de valores de **X**, desde TPC:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Padres(Y_i))$$

$$P(HF, F, CP, E, RXP, D) = P(HF)P(F)P(CP|HF, F)P(E|F)P(RXP|CP)P(D|CP, E)$$

# Inferencia



- Diagnosticar si una persona tiene Cáncer de Pulmón:
  - Sin información previa: P(CP)
  - Rayos X Positivos: P(CP|RXP)
  - Rayos X Positivos, Fumador, No Dipnea: P(CP|RXP,F,~D)
- En todos los casos se puede calcular usando la probabilidad conjunta total y las leyes de probabilidad

# Entrenamiento de Redes Bayesianas

- Varios escenarios:
  - Dando la estructura de la red y todas las variables observables: aprende solo las TPCs.
  - La estructura de la red se conoce, algunas variables están ocultas: método del gradiente descendente (greedy hill-climbing), análogo al aprendizaje neural de la red.
  - La estructura de la red es desconocida, todas las variables son observables: buscar a través del espacio del modelo para reconstruir la topología de la red.
  - Estructura desconocida, todas las variables están ocultas: no se conocen buenos algoritmos para éste propósito.
- Ref. D. Heckerman: Bayesian networks for data mining

# Características

- Modelo grafico
- Construir la red puede ser costoso. Sin embargo, una vez construida la red, adicionar una nueva variable es directo
- Trabajan bien con datos perdidos (sumando o integrando las probabilidades)
- El modelo es robusto a overfitting

# Agenda

- Repaso de probabilidad
- Modelos Bayesianos
- Clasificadores Bayesiano
- Naive Bayes
- Red de creencias
- Clasificación sensible al costo

# Clasificación Sensible al Costo

| Area | Ejemplo |
| --- | --- |
| **Marketing** | Comprador / no Comprador |
| **Medicina** | Enfermo / no Enfermo |
| **Finanzas** | Prestar / no Prestar |
| **Spam** | Spam / no Spam |

# Suponer que los Errores Son Igualmente Costosos Pueden Llevar a Malas Decisiones

**Examples**

| | |
|---|---|
| **Marketing** | El costo de hacerle una oferta a un no comprador es pequeña comparada con no contactar un comprador |
| **Finance** | El costo de un mal prestamo es mayor que negarle un prestamo aun buen cliente |
| **Spam** | Rechazar correo que no sea Spam es más costoso que aceptar correo Spam |

# Matriz de Costos

|  |  | Actual |  |  |
|---|---|---|---|---|
|  |  | Sunny | Snowy | Rainy |
| Predicted | Sunny | 0 | 10 | 15 |
|  | Snowy | 1 | 1 | 11 |
|  | Rainy | 2 | 2 | 2 |

Fuente:  Zadrozny y Abrahams

# Costos Dependientes Fraude con Tarjeta de Créd.

**Real**

**Predicho**

| | Fraude | No fraude |
|---|---|---|
| Rechazo | 20 | - 20 |
| Aprobar | $-x$ | $(0.2)x$ |

x = *valor transación*

Fuente: Aprahams

# Aprendizaje Sensitivo al Costo

- Aprendizaje no sensitivo al costo:

$$\max_{C_i} P(C_i | A_{1,..}.., A_n)$$

- Aprendizaje sensitivo al costo:
  - Escoger acción que minimice el costo esperado

$$\min_{C_i} \sum_{C_j \neq C_i} P(C_j | A_{1,..}.., A_n) Costo(C_j, C_i)$$

  - $Costo(C_j,C_i)$ = costo de clasificar como $C_i$ cuando realmente es $C_j$
- Los dos enfoques son equivalentes cuándo los costos son iguales para todos los errores

# Metacost

UNIVERSIDAD NACIONAL DE COLOMBIA

- Es una algoritmo que permite volver cualquier clasificador sensitivo al costo
- Se debe especificar una matriz de costos
- El algoritmo reetiqueta los ejemplos de entrenamiento de manera que el costo esperado se minimice
- Domingos. *MetaCost: A General Method for Making Classifiers Cost-Sensitive*. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99). 1999.

# Bibliografía


- B. Zadrozny, J. Langford, and N. Abe. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. In Proceedings of the 2003 IEEE International Conference on Data Mining, 2003.
- Alpaydin, E. 2004 Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.
- Tan, Steinbach and Kumar, Introduction to Data Mining, Addison Wesly, 2006
- Alan Abrahams, An Introduction to Cost-Sensitive Learning , Lecture Slides, http://opim.wharton.upenn.edu/~asa28/opim_410_672_spring05/opim_410_guest_lecture_dan_fleder_cost_sensitive_learning.ppt

# Ejemplo

- X variable aleatoria que representa el equipo local
- Y variable aleatoria que representa el ganador
- Probabilidad que equipo 0 gane: $P(Y=0) = 0.65$
- Probabilidad que equipo 1 gane: $P(Y=1) = 0.35$
- Probabilidad de que si el equipo 1 gana esté jugando como local:

$$P(X=1|Y=1) = 0.75$$

- Probabilidad de que si el equipo 0 gana esté jugando como visitante:

$$P(X=1|Y=0) = 0.3$$

# Ejemplo

- Objetivo

  $P(Y=1|X=1)$ probabilidad condicional de que el equipo 1 gane el siguiente juego estando como local, y comparar con $P(Y=0|X=1)$

- Usando Bayes

  $P(Y=1|X=1) = P(X=1|Y=1)\, P(Y=1)/ P(X=1)$ Ley de probabilidad total

  $= P(X=1|Y=1)\, P(Y=1) / P(X=1,Y=1)+P(X=1,Y=0)$

  $= P(X=1|Y=1)\, P(Y=1) / P(X=1|Y=1)P(Y=1) \quad +P(X=1|Y=0)P(Y=0)$

  $= 0.75x0.35/(0.75x0.35 + 0.3x0.65) = 0.5738$

- $P(Y=0|X=1) = 1 - P(Y=1|X=1) = 0.4262$

Equipo1 tiene mas oportunidad de ganar