Clasificación basada en Naïve Bayes

- Familia de modelos basados en Bayes.
- Veremos Clasificador de Naive Bayes.
- También existen las Redes Bayesianas.

Clasificación Basada en Naïve Bayes

- Modelo que busca **modelar la relación probabilística** entre atributos y clase.
- Modelo generativo, asume una distribución conjunta entre X e Y.
- Supuesto: atributos independientes dado la clase (**naive assumption**).

Clasificador Bayesiano

- Esquema probabilístico para resolver problemas de clasificación.

- Probabilidad condicional:

$$P(C \mid A) = \frac{P(A, C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A, C)}{P(C)}$$

- Teorema de Bayes:

$$P(C \mid A) = \frac{P(A \mid C) P(C)}{P(A)}$$

Ejemplo Teorema de Bayes

- Dado:
  - Un doctor sabe que la meningitis produce rigidez de cuello el 50% de las veces.
  - La probabilidad previa de que cualquier paciente tenga meningitis es 1/50,000.
  - La probabilidad previa de que cualquier paciente tenga rigidez en el cuello es de 1/20.
- ¿Si un paciente tiene el cuello rígido, cuál es la probabilidad de que tenga meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Clasificador Naïve Bayes

- Considerar cada atributo como variable condicionalmente independiente de la clase (eso es “naive”).
- Dado un record con atributos $(A_1, A_2,\ldots,A_n)$.
  - La meta es predecir la clase C.
  - Específicamente queremos encontrar el C que maximice $P(C| A_1, A_2,\ldots,A_n)$.
- ¿Podemos estimar $P(C| A_1, A_2,\ldots,A_n)$ directamente de los datos?

Clasificador Naïve Bayes

- Aproximación

  - Computar la probabilidad posterior $P(C \mid A_1, A_2, \ldots, A_n)$ para todos los valores de C usando el Teorema de Bayes.

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

  - Elegir un valor de C que maximice
    $P(C \mid A_1, A_2, \ldots, A_n)$.

  - Equivalente a elegir un valor de C que maximice
    $P(A_1, A_2, \ldots, A_n|C)\ P(C)$.

  - Esto es porque el numerador $P(A_1A_2...A_n)$ es constante para todas las clases.

Clasificador Naïve Bayes

- Asume independencia entre los atributos A_i cuando la clase está dada (independencia condicional):
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$.
  - Se puede estimar $P(A_i \mid C_j)$ para todos los A_i y C_j.
  - Un punto nuevo A, se clasifica como C_j si $P(C_j)\, \Pi\, P(A_i \mid C_j)$ es máxima (en comparación con otros valores de C).

¿Cómo estimar probabilidades 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

- Clase: $P(C_k) = \frac{count(C_k)}{N}$
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10
- Para atributos discretos:

$$P(A_i = b|C_k) = \frac{count(A_{ik} = b)}{count(C_k)}$$

  - donde count(A_{ik}=b) es el número de instancias que tiene el valor b para el atributo A_i y que pertenecen a la clase C_k
  - Ejemplos:
  - P(Status =Married|No) = 4/7
    P(Refund= Yes |Yes) = 0

Laplace Smoothing

- $P(C \mid A_1, A_2, \ldots, A_n)$ se puede ir a cero cuando $|A_{ik}=b| = 0$, osea cuando para alguna clase C_k no hay ningún ejemplo con $A_i=b$.
- En ese caso Naive Bayes le asignaría probabilidad cero a la clase C_k a cualquier ejemplo con $|A_{ik}=b| = 0$, ignorando el valor de los otros atributos (acuérdense que las probabilidades se multiplican).
- Eso no es bueno para la generalización del modelo.
- Laplace Smoothing: soluciona el problema sumándole 1 a todos los conteos para que ninguna probabilidad quede en cero:

$$P(A_i = b|C_k) = \frac{count(A_{ik} = b) + 1}{count(C_K) + values(A_i)}$$

- Donde values(A_i) es la cantidad de categorías del atributo A_i.
- Con Laplace smoothing P(Status =Married|No) = (4+1)/(7+3)

Atributos Numéricos

- ¿Cómo calculamos $P(A_{ik}=b|C_k)$ cuando el atributo A_i es numérico (ej: Taxable income) ?
- Una opción es discretizar el atributo y proceder de la forma anterior.
- Otra solución es asumir que el atributo sigue una distribución Gaussiana y estimar los parámetros de la función de densidad:

$$P(A_i = b|C_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp^{-\frac{(b-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

- Aquí μ_{ik} y σ_{ik} se estiman como la media muestral y la desviación estándar de los ejemplos del atributo A_i cuando la clase es C_k
- Sea A_i = Taxable income y C_k=No, μ_{ik}= mean(125,100,70,120,60,220,75) =110 y σ_{ik}=sd(125,100,70,120,60,220,75)=54.5

P(Taxable Income= 130 |No)= dnorm(x=130,mean=110,sd=54.5) = 0.006843379

Naïve Bayes (Resumen)

- Es robusto ante puntos de ruido aislados.
- Maneja valores faltantes ignorando la instancia durante los cálculos de estimación de probabilidades.
- Robusto a atributos irrelevantes (afectan de igual manera a todas las clases).
- El supuesto de independencia entre atributos puede no ser cierto en todos los casos.
- Las redes Bayesianas o los modelos gráficos dirigidos permiten hacer modelos probabilísticos con supuestos de independencia menos restrictivos.