Universidad de Santiago de Chile
Facultad de Ingeniería
Depto. de Ingeniería Informática

Minería de Datos
Capítulo VI
"Clasificación Bayesiana"

Profesor: Dr. Max Chacón.

Objetivos

- **Cuantificar probabilidad a priori.**
- **Comprender el costo del error de clasificar basado en probabilidad a priori.**
- **Cuantificar el riesgo condicional.**
- **Usar los conceptos anteriores para obtener un método de clasificación a mínimo riesgo condicional para un problema multivariado.**
- **Analizar clasificador Bayesiano simple.**
- **Obtener un clasificador mediante criterio distribuido, usando el concepto de redes de clasificación.**

6.1. Clasificación a priori

Supongamos que se requiere clasificar pacientes que recurren al médico con dolor pectoral, en pacientes que han sufrido un infarto cardiaco y los que su dolor es por otra causa.

De un número grande de observaciones n, se puede obtener que una fracción de ellos n_1 pertenece a la clase c_1 (pacientes con infarto) y la fracción n_2 pertenece a la clase c_2 (pacientes sin infarto)

Con: $\boldsymbol{n=n_1+n_2}$

La *probabilidad a priori* $p(c_i)$ será la probabilidad de que el próximo paciente se clasifique en la clase c_i.

$$p(c_i) = \lim_{n\to\infty} \frac{n_i}{n} \quad \text{con } i=1,2.$$

Para un número grande de observaciones se puede estimar $p(c_i)$ por:

$$\hat{p}(c_i) = \frac{n_i}{n}$$

Las probabilidades $p(c_1)$ y $p(c_2)$ representan el conocimiento a *priori* (en términos estadísticos) de que un paciente tenga infarto o no, antes de que exista el nuevo paciente a clasificar.

Suponiendo que, basado en estas probabilidades (pequeño conocimiento) se quiere clasificar un nuevo paciente, la mejor elección será aquella que asigna el paciente a la clase que tenga mayor probabilidad a priori.

Asignar paciente a la clase c_2 si $p(c_2) > p(c_1)$
En otro caso asignar a c_1.

Para un sujeto en particular, la probabilidad de error de la clasificación será:

$$p(\textit{error de clasificación}) = \begin{cases} p(c_2)\ si\ se\ decide\ C = c_1 \\ p(c_1)\ si\ se\ decide\ C = c_2 \end{cases}$$

Se observa que el error de clasificación es minimizado si se elige c_2 y $p(c_2) > p(c_1)$.

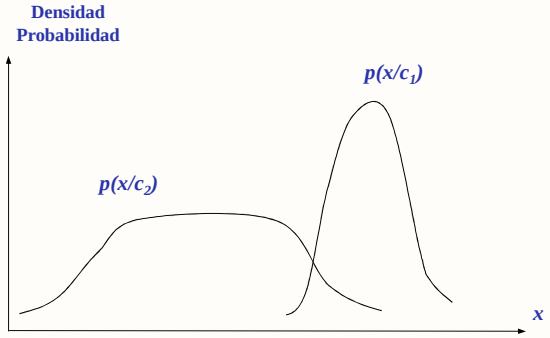
6.2. Clasificación condicionada

Para tomar una decisión más informada es necesario usar alguna característica relevante del paciente que permita distinguir su enfermedad.

Considérese que el médico tratante solicita examen *CKMB*, que la fracción *MB* del nivel *Sérico* de la enzima *Creatinin Fosfokinasa CK* para detectar infarto al miocardio.

Considere x la fracción del nivel sérico de la enzima. El valor de x debe ser considerado como una variable aleatoria que puede ser expresada en términos probabilísticos (también influye el tiempo).

El interés será contar con funciones de densidad de probabilidad condicionales $p(x/c_i)$, con $i=1,2$. Sea c_1 paciente con infarto, c_2 sin infarto



$p(x/c_1)$ es la función de densidad de probabilidad para un valor de x dado que el paciente presenta infarto.

Esta función de densidad de probabilidad es llamada *Verosimilitud* de la clase c_1 con respecto a x y refleja el conocimiento que se tiene de la aplicación de esta función, la cual sugiere que la *verosimilitud* de que un paciente pertenezca a la clase c_1 es grande si $p(x/c_1)$ es grande.

Nota: si las dos funciones de distribución están superpuestas, significa que el conocimiento de la variable x no discrimina entre sano y enfermo.

Para clasificar el paciente en la clase c_i se requiere reunir el conocimiento a priori $p(c_i)$ y el conocimiento de los valores de x para los pacientes que pertenecen a las clases c_i , $p(x/c_i)$.

Para clasificar un paciente dado se requiere la probabilidad a posteriori $p(c_i/x)$, la cual especifica la probabilidad de que el sujeto pertenezca a la clase c_i dado que el valor del nivel sérico es x.

Tener el valor de x dependerá del hecho posterior de que la variable de características x sea medida.

Para dos clases se tiene: $\sum_{i=1}^{2} p(c_i / x) = 1$

Para obtener el valor de $p(c_i/x)$, se requiere conocer las relaciones de probabilidades condicionales.

Sea: $p(c_i;x)$ la función de densidad de probabilidad conjunta, $p(c_i \cap x)$ que es interpretada como la probabilidad de que un paciente pertenezca a la clase c_i y tenga un nivel sérico x.

De la definición de probabilidad condicional:

$$p(c_i;x) = p(c_i/x)p(x) \quad con \quad i = 1,2.$$
$$p(c_i;x) = p(x/c_i)P(c_i) \quad con \quad i = 1,2.$$

$p(x)$ es la probabilidad incondicional de la variable x.

$$p(x) = \sum_{i=1}^{2} p(x/c_i)P(c_i)$$

Usando las definiciones anteriores se tiene el *Teorema de Bayes*;

$$p(c_i/x) = \frac{p(x/c_i)P(c_i)}{\sum_{i=1}^{2} p(x/c_i)P(c_i)}$$

Por lo tanto para conocer $p(c_i/x)$ se requiere el conocimiento a priori y el conocimiento de la verosimilitud de las clases c_i respecto a x.

6.3. Minimización del Riesgo Condicional

"*Un nuevo paciente al cual se han practicado el examen CKMB con un resultado x, se asigna a la clase c_i que tenga el mayor valor de $P(c_i/x)$*".

A esta regla se le denomina Maximización de la Probabilidad (o hipótesis) a Posteriori (MAP).

$$Clase_{MAP} = \arg\max_{c_i}\{p(c_i/x)\}$$

O

$$Clase_{MAP} = \arg\max_{c_i}\{p(x/c_i)\,p(c_i)\}$$

En otras palabras: La regla de clasificación estadística indica que la mejor clasificación será aquella que minimice la probabilidad de error en la clasificación (*Regla de Clasificación Bayesiana*).

Se decide por la clase c_1 si:

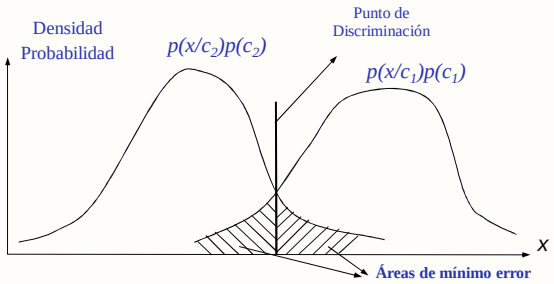
$$p(c_1/x) > p(c_2/x) \Rightarrow \frac{p(x/c_1)P(c_1)}{p(x)} > \frac{p(x/c_2)P(c_2)}{p(x)}$$

$$p(x/c_1)P(c_1) > p(x/c_2)P(c_2)$$

Se decide por la clase c_2 si:

$$p(x/c_2)P(c_2) > p(x/c_1)P(c_1)$$

Usando la función de distribución de probabilidad a posteriori se puede observar el punto óptimo de discriminación.



Se denomina ***Función Discriminante*** *a:*

Si se tienen múltiples (m) clases

$$d_i(x) = P(c_i / x) = \frac{p(x / c_i)P(c_i)}{P(x)} \qquad \text{con: } i=1,2,\ldots m$$

Como $d_i(x)$ se compara con los otros $d_j(x)$, $j=1,2,\ldots,m$, el factor de escala $p(x)$ no necesita ser considerado.

$$d_i(x) = p(x/c_i)P(c_i)$$

En general $d_i(x)$ es una función monótona respecto de x y se puede usar $ln(d_i(x))=L_i(x)$ obteniendo los mismos resultados.

$$L_i(x)=ln\ p(x/c_i)+ln\ p(c_i)$$

La elección de la clase se realiza maximizando la probabilidad a posteriori (MAP).

$$Clase_{MAP} = \arg\max_{c_i} \left\{\ln(p(x/c_i))+\ln(p(c_i)\right\}$$

Ej: Considere un paciente que recibe un examen CKMB con valor 15% (<10% normal, a las 7 horas) y se quiere saber si realmente tiene infarto. De las bases de datos del hospital se sabe que:

- De los pacientes que consultan por dolor agudo al pecho y se les envía a realizar el examen CKMB, el 60% tuvo infarto realmente.

- Además se sabe que el 1% de los pacientes con infarto tenia un valor de 15% de la fracción CKMB y que solo el 0,3% de los que no tuvo infarto tenían un valor de 15% de la encima.

Determine si el paciente tiene o no infarto realmente.

Sol: P(I)=0,6; P($\bar{I}$)=0,4;

p(x=15%/I)=0,01;
p(x=15%/$\bar{I}$)=0,003.

arg max {p(I/x=15%); p($\bar{I}$/x=15%}=

arg max {$p(x=15\%/I)P(I); p(x=15\%/\bar{I})P(\bar{I})$}=

arg max{0,006;0,0012}=arg {0,006}= clase I

∴ El paciente tiene infarto.

6.4. Clasificación Multivariada

Se tienen *n* pacientes portadores de diferentes enfermedades y se requiere clasificarlos en *m* clases c_1, c_2, … ,c_m (enfermedades y caso normal) las cuales se dan en proporciones a priori $p(c_1)$, $p(c_2)$, … ,$p(c_m)$, y se poseen *p* características de los pacientes representadas en el vector de valores reales $\vec{x}=[x_1,x_2,\ldots x_p]$.

Si la muestra utilizada es significativa, se puede suponer que la distribución de las variables aleatorias $\vec{x}$ es una distribución normal multivariada.

En este caso, la probabilidad de obtener un paciente con características $\vec{x}$ que pertenezca a la clase c_i es:

$$p(\vec{x}/c_i)=\frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}}e^{-\frac{1}{2}(\vec{x}-\vec{\hat{\mu}})^T\Sigma_i^{-1}(\vec{x}-\vec{\hat{\mu}})}$$

donde: $\vec{\hat{\mu}}$ = estimación del vector de medias de las *p* características de la clase c_i.

Σ_i: Matriz de varianzas-covarianzas de la clase c_i.

Considerando el estudio univariado, lo que se requiere es:

Dado un paciente que posee un vector de características $\vec{x}$, determinar la clase a la cual pertenece el paciente, $p(c_i/\vec{x})$.

$$Clase_{MAP} = \arg\max_{c_i} \{p(\vec{x}/c_i)\,p(c_i)\}$$

Definiendo la probabilidad a como la función discriminante $d_i(\vec{x}) = p(\vec{x}/c_i)P(c_i)$ y su logaritmo $L_i(\vec{x}) = \ln d_i(\vec{x})$

$$L_i(\vec{x}) = \ln\{p(\vec{x}/c_i)P(c_i)\}$$

Considerando la función densidad de probabilidad como una normal multivariada, se tiene:

$$L_i(\vec{x}) = \frac{p}{2}\ln(2\pi)\ln\{p(c_i)\} - \frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\vec{x}-\hat{\vec{\mu}}_i)^T\Sigma_i^{-1}(\vec{x}-\hat{\vec{\mu}}_i)$$

La clase se obtiene por MAP como:

$$Clase_{MAP} = \arg\max_{c_i}\left\{\frac{p}{2}\ln(2\pi)\ln\{p(c_i)\} - \frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\vec{x}-\hat{\vec{\mu}}_i)^T\Sigma_i^{-1}(\vec{x}-\hat{\vec{\mu}}_i)\right\}$$

Por lo tanto, se asigna el paciente representado por $\vec{x}$ a la clase c_i donde se alcanza mayor valor de $L_i(\vec{x})$.

Esta función de decisión es de tipo cuadrática.

MAP indica elegir la probabilidad a posteriori, $p(c_i/\vec{x})$ de asignar el paciente $\vec{x}$ a la clase elegida, la cual, dado el criterio de decisión, corresponde a la que tiene una mayor probabilidad a posteriori.

6.5. Calificador Bayesiano Ingenuo *("Naive")*.

El clasificador Bayesiano óptimo obtenido por MAP en la sección anterior supone los atributos $\vec{x}$ reales y dependientes entre sí.

En un caso mas general se pueden considerar diferentes tipos de atributos $\{a_1, a_2, ..., a_p\}$ que pueden ser reales, nominales o binarios y el clasificador óptimo será determinado por:

$$Clase_{MAP} = \arg \max_{c_i} \{p(c_i / a_1, a_2 ..., a_p)\}$$

o al usar la verosimilitud y la probabilidad a priori:

$$Clase_{MAP} = \arg \max_{c_i} \{p(a_1, a_2 ..., a_p / c_i)\, p(c_i)\}$$

El problema que presenta este acercamiento al problema de clasificación es que se requiere determinar la verosimilitud de cada uno de los p atributos (conjuntamente) con respecto a las m clases c_i.

Lo cual es una probabilidad muy difícil de obtener, pues se requiere una cantidad muy grande de datos para obtener todas las posibilidades.

Si se consideran todos los a_j ($j=1..p$) reales y que la distribución de la verosimilitud es normal, se obtiene el caso anterior.

La *Clasificación Bayesiana Ingenua* realiza una suposición simple: todas las probabilidades de los atributos son condicionalmente independientes para una clase dada.

Esto significa que dada una clase, la probabilidad de observar la conjunción de los $\{a_1, a_2, ..., a_p\}$ corresponde al producto de las probabilidades individuales de los atributos. Esto es:

$$p(a_1, a_2 ..., a_p / c_i) = \prod_{j=1}^{p} P(a_j / c_i)$$

Sustituyendo esto en el clasificador Bayesiano óptimo obtenido por MAP, se tiene el clasificador Bayesiano ingenuo:

$$Clase_{NBC} = \arg \max_{c_i} \left\{ p(c_i) \prod_{j=1}^{p} P(a_j / c_i) \right\}$$

Se puede observar que en este caso es muy fácil estimar los $p(a_i/c_j)$ del conjunto de datos.

• Para el caso de atributos nominales:

Ej: Suponiendo que a_j puede tomar 4 valores $k \in \{0,1,2,3\}$ entonces:

$$\hat{p}(a_j = k / c_i) = \frac{N^o \, casos \, en \, que \, a_j = k \, en \, la \, clase \, c_i}{Total \, de \, casos \, de \, la \, clase \, i}$$

• Para el caso de atributos reales se puede usar una distribución normal.

Ej: Se estima la media $\bar{a}_j$ y la varianza $\hat{\sigma}_{a_j}$ del atributo y se obtiene la estimación de la probabilidad como:

$$\hat{p}(a_j / c_i) = \frac{1}{\hat{\sigma}_{a_j} \sqrt{2\pi}} e^{-\frac{(a_j - \bar{a}_j)^2}{2\hat{\sigma}^2_{a_j}}}$$

.

El método se aplica primero con un conjunto de datos de los cuales se estiman las probabilidades a priori $p(c_i)$ y las verosimilitudes $p(a_j / c_i)$ para todas las clases i.

Esto se puede considerar la etapa de entrenamiento o aprendizaje del método.

Cuando se presenta una nueva instancia $a_j = k$, de un conjunto desconocido de datos, se determina la clase donde el producto de las probabilidades a priori y verosimilitudes es máxima.

6.6. Aplicación a minería de texto

Para analizar lenguaje natural, se requiere algunas trasformaciones del "Corpus", conjunto de textos para entrenar y evaluar.

En general estas dependen del idioma del texto.

- Normalización: Unificar términos que representan la información y se escriben en forma diferente Ej: "restaurante", "restaurant, "restaurán".
- Tokenización: Separación de las palabras, usualmente palabras (tokens), usualmente separadas por blancos o caracteres especiales. Ej: unidades multi-palabras, nombres propios, fechas, unidades monetarias etc.

- **Stremming:** Algoritmo que permite obtener la raíz de la palabra eliminando terminaciones Ej: “recomendar”, “recomendable”, recomendación” = recomend
- **Lematización:** Algoritmo para encontrar el lema, elimina todas las flexiones de una palabra, esto es, elimina todas las: conjugaciones, grado, persona, genero, numero, etc. Ej: Lema(pésimo)=malo, lema(primero)=primeras.
- **Tratamiento de Negaciones:** Se identifican claramente las negaciones con un símbolo especial, Ej: NOT_.
  - Ej: El servicio no es bueno y la comida es mala = El servicio NOT_es bueno y la comida NOT_buena.

- **Part of Spéech (POS) Tagging (Etiquetado gramatical)**

Agrega a cada termino su etiqueta, cuando cada palabra esta actuando como: Sustantivo, verbo, adjetivo, articulo, adverbio, etc.

- Cada uno de estos pre-procesos son aplicados a todo el Corpus antes de aplicar el método de clasificación.

- La clasificación de texto puede tener varias formas de realizarlo, en general el problema consiste en asignar un texto completo a alguna clase. Ej: clasificar artículos de periódicos por canales (política, deporte, actualidad etc.)

6.6.1. Clasificador Bayesiano para texto.

En este caso se usa el denominado ***Clasificador Bayesiano Ingenuo Miltinomial***.

$$\hat{p}(w_k \mid C_i) = \frac{N^o\, casos\,(w_k, C_i) + 1_i}{|V| + \sum_w N^o\, casos(w_k, C_i)}$$

Donde:

- ***V*** total de palabras en el corpus.
- ***N° casos*** (w_k, C_i) es la cantidad de veces que la palabra w_k aparece el la clase C_i.
- $\sum$***N° casos*** (w_k, C_i) suma de las frecuencias de aparición de cada termino en la clase C_i.

6.6.2. Evaluación de la clasificación de texto.

La evaluación de un clasificador de texto supervisado, se basa en los principios de la recuperación de información.

Para eso se usa la misma matriz confusión que se utiliza para un clasificador, pero esta no compara exactamente la clase verdadera con las respuestas del clasificador. Compara los documentos recuperados con los documentos relevantes:

Documentos	**Relevantes**	**No relevantes**
Recuperados	Verdaderos Positivos	Falsos Positivos
No recuperados	Falsos Negativos	Verdaderos Negativos

- En realidad aquí los documentos no relevantes no recuperados no interesan.
- La proporción de relevantes que son recuperados de todos los recuperados:

$$\text{Precisión} = \frac{Vp}{Vp + Fp}$$

(Valor Predictivo Positivo)

- La proporción de relevantes que son recuperados del total de relevantes:

$$\text{Recall} = \frac{Vp}{Vp + Fn}$$

Exhaustividad

(Sensibilidad)

- Podría ser que existe una gran cantidad de documentos relevantes recuperados (gran precisión), pero también en la recuperación se recuperaron muchos documentos irrelevantes, esto un gran Recall. Luego la recuperación total no es eficiente.
- Existe una métrica que pondera Precisión (P) y Recall (R) promedia en forma armónica las dos medidas.

$$F = \frac{1}{\alpha \frac{1}{p} + (1-\alpha)\frac{1}{R}} = \frac{PR}{\alpha P + (1-\alpha)R}$$

- Si $\alpha = 1/2$ se llama medida F_1

$$F_1 = \frac{2PR}{P + R}$$