Capítulo 4

Estimación máximo a posteriori

4.1 Generalidades

Cuando queremos minimizar la probabilidad de error o minimizar el riesgo de clasificar erróneamente un elemente dentro de una clase, podemos hablar de dividir un espacio de características en M clases diferentes. Si dos regiones R_i y R_j son contiguas, entonces ellas están separadas por una superficie de decisión en dicho espacio.

Para el caso del menor error de probabilidad, este es descrito por la fórmula $P(w_i \mid x) - P(w_j \mid x) = 0$ donde w representa a una clase; en este caso de un lado de la superficie la diferencia es positiva y del otro lado es negativa.

En algunas ocasiones, en lugar de trabajar directamente con probabilidades (o funciones de riesgo), es más conveniente, desde un punto de vista matemático, trabajar con una función equivalente de ellas, por ejemplo $g_i(x) = f(P(w_i \mid x))$ donde $f(\cdot)$ es una función creciente monótona, en ese caso $g_i(x)$ es conocida como una función discriminante. La prueba de decisión se establece como: clasificar x en w_i si $g_i(x) > g_j(x)\ \forall i \neq j$ [11]

4.2 Teoría Bayesiana

Sean X y Y dos variables aleatorias de un experimento aleatorio con un espacio de muestras y una probabilidad P. Supongamos que X toma valores dentro del espacio de muestras S y Y toma valores dentro del espacio de muestras T. Las distribuciones de X y Y son llamadas distribuciones marginales.

Si X y Y son independientes, la probabilidad de que ocurra A y B está dada por el producto de sus probabilidades individuales, es decir $P(A,B) = P(A)P(B)$

Si por el contrario, las variables no son independientes, la probabilidad de que ocurra A dado B, está determinada por la expresión $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

De acuerdo a esta última expresión, tenemos que $P(A,B) = P(B \mid A)P(A)$, y juntando las ecuaciones, llegamos a la regla de Bayes, la cuál está determinada por la expresión $P(A \mid B) = \dfrac{P(A)P(B \mid A)}{P(B)}$

La ley de la probabilidad total dice que si $\{A_1, A_2, A_3, ..., A_n\}$ es un conjunto de variables independientes cuyo espacio es una partición del espacio de muestras, y si una variable aleatoria B es dependiente de las variables $A_1, A_2, A_3, ..., A_n$ entonces la probabilidad de B está dada por $P(B) = \sum_{k=1}^{k=n} P(B \mid A_k)P(A_k)$.

4.3 Caso binario de Segmentación

Tomemos un caso binario [14] en el cuál cada patrón X tenga elementos binarios independientes como se muestra a continuación:

$X = [x_1, x_2, \ldots, x_n]^T \quad x_i = 1 \text{ ó } 0$

Para un problema de dos clases (M = 2), la función discriminante es $d(x) = d_1(x) - d_2(x)$ donde $d_1(x) = \log[p(X|w_1)p(w_1)]$ y $d_2(x) = \log[p(X|w_2)p(w_2)]$ por lo tanto: $d(X) = \log\frac{p(X|w_1)}{p(X|w_2)} - \log\frac{p(w_1)}{p(w_2)}$

Dado que es un problema de 2 clases tenemos $p(w_1) + p(w_2) = 1$ y por lo tanto $d(X) = \log\frac{p(X|w_1)}{p(X|w_2)} + \log\frac{p(w_1)}{1 - p(w_1)}$

Dado que los componentes x_i son independientes: $p(X|w_j) = p(x_1|w_j)p(x_2|w_j)p(x_3|w_j)\cdots p(x_n|w_j) = \prod_{i=1}^{n} p(x_i|w_j)$ y

$d(X) = \sum_{i=1}^{n} \log\frac{p(x_i|w_1)}{p(x_i|w_2)} + \log\frac{p(w_1)}{1 - p(w_1)}$

Dado que los elementos de X son binarios para este caso, simplificando tenemos $p(x_i = 1|w_1) = p_i$ por lo tanto $p(x_i = 0|w_1) = 1 - p_i$ y de manera similar tenemos $p(x_i = 1|w_2) = q_1$ y $p(x_i = 0|w_2) = 1 - q_i$

Entonces podemos decir que: $\log\frac{p(x_i|w_1)}{p(x_i|w_2)} = x_i \log\frac{p_i}{q_i} + (1 - x_i)\log\frac{1 - p_i}{1 - q_i} = \gamma$

Reescribiendo tenemos $\gamma = \log\frac{p(x_i|w_1)}{p(x_i|w_2)} = x_i \log\frac{p_i(1 - q_i)}{q_i(1 - p_i)} + \log\frac{1 - p_i}{1 - q_i}$

Y sustituyendo en la función discriminante tenemos $d(X) = \sum_{i=1}^{n} x_i \log\frac{p_i(1 - q_i)}{q_i(1 - p_i)} + \sum_{i=1}^{n} \log\frac{1 - p_i}{1 - q_i} + \log\frac{p(w_1)}{1 - p(w_1)}$ donde podemos representar a $\log\frac{p_i(1 - q_i)}{q_i(1 - p_i)}$ como w_i y a $\sum_{i=1}^{n} \log\frac{1 - p_i}{1 - q_i} + \log\frac{p(w_1)}{1 - p(w_1)}$ como w_{n+1}.

Por lo tanto tenemos que la función discriminante queda como $d(X) = \sum_{i=1}^{n} w_i x_i + w_{n+1}$ en donde vemos que la función discriminante optima es linear en x_i.

4.4 Aproximación máximo a posteriori-MAP

Para la obtención del estimado de mayor parecido, consideramos θ como un parámetro desconocido; consideramos un vector aleatorio, y estimamos su valor con la condición de muestras ocurridas $x_1, \ldots x_N$ [11]

Siendo $X = \{x_1 ... x_N\}$ nuestro punto inicial es $p(\theta \mid X)$ y del teorema de Bayes tenemos $p(\theta)p(X \mid \theta) = p(X)p(\theta \mid X)$ o $p(\theta \mid X) = \frac{p(\theta)p(X \mid \theta)}{p(X)}$

La probabilidad máximo a posteriori (MAP) estimada $\hat{\theta}_{MAP}$ está definida en el punto $p(\theta \mid X)$ máximo.

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial \theta} p(\theta \mid X) = 0 \text{ ó } \hat{\theta}_{MAP} : \frac{\partial}{\partial \theta}\left(p(\theta)p(X \mid \theta)\right) = 0$$

4.5 Caso aparte: Algoritmo de k-medias

El algoritmo de clasificación conocido como k-medias o isodata es uno de los algoritmos más populares y mejor conocidos [13]. Este algoritmo realiza ajustes iterativos del centro de los diferentes sectores a clasificar.

La idea central del algoritmo es minimizar la distancia entre los patrones y el centro del sector. Normalmente, excepto en casos de un pequeño número de patrones, una búsqueda exhaustiva de todas las particiones de *n* patrones en *c* sectores es prohibitivo. En vez de eso, la distancia mínima local dentro de un sector es buscada iterativamente ajustando los *c* centros de los sectores, m_j, y asigando a cada punto el centro más cercano.

$E = \sum_{j=x_j \in w_j}^{c} \sum \left\| x_i - m_j \right\|^2$ donde E puede ser también visto como el error de la clasificación de los patrones.

El algoritmo puede ser descrito como sigue:

1. Fijar *c* como el número de sectores, *maxiter* como el número máximo de iteraciones permitidas; Δ, el umbral mínimo de la desviación del error en iteraciones sucesivas. Asignar *c* centros iniciales $m_j^{(k)}$ para la iteración *k = 1*.
2. Asignar cada x_i en el conjunto de datos al sector representado por el más cercano $m_j^{(k)}$
3. Calcular para la última clasificación los nuevos centros $m_j^{(k+1)}$ y el error $E^{(k+1)}$.
4. Repetir los pasos 2 y 3 hasta que *k=maxiter* o $\left| E^{(k+1)} - E^{(k)} \right| < \Delta$

Existen variaciones a este algoritmo, dependiendo la elección inicial de los centros, la regla de asignación de patrones, el computo del centro (usualmente calculado como la media del sector) y el criterio de finalización. Normalmente el algoritmo finaliza después de un número pequeño de iteraciones (k < 10). La elección inicial de los centros puede tener una fuerte influencia en la solución final, por lo que se puede realizar de diferentes formas, siendo una de ellas la aproximación estadística o SPSS.