

Tratamiento Estadístico de la Información
Curso de Doctorado en Tecnología de las
Comunicaciones
Universidad Carlos III de Madrid
Derivación del Núcleo de Fisher

Pablo Barrera González
<*pbarrera@tsc.uc3m.es*>

19 de mayo de 2003

Índice

1. Introducción	1
2. Motivación del Núcleo de Fisher	2
3. La función de Núcleo de Fisher	2
4. El Núcleo de Fisher en MOM Discretos	4
4.1. Probabilidades de emisión	5
4.2. Probabilidades de transición	8
5. Comentarios y conclusiones	9

Resumen

Las máquinas de vectores soporte representan, en la actualidad, el estado del arte de los métodos de clasificación. El problema que las mismas presentan es su aplicación sobre el que no se pueda definir, de forma directa, una función de medida, como pueden ser el caso de la utilización de secuencias. Una solución a esta aparente limitación es el núcleo de Fisher. Con él se puede construir una función de medida a partir de un modelo de probabilidad generativo sobre el conjunto de datos a tratar. Este trabajo presenta el funcionamiento del núcleo de Fisher y proporciona las expresiones necesarias para su aplicación en diferentes entornos, discutiendo, a su vez, el significado de las mismas.

1. Introducción

La capacidad de clasificación que proporcionan las máquinas de vectores soporte (*MVS*) está limitada por nuestra habilidad para escoger y diseñar funciones de núcleo adecuadas. En el caso de las secuencias, existe un problema para definir una métrica al considerar secuencias de diferentes tamaños o con pérdidas de fragmentos en las mismas. En la práctica se han conseguido resolver estas trabas en los problemas de clasificación empleando métodos probabilísticos, como pueden ser los Modelos Ocultos de Markov (*MOM*) [1].

Estos métodos de probabilidad generativos resuelven estos problemas, ya que son capaces de asignar una verosimilitud a cualquier secuencia para un determinado modelo (normalmente asociado a una clase de datos). Entrenando adecuadamente los modelos, se pueden construir funciones de discriminación basadas en las medidas de verosimilitud que proporcionan estos. Aún así, el enfoque de los métodos discriminativos, como las *MVS*, construyen fronteras de decisión más flexibles y precisas, lo que se refleja en una buena clasificación y generalización, en muchos casos, mejor que la que proporcionan los métodos basados en modelos probabilísticos. Un clasificador óptimo debería unir estos dos enfoques.

El *núcleo de Fisher* proporciona una manera natural de llevar a cabo esta unión, estableciendo un marco general para extraer una función de núcleo a partir de un modelo de probabilidad generativo. Esta función de núcleo representa la métrica usada entre las muestras, una forma de indicar lo parecidas o distintas que son dos secuencias entre sí.

A continuación, en este texto, se estudiará más detenidamente las motivaciones y la formulación del núcleo de Fisher. Posteriormente, se presentará el desarrollo del vector de puntuaciones de Fisher, eje del cálculo del núcleo de Fisher. El desarrollo se hará empleando Modelos Ocultos de Markov como modelo de probabilidad generativo.

2. Motivación del Núcleo de Fisher

Los modelos de probabilidad generativos son capaces de trabajar con secuencias de tamaño variable y afrontar pérdidas de fragmentos en las mismas. Pueden emplearse incluso para clasificación, construyendo funciones de discriminación. Estas funciones serán óptimas cuando los modelos estén perfectamente ajustados a los conjuntos o clases de secuencias con las que se trate. En un caso general, esta situación no se dará, por ejemplo, por trabajar con conjuntos limitados de muestras.

Las máquinas de vectores soporte están demostrando su valía para multitud de aplicaciones, con unos resultados no alcanzados anteriormente por otros sistemas de clasificación. Dada la generalidad con la que están contruidos estos métodos, puede aplicarse su capacidad a cualquier problema en el que esté definido un *núcleo adecuado*. Lo que no es tan sencillo es cómo encontrar este *núcleo adecuado* para un problema en particular. El principal problema a la hora de encontrar un núcleo es que resulta necesario tener algún tipo de métrica sobre los datos tratados, alguna forma de decir lo cerca o lejos que están dos muestras, en el fondo, lo parecidas que son. En el caso de las secuencias, dicha métrica no está definida, lo que impide emplear las MVS con este tipo de datos. Otras aproximaciones han intentado tratar una secuencia como un vector o extraer de ella una serie de parámetros, sin obtener, en general, buenos resultados. En primer lugar los símbolos que forman la secuencia pueden tener una representación arbitraria (letras, índices, etc) sobre la cual no se puede construir una métrica. Si se aborda la extracción de parámetros, el problema está en saber cuales son los necesarios, tarea, a priori, desconocida.

El núcleo de Fisher proporciona un marco general para unir los dos enfoques anteriores, los modelos de probabilidad generativos con las máquinas de vectores soporte. El objetivo de esta unión es compensar las carencias de cada uno de los enfoques, sacando partido de sus ventajas individuales. El núcleo de Fisher proporciona una función de núcleo y por tanto un tipo de métrica, sobre un conjunto de secuencias, lo que habilita a las máquinas de vectores soporte para trabajar con este tipo de datos. Por otro lado emplea la capacidad discriminativa, que demuestran las MVS, en los problemas en los que la utilización únicamente de los modelos de probabilidad no proporciona resultados aceptables.

Por ésto, el núcleo de Fisher aparece como un método muy interesante, frente a las alternativas antes comentadas, para trabajar con secuencias y máquinas de vectores soporte, cuando la calidad de los métodos basados en modelos es insuficiente.

3. La función de Núcleo de Fisher

El objetivo del *núcleo de Fisher* es extraer (*derivar*) una función de núcleo de un modelo de probabilidad generativo. Ya se ha hablado del interés de este núcleo, pero nada se ha dicho de cómo conseguirlo. La función debe reflejar, de alguna manera, una métrica o relación entre las muestras con las que se trabaje, sean éstas del tipo que sea. La medida $K(X, Y)$ indicará lo cerca o lo lejos que están las muestras X e Y según dicha métrica.

Una función de núcleo puede verse, de manera general, como un producto interno en un espacio de características, sobre el que han sido proyectadas las muestras de entrada [2]:

$$K(X, Y) = \bar{\Phi}(X)^T \Sigma \bar{\Phi}(Y) \quad (1)$$

La forma en la que se defina el producto interno en este espacio, llevará asociada una métrica en el mismo. El núcleo de Fisher extrae esta métrica de un modelo de probabilidad generativo. Además, si nuestro objetivo es la clasificación, los modelos deben incluir la clase como una variable latente [3].

Un modelo de probabilidad generativo proporciona una medida de la verosimilitud de que una muestra haya sido o no generada por él. Dicha verosimilitud se utiliza, en una aproximación tradicional, para buscar cuál es el modelo que proporciona la mayor verosimilitud de entre los disponibles. Con esta información se puede construir un clasificador de tipo *MAP* (*máximo a posteriori*). Al contrario que en esta aproximación, lo que interesa ahora, para buscar una métrica entre muestras, es encontrar las diferencias en el proceso de generación. Lo lejos o cerca que están del modelo (verosimilitud) no proporciona información sobre si se parecen entre sí. Dos valores de verosimilitud parecidos pueden corresponder a dos secuencias radicalmente distintas.

Una forma de capturar el proceso de generación es emplear el gradiente del logaritmo de la verosimilitud, con respecto a los parámetros del modelo. Cada una de las derivadas parciales del gradiente indicará cómo afecta un parámetro en particular al cálculo de la verosimilitud para una muestra. El espacio del gradiente tiene la ventaja de que además preserva todas las suposiciones que contiene el modelo sobre el proceso de generación de muestras.

Considérense todos los posibles modelos, definidos por el vector de parámetros $\bar{\theta}$, tomando de entre todos los parámetros posibles, Θ :

$$P(X|\bar{\theta}); \quad \bar{\theta} \in \Theta \quad (2)$$

Esta clase de modelos de probabilidad definen una variedad de Riemann, M_Θ , con una métrica local dada por la matriz de información de Fisher, F [4]:

$$F = E_X\{\bar{U}_X \bar{U}_X^T\} \quad (3)$$

La esperanza se realiza sobre los vectores \bar{U}_X , conocidos como puntuaciones de Fisher:

$$\bar{U}_X = \nabla_{\bar{\theta}} \log P(X|\bar{\theta}) \quad (4)$$

Para simplificar se ha eliminado la dependencia de F y \bar{U}_X de los parámetros $\bar{\theta}$, o lo que es lo mismo, de la posición en la variedad.

La métrica local M_Θ define una distancia entre el modelo actual, $P(X|\bar{\theta})$ y un modelo cercano, $P(X|\bar{\theta} + \bar{\delta})$. Esta distancia vendrá dada por $D(\bar{\theta}, \bar{\theta} + \bar{\delta}) = 1/2 \bar{\delta}^T F \bar{\delta}$, que también aproxima la distancia *Kullback-Leibler* o divergencia asimétrica [5] entre dos modelos para un $\bar{\delta}$ suficientemente pequeño.

La puntuación de Fisher, $\bar{U}_X = \nabla_{\bar{\theta}} \log P(X|\bar{\theta})$ proyecta una muestra X en un punto del espacio del gradiente de la variedad M_Θ , vector \bar{U}_X . A esto se le conoce como *proyección de la puntuación de Fisher*. \bar{U}_X también puede usarse para definir la dirección de máxima pendiente en la función $\log P(X|\bar{\theta})$ para una muestra X en la variedad M_Θ . Este gradiente se conoce como el *gradiente natural* [4] y se obtiene sustituyendo $\bar{\phi}_X$ por $F^{-1} \bar{U}_X$. La función de proyección $X \mapsto \bar{\phi}_X$ es la *proyección natural* de las muestras en los vectores de características. El producto

interno en el espacio de $\bar{\phi}$ relativo a la métrica local de Rieman define el núcleo natural, que puede expresarse en función de los vectores de puntuación de Fisher de la siguiente forma:

$$K(X_i, X_j) \propto \bar{\phi}_{X_i}^T F \bar{\phi}_{X_j} = (F^{-1} \bar{U}_{X_i})^T F (F^{-1} \bar{U}_{X_j}) = \bar{U}_{X_i}^T F^{-1} \bar{U}_{X_j} \quad (5)$$

Éste es el llamado núcleo de Fisher. Su nombre se debe al importante papel que juegan las puntuaciones de Fisher en su construcción. Las propiedades del mismo se pueden encontrar en [3].

Las versiones cuadráticas, cúbicas, etc, del núcleo de Fisher también se pueden manejar para un caso no lineal:

$$\hat{K}(X, Y) = (1 + K(X, Y))^p \quad (6)$$

$$(7)$$

Incluso se puede trabajar con núcleo gaussianos recurriendo a una medida de distancia natural entre vectores [6]. Ésta está dada por:

$$D^2(X, Y) = \frac{1}{2} (\bar{U}_X - \bar{U}_Y)^T F^{-1} (\bar{U}_X - \bar{U}_Y) \quad (8)$$

Lo que permite emplear el siguiente núcleo gaussiano:

$$K(X, Y) = e^{-D^2(X, Y)} \quad (9)$$

Una demostración más completa de las bases de la derivación de una función de núcleo a partir de un modelo de probabilidad generativo puede encontrarse en [3] y [6].

4. El Núcleo de Fisher en MOM Discretos

A la hora de aplicar el núcleo de Fisher en un problema concreto, es necesario calcular los vectores de puntuación, llamados hasta ahora \bar{U}_X . La formulación para extraer estos vectores de puntuación dependerá de modelo probabilístico que estemos usando en particular, así como de los parámetros del mismo.

En el caso que nos interesa, trabajaremos con modelos ocultos de Markov discretos ([1]). Los modelos se representan como una serie de estados en los que se permite emitir uno de entre una serie de símbolos posibles. Para definir estos modelos se emplean los siguientes parámetros:

- La probabilidad de saltar a un estado, q_j , dado que estás en el estado q_i , $a_{q_i q_j}$ (matriz A).
- La probabilidad en emitir un símbolo o dado que estás en el estado q_j , $b_{q_j}(o)$ (matriz B).
- La probabilidad de encontrarte en el estado q_j al inicio de la secuencia, π_{q_j} (vector π).

Para simplificar nos referimos al conjunto de todos estos parámetros como λ , $\lambda = [A, B, \pi]$. Calcular el vector de puntuaciones consiste únicamente en obtener el gradiente del logaritmo de la verosimilitud de una secuencia con respecto a estos parámetros.

$$\bar{U}_X = \nabla_{\lambda} \log P(O|\lambda) \quad (10)$$

Para calcular los vectores de puntuación de Fisher deberá calcularse el gradiente del logaritmo de la verosimilitud de cada secuencia con respecto a todos los parámetros del modelo, o por lo menos con respecto a aquellos que influyan de manera importante en el proceso de generación. Aunque en [3] y [6] sólo se consideran como parámetros del modelo las probabilidades de emisión, aquí vamos a presentar el desarrollo del vector de puntuaciones tanto para estas como para las probabilidades de transición.

Para el caso de los MOM discretos, la verosimilitud para una secuencia, $O = o_1, o_2, \dots, o_l$, cualquiera, dado el modelo λ , se puede calcular sin más que:

$$P(O|\lambda) = \sum_{Q_j} \prod_i P(o_i|q_i, B)P(q_i|q_{i-1}, A) = \sum_{Q_j} \prod_i b_{q_i}(o_i)a_{q_{i-1}q_i} \quad (11)$$

donde $Q_j, Q_j = Q_1, \dots, Q_n$, representa todos los posibles caminos que puede seguir la secuencia en su proceso de generación dentro del modelo. En cada uno de estos caminos, la probabilidad de que se genere la secuencia de observaciones O en particular, y no otra cualquiera, se calcula como la probabilidad de haber estado en todos los estados de la secuencia Q_j , y haber emitido, en cada uno de ellos, el símbolo o_i , $\prod_i b_{q_i}(o_i)a_{q_{i-1}q_i}$.

El desarrollo del gradiente se puede separar, en dos partes, la dependiente de las probabilidades de emisión y la dependiente de las probabilidades de transición. No hemos considerado el vector de parámetros π , puesto que en general se suele fijar un estado del modelo como *inicial*. A continuación se desarrollará el gradiente para cada una de ellas. El vector de puntuaciones final se formará tomando las componentes de cada una de las partes.

4.1. Probabilidades de emisión

En primer lugar se va a desarrollar las componentes del vector de puntuaciones de Fisher que se derivan a partir de las probabilidades de emisión del modelo. Por cada una de las probabilidades obtendremos una componente del vector de puntuaciones. Cada una de éstas, será la derivada parcial, con respecto a dicho parámetro en particular, del logaritmo de la verosimilitud de la secuencia:

$$\bar{U}_O(\tilde{q}, \tilde{o}) = \frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \log P(X|\lambda) \quad (12)$$

Comenzamos a derivar esta expresión por el logaritmo:

$$\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \log P(O|\lambda) = \frac{1}{P(O|\lambda)} \frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} P(O|\lambda) \quad (13)$$

A continuación, podemos sustituir $P(O|\lambda)$ por la expresión obtenida del modelo,

$$\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \log P(O|\lambda) = \frac{1}{P(O|\lambda)} \frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \sum_{Q_j} \prod_i b_{q_i}(o_i)a_{q_{i-1}q_i} \quad (14)$$

$$= \frac{1}{P(O|\lambda)} \sum_{Q_j} \frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \prod_i b_{q_i}(o_i)a_{q_{i-1}q_i} \quad (15)$$

Al cambiar la posición de la derivada parcial con la del sumatorio, podemos, para simplificar, abordar el cálculo de ésta únicamente. La parcial del productorio quedará como sigue:

$$\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} = \sum_k \left[\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} b_{q_k}(o_k) \right] a_{q_{i-k}q_k} \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (16)$$

Aquí es necesario tener en cuenta que los parámetros b no son independientes entre sí. Todos ellos están sujetos a la condición de que las probabilidades de emisión para todos los símbolos en un estado deben sumar 1:

$$\sum_n b_q(o_n) = 1 \quad (17)$$

Obviamente, viendo la expresión anterior, podemos escribir una probabilidad de emisión en función de las demás de la siguiente forma:

$$b_q(\tilde{o}) = 1 - \sum_{n/o_n \neq \tilde{o}} b_q(o_n) \quad (18)$$

Considerando esto, las derivadas parciales de antes quedarán de la siguiente forma:

$$\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} b_q(o) = \begin{cases} 1, & \text{si } \tilde{q} = q, \tilde{o} = o \\ -1, & \text{si } \tilde{q} = q, \tilde{o} \neq o \\ 0, & \text{si } \tilde{q} \neq q \end{cases} \quad (19)$$

Podemos, ahora, substituir la derivada parcial obtenida en (16), con lo que dicha expresión queda como:

$$\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} = \sum_k \left[\frac{\partial}{\partial b_{\tilde{q}}(\tilde{o})} b_{q_k}(o_k) \right] a_{q_{i-k}q_k} \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (20)$$

$$= \sum_k [2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} - \delta_{q_k, \tilde{q}}] a_{q_{i-k}q_k} \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (21)$$

$$= \sum_k \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} - \delta_{q_k, \tilde{q}}}{b_{q_k}(o_k)} b_{q_k}(o_k) a_{q_{i-k}q_k} \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (22)$$

$$= \sum_k \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} - \delta_{q_k, \tilde{q}}}{b_{q_k}(o_k)} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (23)$$

$$(24)$$

El resultado anterior se introduce ahora en la expresión (15):

$$\bar{U}_O = \frac{1}{P(O|\lambda)} \sum_{Q_j} \sum_k \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} - \delta_{q_k, \tilde{q}}}{b_{q_k}(o_k)} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (25)$$

El orden de los sumatorios se puede cambiar y separar en dos partes la división:

$$\bar{U}_O = \frac{1}{P(O|\lambda)} \left(\sum_k \sum_{Q_j} \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}}}{b_{\tilde{q}}(\tilde{o})} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} - \sum_k \sum_{Q_j} \frac{\delta_{q_k, \tilde{q}}}{b_{\tilde{q}}(o_k)} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} \right) \quad (26)$$

De esta forma tenemos dos miembros diferenciados, uno que depende de $\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}}$ y otro que sólo depende de $\delta_{q_k, \tilde{q}}$. El productorio $\prod_i b_{q_i}(o_i) a_{q_{i-1}q_i}$ es igual a la probabilidad de la observación para un camino dado el modelo, o lo que es lo mismo, $P(O|Q_j, \lambda)$.

$$\bar{U}_O = \frac{1}{b_{\tilde{q}}(\tilde{o})} \sum_k \sum_{Q_j} 2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} \frac{P(O|Q_j, \lambda)}{P(O|\lambda)} - \sum_k \frac{1}{b_{\tilde{q}}(o_k)} \sum_{Q_j} \delta_{q_k, \tilde{q}} \frac{P(O|Q_j, \lambda)}{P(O|\lambda)} \quad (27)$$

Los cocientes de probabilidad sumados para todos los posibles caminos del modelo y ponderados por las deltas, no son más que las esperanzas de estar en un instante en un estado emitiendo un símbolo en concreto.

$$\bar{U}_O = \frac{1}{b_{\tilde{q}}(\tilde{o})} \sum_k 2E\{\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} | O, \lambda\} - \sum_k \frac{1}{b_{\tilde{q}}(o_k)} E\{\delta_{q_k, \tilde{q}} | O, \lambda\} \quad (28)$$

$$= \frac{2\xi(\tilde{q}, \tilde{o})}{b_{\tilde{q}}(\tilde{o})} - \xi'(\tilde{q}) \quad (29)$$

donde,

$$\xi(\tilde{q}, \tilde{o}) = \sum_k E\{\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} | O, \lambda\} \quad (30)$$

$$\xi'(\tilde{q}) = \sum_k \frac{1}{b_{\tilde{q}}(o_k)} E\{\delta_{q_k, \tilde{q}} | O, \lambda\} \quad (31)$$

$\xi(\tilde{q}, \tilde{o})$ representa el número de veces medio que hemos estado en un estado determinado emitiendo un símbolo en concreto. $\xi'(\tilde{q})$ será el número medio de veces que se está en cada estado ponderado por la probabilidad de la observación en cada estado.

Estos valores pueden calcularse a partir de las matrices α y β del algoritmo hacia delante y hacia atrás [1].

4.2. Probabilidades de transición

Se puede repetir este desarrollo teniendo en cuenta, únicamente, las probabilidades de transición. Partimos de la misma expresión que antes cambiando únicamente la parcial:

$$\bar{U}_O(\tilde{q}, \tilde{\sigma}) = \frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} \log P(O|\lambda) \quad (32)$$

Al igual que en el caso anterior se puede introducir la expresión (11), y desarrollar la parcial. De esta forma se obtiene:

$$\frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} \log P(O|\lambda) = \frac{1}{P(O|\lambda)} \sum_{Q_j} \frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (33)$$

Resolviendo solo la parcial se llega a:

$$\frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} \prod_i b_{q_i}(x_i) a_{q_{i-1}q_i} = \sum_k \left[\frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} a_{q_{i-k}q_k} \right] b_{q_k}(o_k) \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (34)$$

$$(35)$$

De nuevo puede verse que éstas no son independientes entre sí, ya que la suma de las probabilidades de transición en cada estado debe ser 1, estableciendo una dependencia entre las derivadas parciales, como pasaba en la expresión (19), con lo que podemos sustituir la derivada de la siguiente forma:

$$\frac{\partial}{\partial a_{\tilde{q}\tilde{q}'}} \prod_i b_{q_i}(x_i) a_{q_{i-1}q_i} = \sum_k [2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{\sigma}} - \delta_{q_k, \tilde{q}}] b_{q_k}(o_k) \prod_{i \neq k} b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (36)$$

$$= \sum_k \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{\sigma}} - \delta_{q_k, \tilde{q}}}{a_{q_{i-k}q_k}} \prod_i b_{q_i}(o_i) a_{q_{i-1}q_i} \quad (37)$$

$$(38)$$

Volviendo a introducir este valor en la expresión original se obtiene:

$$U_O = \frac{1}{P(O|\lambda)} \sum_{Q_j} \sum_k \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} - \delta_{q_k, \tilde{q}}}{a_{q_k-1q_k}} \prod_i b_{q_i}(x_i) a_{q_i-1q_i} \quad (39)$$

$$= \frac{1}{P(O|\lambda)} \left(\sum_k \sum_{Q_j} \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}}}{a_{\tilde{q}\tilde{q}'}} \prod_i b_{q_i}(o_i) a_{q_i-1q_i} - \sum_k \sum_{Q_j} \frac{\delta_{q_k, \tilde{q}}}{a_{\tilde{q}q_k}} \prod_i b_{q_i}(o_i) a_{q_i-1q_i} \right) \quad (40)$$

$$= \frac{1}{a_{\tilde{q}\tilde{q}'}} \sum_k \sum_{Q_j} \frac{2\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}}}{P(O|\lambda)} P(O|Q_j, \lambda) - \sum_k \frac{1}{a_{\tilde{q}q_k}} \sum_{Q_j} \frac{\delta_{q_k, \tilde{q}}}{P(O|\lambda)} P(O|Q_j, \lambda) \quad (41)$$

$$= \frac{1}{b_{\tilde{q}\tilde{q}'}} \sum_k 2E\{\delta_{q_k, \tilde{q}} \delta_{x_k, \tilde{o}} | O, \lambda\} - \sum_k \frac{1}{a_{\tilde{q}q_k}} E\{\delta_{q_k, \tilde{q}} | O, \lambda\} \quad (42)$$

$$= \frac{2\tau(\tilde{q}, \tilde{o})}{a_{\tilde{q}\tilde{q}'}} - \tau'(\tilde{q}) \quad (43)$$

donde,

$$\tau(\tilde{q}, \tilde{o}) = \sum_k E\{\delta_{q_k, \tilde{q}} \delta_{o_k, \tilde{o}} | O, \lambda\} \quad (44)$$

$$\tau'(\tilde{q}) = \sum_k \frac{1}{b_{\tilde{q}}(o_k)} E\{\delta_{q_k, \tilde{q}} | O, \lambda\} \quad (45)$$

5. Comentarios y conclusiones

Antes de proceder a revisar que posibles conclusiones

Para concluir este trabajo vamos a presentar una serie de conclusiones obtenidas tras estudiar tanto la formulación del núcleo de Fisher como algunos de los artículos que lo desarrollan.

En primer lugar recordemos que la función del núcleo de Fisher se presenta de la siguiente forma:

$$K(X, Y) = \bar{U}_X F^{-1} U_Y \quad (46)$$

Aunque con la proyección de la puntuación de Fisher obtenemos vectores, la función de núcleo incluye una matriz de regularización entre ellos. En algunos desarrollos se elimina, como simplificación [3] o se consideran únicamente los valores de su diagonal. En según que casos dicha aproximación puede no ser válida. En general los vectores de puntuación de Fisher deberán ser regularizados antes de poder emplearse dentro de una función de núcleo. Si no se procede así puede que ciertas componentes sean mucho más importantes que otras, a la hora de evaluar

la función, aunque la información que contengan no sea la más representativa. También hay que considerar que sí existe correlación entre las componentes del vector, por lo que la aproximación de la matriz de información de Fisher por su diagonal resulta incompleta. El problema de abordar el cálculo de la matriz de información es su dimensión, en general alta, tanto para calcularla como para invertirla. Una posible solución es aproximar F por una matriz diagonal por bloques, dado que las componentes del vector solo guardan una relación importante con las que derivan del mismo estado.

Para el cálculo del vector de puntuaciones de Fisher hemos introducido también las probabilidades de transición del modelo de Markov. Hasta ahora no he encontrado referencias acerca de resultados obtenidos empleando estas componentes del vector de puntuaciones. En general se ha demostrado experimentalmente que la matriz de probabilidades de transición no tiene gran importancia en los problemas de discriminación. Variaciones en esta matriz que respeten la topología del modelo no se reflejan en cambios importantes en los errores de clasificación. Aún así no, esta demostración no tiene una base matemática suficientemente clara, por lo que no parece apropiado desechar las probabilidades de transición tan a la ligera, sobre todo teniendo en cuenta que el problema que queremos resolver aquí es ligeramente diferente al que hemos citado anteriormente. En nuestro caso lo que estamos buscando es encontrar diferencias en el proceso de generación entre un par de muestras dadas.

En cuanto al cálculo respecto a las probabilidades de emisión, pueden verse diferencias con el presentado en [6]. En este artículo Jaakkola parte de una suposición, a mi modo de ver, muy restrictiva. Él asume que $\sum_o P(o, q) = 1$. Lo único que podemos suponer de una forma general es $\sum_o P(o|q) = 1$. Por este motivo aquí se ha presentado un desarrollo completo del vector de puntuaciones de Fisher, sin ninguna suposición adicional.

Me tomaré la libertad, por último, de criticar la nomenclatura que he empleado en este texto para hacer los desarrollos. Pues si bien es la más comúnmente empleada en las publicaciones sobre Modelos Ocultos de Markov, por lo menos en aquellas que los aplican para procesado de habla, la encuentro especialmente farragosa y poco intuitiva. Aún así, decidí emplearla precisamente por ser la más comúnmente empleada, con la esperanza de simplificar, de esta forma, la comprensión a posibles lectores acostumbrados a tratar con MOM.

Referencias

- [1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Febrero 1989.
- [2] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, 1998.
- [3] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," 1998.
- [4] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.
- [5] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

- [6] T. Jaakkola, M. Diekhans, and D. Haussler, “A discriminative framework for detecting remote protein homologies,” 1998.