Ejercicios PEE II (Multimodal Interaction in PR)

Autor: Raúl García Crespo

1.Briefly explain the differences between Classification and Structured Output Prediction. Cite two application examples each paradigm.

Una de las diferencias entre Clasificación y Predicción Estructurada Estadística (PEE) son los tipos de hipótesis que pueden ser generadas por cada uno de los tipos de marco para el reconocimiento de patrones. En el caso de la Clasificación, las posibles hipótesis de salida del sistema toman valores finitos e independientes, es decir:

$$H = \{1, \dots, C\}$$

Donde H es el conjunto de posibles hipótesis y C el número de posibles valores.

Mientras que en el caso de PEE, las hipótesis están estructuradas en forma de secuencia o grafo a partir de un conjunto de valores tomados de un espacio H posiblemente infinito. Este agrupamiento en secuencias se debe a que las distintas hipótesis poseen una dependencia entre ellas que es necesario tener en cuenta.

Otra diferencia la encontramos en la definición de los modelos para resolver estos dos tipos de problemas. Para el caso de la Clasificación, basta con resolver de forma trivial y exhaustiva:

$$h = \underset{h \in H}{\operatorname{argmax}} P(h|x)$$

Donde *x* representa una muestra de entrenamiento a clasificar.

Sin embargo para el caso de PEE esta fórmula puede ser difícil de estimar directamente y suele ser recomendable aplicar Bayes para descomponerla de la siguiente forma:

$$h = \underset{h \in H}{\operatorname{argmax}} P(x|h) \cdot P(h)$$

Donde $P(x|h)$ sería la verosimilitud del modelo, la cual es fácilmente estimada a partir de pares de entrenamiento muestra/etiqueta; y donde $P(h)$ es la probabilidad a priori que pude ser estimada empleando de nuevo las muestras de entrenamiento.

La Clasificación tiene aplicaciones en la clasificación de imágenes etiquetadas, como las que podemos encontrar en competiciones como ImagNet; o en biometría, donde por ejemplo se emplea para reconocer a una persona a partir de una huella dactilar.

Por otro lado, PEE tiene aplicaciones en el reconocimiento de automático del habla, donde cada palabra detectada por el modelo depende de las anteriores y/o las posteriores; o en el etiquetado sintáctico de oraciones, donde las etiquetas deben aparecer en secuencias concretas.

2.Justify why the naive Bayes decomposition of Eq.(5) is adequate for karyotype recognition problem.

El problema del reconocimiento de cariotipo nos pide que, a partir de un set de imágenes desordenadas las cuales representan los cromosomas no sexuales de un cariotipo, etiquetemos cada una de las imágenes con etiquetas del 1 al 22 en base al cromosoma que representan. En conjunto, se generará una hipótesis en forma de secuencia de 22 etiquetas:

$$h = h_1^{22} \in H$$

Siendo *H* el conjunto de todas las posibles combinaciones de las 22 etiquetas.

La aproximación propuesta por naive Bayes para resolver la verosimilitud de este problema es la siguiente:

$$P(x|h) = P\left(x_1, x_2 \dots, x_{22} \middle| h_1, h_2, \dots, h_{22}\right) \approx \prod_{i=1}^{22} P(x_i|h_i)$$

Esto puede justificarse dado que si aplicamos la regla de la cadena:

$$P(x|h) = P(x_1|h) \cdot P(x_2|x_1, h) \dots \cdot P(x_{22}|x_1, .., x_{21}, h)$$

Donde la probabilidad de cada elemento de *x* depende de la hipótesis y de todos los valores de *x* anteriores.

El problema del reconocimiento de cariotipo puede interpretarse como un problema de clasificación de imágenes, donde es necesario etiquetar cada una de las imágenes desordenadas representando un cromosoma con la etiqueta de su cromosoma correspondiente. Es posible realizar esta asunción dado que la forma de un cromosoma no depende en ningún modo de la forma del resto de cromosomas, por lo que solo es necesaria su imagen para reconocerlo. Una vez obtenidas las hipótesis de etiqueta asociadas a cada imagen, estas se agrupan en la hipótesis asociada al cariotipo. De esta forma se asume que la hipótesis h_i asociada a cada x_i es independiente del resto de valores de h y x, y por lo tanto:

$$P(x|h) = P(x_1|h_1) \cdot P(x_2|h_2) \dots \cdot P(x_{22}|h_{22}) = \prod_{i=1}^{22} P(x_i|h_i)$$

3.Briefly explain all the steps and assumptions needed to derive Eq.(9) from Eq.(7).

Cuando tenemos en cuenta la historia pasada y el feedback del usuario en la generación de nuevas hipótesis, la hipótesis más probable puede obtenerse con:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} P(h|x, h', f)$$

Donde h' representa los valores pasados de h y f el feedback del usuario.

Aplicando el teorema de Bayes obtenemos que:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} P(h|x, h', f) = \underset{h \in H}{\operatorname{argmax}} \frac{P(h, x, h', f)}{P(x, h', f)}$$

Dado que el denominador no depende de h puede extraerse de la fórmula:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} P(h, x, h', f)$$

Y aplicando la regla de la cadena obtenemos:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} P(h, x, h', f) = \underset{h \in H}{\operatorname{argmax}} P(h')P(f|h')P(h|f, h')P(x|f, h', h)$$

Si sustituimos el feedback por su decodificación y extraemos de la fórmula los elementos que de nuevo no dependen de h:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} \; P(h|d, h')P(x|d, h', h)$$

Con lo que se consigue la fórmula de la Eq.(8) de las transparencias. Si ahora consideramos que la probabilidad $P(x|d, h', h)$ es independiente de d y h', dado que el propósito de d es alterar un elemento o una parte de h' para generar una nueva h y no afectan de manera directa a x entonces:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} \; P(h|d, h')P(x|h)$$

4. Write a C (or pseudo-code) algorithm that implements a greedy solution to the interactive (pasive, left-to-right) search problem discussed in pages 23-27 (see the basic idea in page 18).

Asumiendo que tenemos acceso a una matriz de verosimilitud para cada muestra de test P [] donde $P[i,c] = P(x_i|c)\ para\ 1 \leq i \leq 22\ y\ 1 \leq c \leq 22$ podemos generar el siguiente pseudo-código para implementar el algoritmo greedy sobre el problema de búsqueda interactiva left-to-right:

//Variables conteniendo el feedback (posición del primer error y su corrección) y la hipótesis //pasada (el último cariotipo propuesto)

$feedback, lastKario;$

//Extraemos la posición y la corrección del primer error por la izquierda

$c, i = feedback;$

//Corregimos la posición i de la última hipótesis según la información del feedback

$newKario = lastKario;$

$newKario[i] = c;$

//Guardamos las etiquetas usadas previas a la posición corregida por el usuario

$usedLabels = newKario[1, \dots, i];$

//Aplicamos el algoritmo greedy sobre el resto del cariotipo

$\text{bestScores} = [];$

$for\ j\ in\ [i+1, .., 22]:$

$$c_j = \max_{c \in \{1,..,22\}} \frac{P(x_j|c) \cdot P(c)}{P(x_j)};$$

$$bestScores.append\left(\left(c_j, j\right)\right);$$

//Ordenamos los scores

$sort(bestScores);$

//En función del orden de los scores, clasificamos las imágenes restantes utilizando la matriz de verosimilitud

$for\ c, i\ in\ bestScores:$

$$c = \underset{k \in \{"1",\dots,"22"\} - usedLabels}{\text{argmax}} P[i,k];$$

$usedLabels.append(c);$

$newKario[i] = c;$

$return\ newKario$

Con este pseudo-código obtenemos en la variable "*newKario*" un array de 22 elementos donde cada posición contiene la etiqueta asignada a la imagen que se encuentra en esa misma posición.

6. Briefly explain all the steps and assumptions needed to derive Eq.(19) from Eq.(7).

Partiendo de nuevo de la expresión:

$$\hat{\mathrm{h}} = \underset{h \in H}{\operatorname{argmax}} P(h|x, h', f) \quad (1)$$

En este caso la decodificación de f no es determinista y debe incluirse como una nueva variable oculta:

$$\hat{\mathrm{h}} = \underset{h \in H}{\operatorname{argmax}} P(h|x, h', f) = \underset{h \in H}{\operatorname{argmax}} \sum_{d} P(h, d|x, h', f) \quad (2)$$

Usando la aproximación por moda que encontramos en página 1 de las transparencias:

$$\sum_{x} \mathrm{P}(\mathrm{x}, \dots) \approx \max_{x} P(x, \dots)$$

Podemos pasar a la expresión:

$$\hat{\mathrm{h}} = \underset{h \in H}{\operatorname{argmax}} \sum_{d} P(h, d|x, h', f) \approx \underset{h \in H}{\operatorname{argmax}} \max_{d} P(h, d|x, h', f) \quad (3)$$

Y aplicando Bayes:

$$P(h, d|x, h', f) = \frac{P(h, d, x, h', f)}{P(x, h', f)} \quad (4)$$

Al sustituir la expresión (4) en (3) el denominador no depende de h ni d por lo que se puede omitir:

$$\hat{\mathrm{h}} = \underset{h \in H}{\operatorname{argmax}} \max_{d} P(h, d|x, h', f) = \underset{h \in H}{\operatorname{argmax}} \max_{d} P(h, d, x, h', f) \quad (5)$$

Aplicando la regla de la cadena y omitiendo los términos que no dependen de las variables ocultas:

$$\hat{\mathrm{h}} = \underset{h \in H}{\operatorname{argmax}} \max_{d} P(h, d, x, h', f) =$$

$$= \underset{h \in H}{\operatorname{argmax}} \max_{d} P(h') \cdot P(d|h') \cdot P(f|d, h') \cdot P(h|f, d, h') \cdot P(x|h, f, d, h') =$$

$$= \underset{h \in H}{\operatorname{argmax}} \max_{d} P(d|h') \cdot P(f|d, h') \cdot P(h|f, d, h') \cdot P(x|h, f, d, h') \quad (6)$$

Asumiendo las siguientes independencias:

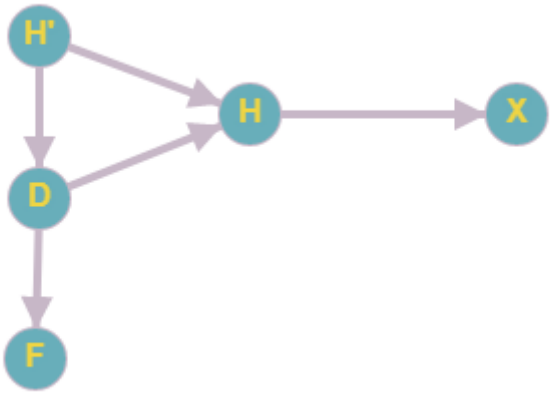
- $P(f|d,h')$ solo depende de d. La decodificación ya contiene la información de la hipótesis pasada, bloqueando la dependencia con h' .
- $P(h|f,d,h')$ solo depende de d y h'. La decodificación ya contienen la información del feedback, al ser una forma derivada de este; pero la hipótesis pasada sí que es necesaria para generar la nueva ya que aplicaremos los cambios propuestos por el feedback sobre ella.
- $P(x|h,f,d,h')$ solo depende de h, que ya posee la información de la decodificación, el feedback y la historia pasada al haber sido generada en base a todos ellos.

Obtenemos la expresión final:

$$\hat{h} = \underset{h \in H}{\operatorname{argmax}} \max_{d} P(d|h') \cdot P(f|d,h') \cdot P(h|f,d,h') \cdot P(x|h,f,d,h') \approx$$

$$\approx \underset{h,d}{\operatorname{argmax}} P(d|h') \cdot P(f|d) \cdot P(h|d,h') \cdot P(x|h)$$

Las independencias asumidas para este último paso pueden visualizarse al expresar el modelo como la siguiente red Bayesiana:



7. Briefly explain under which conditions the solution given by Eq.(22-23) may be optimal. Do the same conditions hold for the optimality of the solution given by Eq.(20-21)? Why? Use the karyotyping example to illustrate your (otherwise general) responses.

Las expresiones:

$$\{d_1, \dots, d_n\} = \text{n} - \underset{d}{\text{best}}\, P(f|d) \cdot P(d|h')$$

$$\hat{\text{h}} \approx \underset{h}{\text{argmax}} \max_{1 \le i \le n} P(f|\hat{d}_i) \cdot P(\hat{d}_i|h') \cdot P(x|h) \cdot P(h|\hat{d}_i, h')$$

Proponen una alternativa al cálculo de $(\hat{h}, \hat{d})$ que no exige una optimización conjunto de la expresión obtenido al final del ejercicio 6. Con la primera de estas 2 expresiones se obtienen las n decodificaciones más probables del feedback y con la segunda obtenemos $\hat{h}$ en función de estos valores.

Al obtener $\hat{h}$ en base a un subconjunto de n posibles valores de d, la solución óptima se obtendrá cuando n sea igual al número total de decodificaciones que puede adoptar como valor d. En el problema de reconocimiento de cariotipo, la decodificación del feedback se representa en forma de dos valores: la primera posición de una etiqueta incorrecta y el valor de la etiqueta correcta. Por lo tanto, en total, la decodificación podrá tomar tantas formas posibles como combinaciones de los dos valores que la componen. Dado que existen 22 posibles posiciones para las etiquetas y tenemos 22 tipos de etiquetas, el número de posibles combinaciones es de $22 \times 22 = 22^2$; por lo tanto cuando n sea igual 22^2, el valor de $\hat{\text{h}}$ obtenido a partir de estas expresiones será óptimo.

Para las expresiones:

$$\hat{d} = \underset{d}{\text{argmax}}\, P(f|d) \cdot P(d|h')$$

$$\hat{h} \approx \underset{h}{\text{argmax}}\, P(x|h) \cdot P(h|h', d)$$

Se obtiene primero el valor “óptimo” de la decodificación y a partir de ese valor se obtiene la hipótesis óptima, sin embargo puede existir una combinación de valores de d y h que genere una hipótesis óptima mejor pero que no se explore porque el valor necesario de d no se ha obtenido en la primera expresión por no ser “óptimo”. No es posible aplicar las mismas condiciones que las aplicadas a las expresiones anteriores para garantizar que la hipótesis siempre sea la óptima y además no existe ninguna otra forma de garantizar que esto ocurra, ya que la hipótesis óptima solo se conseguirá en aquellos casos concretos en los que el valor de d en la primera expresión sea parte de la optimización conjunta óptima de $(\hat{h}, \hat{d})$.

8. Briefly explain the concepts and main differences between Active and Passive interaction protocols.

En un protocolo de interacción, una vez el sistema ha generado una hipótesis, el usuario debe interaccionar con ella, corrigiendo aquellos elementos mal clasificados y dando instrucciones al sistema para corregirlos La pasividad de un protocolo de interacción se define en función de quien decide (el usuario o el sistema) que muestras de la secuencia de la hipótesis necesitan ser revisadas y corregidas.

En el caso de la interacción **pasiva** es el usuario el encargado de decidir qué elementos corregir. Este tipo de protocolo garantiza un resultado óptimo ya que, suponiendo que el usuario posee una información perfecta del resultado y no se equivoca, deberá supervisar cada elemento de la hipótesis y corregir aquellos erróneos hasta que no haya ninguno.

Existen dos variantes de la interacción pasiva:

- En **left-to-right** el sistema ordena los elementos que componen la hipótesis de acuerdo a la probabilidad max-posterior, de manera que el usuario tendrá que supervisar dichos elementos en ese mismo orden, deteniéndose cada vez que encuentre un error y comunicándolo al sitema para elaborar una nueva hipótesis. En esta variante, el usuario en cada iteración solo debe supervisar todos los elementos previos a aquel en el que se ha detectado un error.
- En **Desultory** el usuario supervisa siempre todos los elementos de la hipótesis y decide cuál de ellos, si es que hay alguno, es el que necesita ser corregido con mayor urgencia, independientemente de su orden en la hipótesis.

Como ya se ha dicho ambas variantes del protocolo pasivo continúan hasta que la hipótesis generada no contenga errores, obteniendo un resultado “perfecto”.

Para el caso de protocolos de interacción **activa**, es el sistema quien decide que muestras necesitan ser revisadas y, si corresponde, corregidas. Para ello, cuando el sistema genera una hipótesis asigna un valor de confianza a cada elemento de esta. En base a estos valores, el sistema selecciona aquel el que tiene menor confianza para que el usuario lo supervise. El protocolo pasivo finaliza cuando el valor de confianza para cada elemento está por encima de un umbral.

En este tipo de protocolos no se garantizan resultados perfectos, pues el sistema puede generar una hipótesis equivocada pero haber asignado un valor de confianza alto al elemento erróneo, por lo que nunca se le propone al usuario que lo corrija. A cambio de este posible error en la precisión final se alivia la carga del usuario, ya que no es necesario que revise todos los elementos de la hipótesis, solo aquellos propuestos por el sistema.