Pontificia Universidad Católica de Chile
Escuela de Ingeniería, Depto. de Ciencia de la Computación
IIC-2612 Inteligencia Artificial

Noviembre 2013

Interrogación 3

Tiempo: 100 minutos, SIN APUNTES

Nombre:_____

1. (*20 puntos*) Responda las siguientes preguntas fundamentando brevemente su respuesta.

a. En clases vimos un método para podar un árbol de decisión utilizando un set de validación, donde la búsqueda por nodos a podar partía desde los nodos hojas hacia arriba en el árbol. Otra posibilidad es podar el árbol partiendo desde el nodo raíz hacia abajo. Asumiendo que en ambos casos se usa el mismo criterio y métrica para decidir si un nodo es podado, ¿ambas estrategias de poda deberían arrojar el mismo árbol?.

Falso, es posible que la búsqueda en una rama del árbol partiendo desde la hoja o la raíz se detenga en un nodo intermedio distinto. En general, la estrategia top-down de poda (partiendo de la raíz) produce árboles más pequeños, i.e., la poda es más agresiva que la poda bottom-up (partiendo desde nodo hoja).

b. Un amigo le indica que tiene un muy buen algoritmo de clasificación, el cual entrega un rendimiento de 100 % en el set de entrenamiento, ¿Qué comentaría a su amigo?

Comentaría que la medición del rendimiento, o capacidad de generalizar de un algoritmo de clasificación, debe utilizar un set independiente, o de test, y que no es adecuado usar para este fin el set de entrenamiento pues la estimación típicamente sobreestimará rendimiento.

c. De existir, el algoritmo ID3 visto en clases para entrenar árboles de decisión garantiza encontrar un árbol óptimo (medido por error en set de entrenamiento)?.

Falso, al momento de decidir el atributo para cada split del árbol la construcción es codiciosa (greedy), por tanto, sólo garantiza encontrar óptimos locales.

d. De existir, el algoritmo visto en clases para entrenar el clasificador de naive Bayes garantiza encontrar un modelo óptimo (medido por error en set de entrenamiento)?.

Falso, la construcción encuentra sólo una hipótesis que máximiza la probabilidad a posteriori bajo varias aproximaciones, por ejemplo, estimar las funciones de probabilidad utilizando frecuencias observadas, o asumir una serie de independencias probabilísticas.

e. De existir, el algoritmo visto en clases para entrenar perceptrones garantiza encontrar un modelo óptimo (medido por error en set de entrenamiento)?.

   Verdadero, la existencia implica que se trata de de un problema linealmente separable, además para el perceptron la métrica de error es convexa, por tanto, para este caso la búsqueda usando gradiente garantiza encontrar un óptimo global.

f. De existir, el algoritmo de backpropagation visto en clases para entrenar una red neuronal con capa oculta garantiza encontrar una hipótesis óptima (medida por error en set de entrenamiento)?.

   Falso, para el caso de una capa oculta la métrica de error no es convexa, por tanto, la técnica del gradiente estocástico sólo garantiza encontrar un óptimo local.

g. La estimación del error de un clasificador en el set de entrenamiento entrega una estimación pesimista del error real de este clasificador.

   Falso, la estimación es optimista pues puede existir sobreajuste (overfitting).

h. La estimación del error de un clasificador en el set de validación entrega una estimación pesimista del error real de este clasificador.

   Falso, la estimación es optimista pues puede existir sobreajuste al set de validación (overfitting).

i. La estimación del error de un clasificador en el set de test entrega una estimación pesimista del error real de este clasificador.

   Falso, la estimación no es sesgada.

j. Un algoritmo de aprendizaje supervisado se caracteriza por operar con datos no rotulados y encontrar patrones relevantes semánticamente.

   Falso, como discutimos en clases, los patrones no tienen necesariamente una interpretación semántica.

2. (*16 puntos*) Árboles de Decisión

Una expresión comúnmente utilizada para medir similitud entre 2 distribuciones de probabilidad es la denominada Kullback-Leibler divergence or KLD. Para 2 distribuciones de probabilidad $p(x)$ y $q(x)$, KLD se define como:

$$KLD(p(x), q(x)) = -\sum_x p(x) log_2 \frac{q(x)}{p(x)}.$$

Una expresión comúmmente utilizada para estimar la relación entre 2 variables aleatorias x e y es la denominada información mutua or $I(x, y)$, la cual se define como:

$$I(x, y) = H(x) - H(x/y)$$

donde $H(x) = \sum_x p(x) log_2 \frac{1}{p(x)}$ es la entropía de x; $H(x/y) = \sum_{y_i} p(y_i) H(x/y_i)$ es la entropía condicional de x dado y; y $H(x/y_i) = \sum_x p(x/y_i) log_2 \frac{1}{p(x/y_i)}$ es la entropía condicional específica de x dado y_i.

a. (*4 pts*) Determine la relación que conecta en forma estrecha KLD con ganancia de información:$GI(c, x)$.

b. (*4 pts*) Determine la relación que conecta en forma estrecha $I(x, y)$ con $GI(c, x)$.

c. (*4 pts*) Suponga que para la construcción de un árbol de decisión se decide cambiar el uso de $GI(C, A_i)$ como métrica para seleccionar el atributo A_i a ser utilizado en cada nodo del árbol. La nueva métrica se define como:

$$KLD(p(A_i, C), p(A_i)p(C))$$

donde A_i se refiere al atributo testeado y C indica la variable clase. Indique posibles ventajas o desventajas de esta métrica comparada con usar $GI(C, A_i)$.
Obs:

$$KLD(p(x, y), q(x, y)) = -\sum_x \sum_y p(x, y) log_2 \frac{q(x, y)}{p(x, y)}.$$

d. (*4 pts*) Suponga que para la construcción de un árbol de decisión se decide cambiar el uso de $GI(C, A_i)$ como métrica para seleccionar el atributo A_i a ser utilizado en cada nodo del árbol. La nueva métrica se define como:

$$I(C, A_i)$$

donde A_i se refiere al atributo testeado y C indica la variable clase. Indique posibles ventajas o desventajas de esta métrica comparada con usar $GI(C, A_i)$.

Solución:

La relación entre las 3 métricas indicadas es que son equivalentes. Por tanto, no hay mayor ventaja o desventaja de usar cualquiera de ellas. Hay distintas formas de demostrar esto, la clave en la demostración es el concepto que discutimos en clases sobre el significado de la ganancia de información. La ganancia de información de la clase respecto de un atributo A_i es la reducción de la entropía de la clase al considerar los valores de este atributo.

Según definición en clases: $GI(C, A_i) = H(C) - \sum_{a_i \in A_i} \frac{|C_{a_i}|}{|C|} H(C_{a_i})$. Esto es equivalente a:

$$\begin{aligned} GI(C, A_i) =& H(C) - \sum_{A_i} p(A_i)(\sum_C p(C/A_i) log_2 p(C/A_i)) \\ =& H(C) - H(C|A_i) \end{aligned}$$

Así podemos demostrar los requerido:
a)

$$KLD(C, A_i) = -\sum_{C}\sum_{A_i} p(C, A_i) log_2 \frac{p(C)p(A_i)}{p(C, A_i)}$$
$$KLD(C, A_i) = -\sum_{C}\sum_{A_i} p(C, A_i)(log_2 p(C) + log_2 p(A_i) - log_2 p(C, A_i))$$
$$KLD(C, A_i) = -\sum_{C}\sum_{A_i} p(C, A_i) log_2 p(C) - \sum_{C}\sum_{A_i} p(C, A_i)(log_2 p(A_i) - log_2 p(C, A_i))$$
$$KLD(C, A_i) = -\sum_{C} p(C) log_2 p(C) - \sum_{C}\sum_{A_i} P(C|A_i)P(A_i)((log_2 p(A_i) - log_2 p(C|A_i) - log_2 p(A_i))$$
$$KLD(C, A_i) = H(C) + \sum_{C}\sum_{A_i} p(C|A_i)p(A_i)(log_2 p(C|A_i))$$
$$KLD(C, A_i) = H(C) + \sum_{A_i} p(A_i) \sum_{C} p(C|A_i) log_2 p(C|A_i)$$
$$KLD(C, A_i) = H(C) - H(C|A_i) = GI(C, A_i)$$

b) Directa por la definición de información mutua.

c) No hay ventajas/desventajas pues son equivalentes, según demostración en a)

d) No hay ventajas/desventajas pues son equivalentes, según demostración en b)

3. (*16 puntos*) Naive y Redes de Bayes

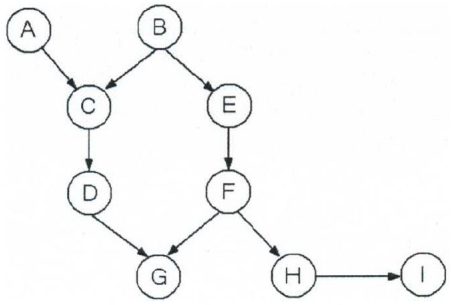
a. (*4 pts*) Se tiene un set de entrenamiento de D documentos, los cuales pueden pertenecer a una de C posibles clases. Cada documento contiene n palabras: $X_1, ..., X_n$, donde X_i es la palabra en la posición i en el documento. Para predecir la clase de cada documento se decide usar el siguiente modelo: $P(C|X_1...X_n) \propto P(X_1...X_n|C)P(C) = P(C)\prod_{i=1}^{n} P(X_i|C)$. Como es usual en modelos del tipo Naive Bayes para el caso de datos discretos, podemos asumir que todas las distribuciones son multinomiales, en este caso sobre un vocabulario V de tamaño $|V| = 3000$ palabras y un total de clases $|C| = 10$. Considerando el caso en que se tiene un set de entrenamiento de 1000 documentos, cada uno de 100 palabras, indique 1 ventaja y 1 desventaja de este modelo comparado al modelo tradicional de naive Bayes.
Solución:

Lo indicado en el enunciado se diferencia del modelo convencional en que cada variable representa la posición de la palabra en el texto, mientras que el modelo convencional sólo considerar la ocurrencia de cada palabra en el texto, no importando su posición.

Así, la principal desventaja del modelo propuesto es que será muy díficil poder modelar las funciones de probabilidad relevantes para realizar inferencia. Básicamente se requeriría documentos con cada palabra en cada posición posible.

La principal ventaja es que el modelamiento de la posición relativa y espacial de las palabras en un documento ayuda a la discriminatividad. Por ejemplo, que la palabra Inteligencia vaya seguida de Artificial, es una información relevantes acerca de la temática de cierto documento.

b. (*4 pts*) Considere la siguiente red de Bayes, en la cual todas las variables son binarias:



¿Cuál es el número de parámetros independientes de la red?
Solución:
La factorización relevante es: $p(A)p(B)p(C|A,B)p(E|B)p(D|C)p(F|E)p(G|D,F)p(H|F)p(I|H)$
1+1+4+2+2+2+4+2+2=20.

c. (*4 pts*) Para la red anterior ¿Cuál es el número de parámetros que se necesita para especificar $p(F|A,B,C,D,E,G,H,I)$.
Solución:
Primero observamos que dado E, F es independiente de B, y dado G y H, es independiente de todas las

restantes variables, por tanto:

$$p(F|A,B,C,D,E,G,H,I) = p(F|E,G,H)$$

Para este caso se necesitarían 8 parámetros, las respuestas que pusieron esto y realizaron la justificación anterior tienen 2 puntos.

La respuesta correcta requiere un poco más de trabajo:

$$\begin{aligned} p(F|A,B,C,D,E,G,H,I) &= p(F|E,G,H) \\ &= \frac{p(G,H|E,F)p(F|E)}{p(G,H|E)} \\ &= \frac{p(G,H|F)p(F|E)}{p(G,H|E)} \\ &= \frac{p(G|F)p(H|F)p(F|E)}{\sum_F p(G,H,F|E)p(F|E)} \\ &= \frac{p(G|F)p(H|F)p(F|E)}{\sum_F p(G|F)p(H|F)p(F|E)} \end{aligned}$$

Con lo cual se requieren 6 parámetros.

d. (*4 pts*) Suponga un modelo convencional de naive Bayes donde existen 4 atributos binarios (x_1, x_2, x_3, x_4) que son usados para predecir el valor de una clase $C \in [1, 2, 3, 4]$. Suponga que se decide agregar a este modelo relaciones causales que ligan cada atributo x_{i-1} a x_i, $i \in [2, 4]$. Respecto al modelo convencional de naive Bayes, ¿Cómo se ve afectado el número de parámetros que se tiene que estimar para este caso, asumiendo que todas las distribuciones consideradas son multinomiales?.

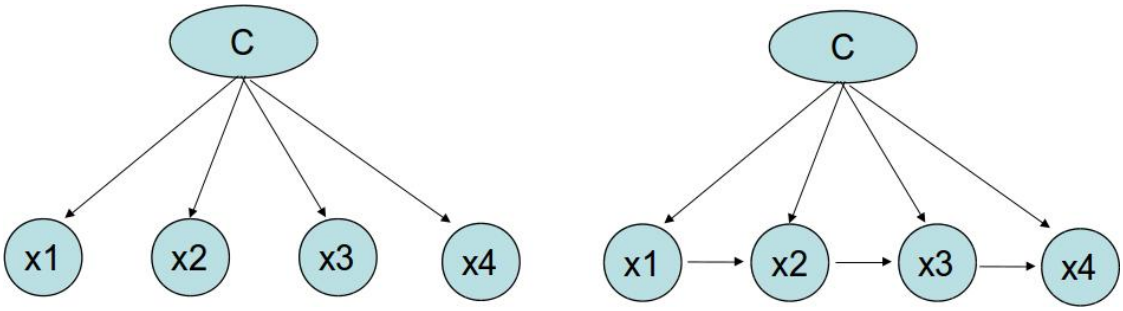
Solución:
Para el modelo convencional las funciones relevantes son:

$$p(C), P(x_1|C), \ldots, P(x_4|C)$$

Bajo el modelo alternativo las funciones relevantes son:

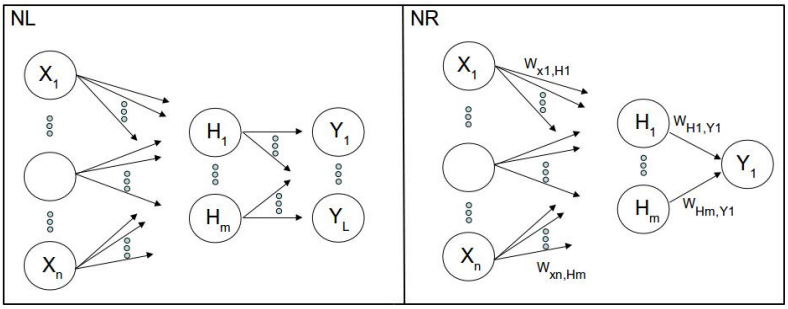
$$p(C), P(x_1|C), \ldots, P(x_4|C), P(x_4|x_3), P(x_3|x_2), P(x_2|x_1).$$

Por tanto, se agrega la estimación de 6 parámetros correspondientes a: $P(x_4|x_3), P(x_3|x_2)$ y $P(x_2|x_1)$.
Lo anterior es fácil de entender al comparar las redes de Bayes correspondientes a cada modelo:



4. (*16 puntos*) Redes Neuronales

Considere las estructuras de red neuronal en el lado izquierdo (NL) y en el lado derecho (ND) de la siguiente figura.



Ambas redes son del tipo feed-forward con una capa oculta. NL consiste de n unidades en la capa de entrada, m en la capa oculta y L en la capa de salida; mientras que NR contiene n unidades en la capa de entrada, m en la capa oculta y sólo 1 unidad en la capa de salida.
Suponga que se requiere modelar un problema con n variables de entrada y L variables de salida. Dos posibles alternativas para modelar este problema son: i) Usar directamente NL, ii) Utilizar L redes del tipo NR, donde cada una de estas redes es entrenada para modelar una de las posibles L variables de salida.

a. (*4 pts*) Compare la complejidad computacional de entrenar i) e ii) para el caso de un set de entramiento de D registros y p épocas de entrenamiento. ¿Cuál solución tiene menor complejidad?. Fundamente su respuesta.

Solución:
NL: $O\left(D * P * (n * m + n * m + m * L + m * L)\right)$

NR: $O\left(D * P * L * (n * m + n * m + m + m)\right)$

b. (*4 pts*) Compare la solución i) e ii) en términos de su poder representacional, ¿Cuál es mayor?. Fundamente su respuesta.

Solución:
Parámetros NL:n*m + m*L
Parámetros NR:L(n*m + m)
NR tiene un mayor número de parámetros por tanto un mayor poder representacional.

c. (*4 pts*) Dejando de lado argumentos relacionados con el número de parámetros de ambas redes, ¿Cuál de las soluciones tiene mayor posibilidad de sobre-ajustar el set de entrenamiento?. Fundamente su respuesta.

Solución:
NL tiene menos posibilidad de sobreajustar pues la representación aprendida entre la capa de entrada y capa oculta (pesos que conectan X con H) es común para la generación de todas las salidas Y. Éste no es el caso para la red NR, donde cada una de las L redes debe ser entrenada en forma totalmente independiente. El hecho de compartir parte de la representación en NL evita el overfitting al evitar aprender una

salida específica sino todas ellas al mismo tiempo (L variables de salida).

d. (*4 pts*) Para el caso de la red NR derive una expresión para actualizar el peso $W_{X1,H1}$ usando el método del gradiente estocástico visto en clases. Asuma que las unidades usan activación sigmoidal dada por: $\sigma(x) = \frac{1}{1+e^{(-x)}}$.
Ayuda: $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$.

Solución:

La función de error está dada por: $E = \frac{1}{2}\sum_D (t_1 - Y_1)^2$, donde t_1 es la salida deseada.

$$\frac{dE}{dW_{X_1,H_1}} = \frac{dE}{dY_1}\frac{dY_1}{dnet_{Y1}}\frac{dnet_{Y1}}{dH_1}\frac{dH_1}{dnet_{H1}}\frac{dnet_{H1}}{dW_{X_1,H_1}}$$

$$\frac{dE}{dW_{X_1,H_1}} = -(t_1 - Y_1)Y_1(1-Y_1)W_{H_1,Y_1}H_1(1-H_1)X_1$$