Práctico 3

Aprendizaje bayesiano. Aprendizaje por casos. Metodología.

Ejercicio 1

Suponga que cuenta con un conjunto de entrenamiento D sin ruido, y se considera un espacio H en donde las hipótesis a priori, cuanto más generales, mayor probabilidad tienen de ser el concepto objetivo. Indique si las siguientes afirmaciones son verdaderas o falsas. Justifique.

i. Find-S da una hipótesis MAP.
ii. Find-S da una hipótesis ML.
iii. Candidate-Elimination con votación es un clasificador bayesiano óptimo.

Ejercicio 2

Se desea aplicar el principio MDL a un espacio de conjunciones de hasta *n* atributos booleanos; por ejemplo: *Soleado* $\wedge$ *SinCambios*. Se tiene un conjunto de m ejemplos.

Cada hipótesis se transmite listando sus atributos; cada atributo se codifica utilizando log(n) bits. Dada una hipótesis h, la codificación de un ejemplo tiene largo cero si h lo clasifica correctamente, y largo log(m) en caso contrario (para indicar cuál ejemplo es errado).

a) Dé la expresión a minimizar según el principio MDL.

b) ¿Es posible construir un conjunto de entrenamiento que tenga una hipótesis consistente pero haga que MDL elija otra menos consistente? Justifique.

c) Plantee distribuciones para P(h) y P(D|h) bajo las cuales el algoritmo MDL da como resultado una hipótesis MAP.

Ejercicio 3

A partir de la aplicación de la «ley de inclusión financiera», se decide monitorear cómo se comportan los consumidores respecto a los medios de pago. Se recaba el siguiente conjunto de ejemplos, siendo «Medio» el atributo objetivo:

#	Edad	Medio	Gasto	Mercancía	Sexo
1	18	efectivo	80	1ra. necesidad	M
2	35	débito	12	1ra. necesidad	F
3	55	crédito	180	1ra. necesidad	M
4	73	efectivo	80	cultura	?
5	45	crédito	540	electrodomésticos	M
6	27	débito	150	electrodomésticos	F

Los atributos numéricos son enteros, edad $\in$ [18, 80], gasto $\in$ [0, 1000], sexo $\in$ {M, F}, medio $\in$ {efectivo, débito, crédito} y mercancía $\in$ {1ra. necesidad, electrodomésticos, cultura}.

a) Explique cómo se podrían tratar los atributos numéricos y los atributos faltantes según se apliquen los siguientes algoritmos: ID3, Naive Bayes y KNN.

b) Defina lo necesario para aplicar 3-NN sobre el conjunto de ejemplos. Justifique.

c) Aplique los algoritmos Naive Bayes e ID3 sobre el conjunto de datos, ahora preprocesados, en donde se consideran tres rangos de edades: <30, 30-60, >60; y tres rangos para los gastos: <100, 100-500, >500.

d) Compare la solución por rangos utilizada en (c) con su solución en (a).

e) Clasifique a la siguiente instancia según sus tres clasificadores obtenidos en (b) y (c).

#	Edad	Medio	Gasto	Mercancía	Sexo
7	38	?	950	electrodomésticos	M

Ejercicio 4

Suponga que quiere implementar un detector de correo no deseado y cuenta con el siguiente conjunto de entrenamiento:

#	Agendado	Idioma	Para	Venta	Deseado
1	Sí	Inglés	Sí	Sí	Sí
2	No	Otro	No	No	Sí
3	No	Español	No	Sí	Sí
4	Sí	Otro	Sí	No	No
5	No	Español	Sí	Sí	No

Considere al ejemplo:

#	Agendado	Idioma	Para	'Venta'	Deseado
6	Sí	Inglés	No	Sí	??

a) ¿Cómo clasificaría un clasificador bayesiano sencillo al correo #6? ¿Con qué probabilidad?

b) ¿Qué modificación debería hacer en la parte anterior, si le informan que el atributo «Idioma» puede tomar el valor «francés»?

c) ¿Cómo implementaría un 2-NN? Ejemplifique su solución clasificando a #6.

Ejercicio 5

Se quiere resolver el siguiente problema de clasificación: «¿Qué días Pedro compra paltas?», y se cuenta con el siguiente conjunto de entrenamiento:

#	Precio ($/Kg)	Ciudad	Mes	Humor	Calidad	Peso (Kg)	¿Compra?
1	100	Salto	Febrero	Bueno	Buena	0,300	Sí
2	140	Florida	Marzo	Malo	Media	0,200	No
3	80	Salto	Marzo	Malo	Buena	0,500	Sí
4	160	Durazno	Agosto	Bueno	?	0,700	Sí
5	200	Salto	Octubre	Bueno	Buena	0,100	No

donde Ciudad$\in${Salto, Florida, Durazno}, Humor$\in${Bueno, Malo} y Calidad$\in${Buena, Media, Mala} y Precio y Peso son atributos numéricos.

a) Dé dos formas de manejar el valor faltante de la instancia 4.

b) Considere un clasificador que responda afirmativamente únicamente a las paltas con peso mayor a 600 gramos y al conjunto de entrenamiento dado:

   i. Dé la matriz de confusión.

ii. Calcule la estimación del acierto micro y macro.

iii. Dé la estimación de la precisión y el *recall* de la clase positiva.

c) ¿Qué modificaciones le realizaría al corpus para aplicar el algoritmo KNN? Justifique

d) Explique una ventaja y una desventaja de aumentar el valor de K en el algoritmo KNN.

e) Dé el resultado de clasificar a la siguiente instancia según el algoritmo Naïve Bayes, discretizando a los atributos continuos como crea conveniente:

#	Precio	Ciudad	Mes	Humor	Calidad	Peso	¿Compra?
6	110	Salto	Marzo	Malo	Buena	0,250	?

Ejercicio 6

Considere el siguiente conjunto de datos de estudiantes, y considere el problema de aprender a clasificar cuáles aprobaron el año.

#	Nombre	Turno	Edad	Nota anterior	Sexo	Estatura	Aprueba
1	Luisa	Vespertino	16	10	Fem	1,79	Sí
2	Ana	Vespertino	15	3	Masc	1,65	Sí
3	Juanjo	Matutino	14	4	Masc	?	No
4	Pedro	Vespertino	16	9	Masc	?	Sí
5	Ramiro	Vespertino	15	3	?	1,65	Sí
6	Daniel	Matutino	15	8	Masc	?	No
7	Eduardo	?	14	10	Masc	?	Sí
8	Aiala	?	15	9	Fem	?	Sí
9	Dina	Matutino	15	7	Fem	?	No
10	Mathias	Vespertino	14	11	Masc	1,79	Sí
11	Diego	Matutino	15	5	Masc	?	Sí
12	Santiago	Vespertino	15	6	Masc	1,67	Sí
13	Luis	Maturino	15	6	Masc	1,72	No
14	Alejandra	Vespertino	16	6	Fem	1,73	Sí

a) ¿Cómo completaría los atributos faltantes para cada uno de los atributos?

b) ¿Qué sugeriría hacer con el atributo estatura?

c) ¿Cuál sería el resultado de transfomar el atributo «Turno» utilizando one-hot-encoding?

d) Divida el conjunto de datos en entrenamiento y evaluación, utilizando estratificación.

e) Estandarice los valores de edad y nota. Aplique la estandarización al conjunto de evaluación ¿Es lo mismo esto que aplicar la estandarización a todo el dataset?

f) Suponga un clasificador trivial que predice la clase mayoritaria en el conjunto de entrenamiento. Suponga ahora que nuestra tabla es el conjunto total de entrenamiento, y realice 3-fold cross validation. Reporte precisión, recall, medida F, así como la desviación estándar de cada una de ellas.

Ejercicio 7

a) Suponga que su conjunto de entrenamiento tiene 8000 instancias y se desea ajustar el hiperparámetro α de un cierto método de aprendizaje automático. Describa el proceso de validación cruzada con 5 folds y accuracy como medida de rendimiento.

b) Si cuenta con un conjunto de datos con 9200 instancias, divididas en tres clases A, B y C con 8500, 500 y 200 instancias, respectivamente, y ordenadas por clase: ¿cómo dividiría a las instancias en los conjuntos de entrenamiento y evaluación? Justifique.

Se realiza la evaluación de un modelo y se obtiene la siguiente matriz de confusión (en las filas están los valores correctos de las instancias):

	A	**B**	**C**
A	910	5	5
B	5	20	20
C	6	4	15

i) Calcule accuracy, precisión, recall y medida F1 para cada una de las clases

ii) Obtenga los valores de macro y micro average de las medidas. Analícelas.

Ejercicio 8

Considere una clasificador binario h que, cuando es evaluado en un conjunto de 100 instancias, clasifica correctamente a 83 de ellas. ¿Cuál es la desviación estándar y el intervalo de confianza de 95% para la tasa real de error? ¿Y si clasifica correctamente 830 de 1000 instancias?