Aprendizaje Automático (2014-2015)

Grado en Ingeniería Informática

Universidad de Granada

Cuestionario T1

Mª Cristina Heredia Gómez

29 de marzo de 2015

Índice

Índice de figuras

1. (0.5 puntos) Decir cuáles de los siguientes casos son un problema de regresión o de clasificación e indicar si estaremos más interesado en inferencia o en predicción. Identificar también los valores de n (tamaño muestra) y p (número de predictores):

1.1. Recopilamos un conjunto de datos de las 500 empresas españolas más grandes. Por cada compañía recogemos: beneficio anual, número de empleados, tipo de industria y sueldo del director. Estamos interesados en comprender que factores afectan al sueldo del director.

En este caso buscamos...estimar los factores que influyen en el sueldo del director, luego sería un problema de regresión donde estaríamos más interesados en inferir los factores que afectan al sueldo del director que en predecirlos. Tendríamos: n=500 , p=4

1.2. Estamos considerando lanzar un nuevo producto y deseamos conocer si será un éxito o un fracaso. Para ello recogemos datos de 20 productos semejantes ya existentes en el mercado. Para cada producto medimos: a) si fue un éxito o un fracaso; b) precio del producto; c) presupuesto de marketing; d) precio de oferta inicial y otras diez variables más.

Se trata de un problema de clasificación, pues queremos que la salida sea: será éxito o fracaso en el cual estaremos interesados en predecir este resultado usando cosas que ya conocemos de otros productos. Tendríamos: n=20 , p=15 (14 más si será éxito o fracaso)

1.3. Estamos interesados en predecir el % de variación del euro respecto de los porcentajes de variación semanales de los mercados europeos. Para ello recogemos datos semanales de todo el 2012. En cada semana medimos el % de cambio del euro, el % de cambio de la bolsa Alemana, el % de cambio de la Bolsa Inglesa y el % de cambio de la bolsa Francesa.

Se trata de un problema de regresión, pues la salida será una variable cuantitativa. Además estaremos interesados en hacer una predicción, pues buscamos la variación futura del euro respecto a los mercados europeos, basándonos en datos que ya conocemos. Tendríamos: n=53 , p=4

2. (1 punto) Identificar dos aplicaciones empresariales (no comentadas en clase) en las que considere que las técnicas de Aprendizaje Automático serán útiles. Describir brevemente cada una de ellas, el interés de la misma y el problema que se resuelve. Identificar algunas de las variables que considere más importantes al problema. Describir un caso en que regresión será la técnica a aplicar (decir además si el problema es más de inferencia o de predicción) y otro de clasificación.

Un posible caso en el que la regresión será la técnica a aplicar, será por ejemplo predecir el número de personas que se espera que visiten un destino turístico y no otro, para hacer mejores ofertas y más márketing en unos paquetes de viajes u otros, para una agencia de viajes. Aquí estaríamos antes un problema de predicción en el que estaríamos tratanto de predecir cuántas personas visitará cada destino a partir de unos datos conocidos. Algunas variables en este problema serían: número de turistas en los destinos otros años, el clima de este año, las prioridades de la gente de este año, la economía actual(sueldo mensual medio), la situación política y social de los diferentes destinos ofertados...
En cuanto a un problema de clasificasión, yo creo que un buen ejemplo sería la biometría, pues ahí tenemos datos de un individuo y queremos identificarlo; es decir, queremos saber quien es, clasificarlo. Aquí ni la variable de salida ni la mayoría de los datos que se recojen de una persona para llevar a cabo su identificación son variables cuantitativas. Algunas variables de este problema serían: iris del ojo de la persona, huella dacilar, frecuencia y tono de la voz, caligrafía...

3. (0.5 puntos) Describir las diferencias entre las aproximaciones supervisadas paramétricas y las no-paramétricas. ¿Cuáles son las ventajas de la aproximación paramétrica en regresión y clasificación? ¿Cuáles las desventajas? Justificar la respuesta.

La diferencia entre las aproximaciones supervisadas paramétricas y no paramétricas es principalmente, que en las aproximaciones paramétricas dependen de unos parámetros que yo tengo que conocer a priori, mientras que las no paramétricas no.
Las ventajas de la aproximación paramétrica es que es sencilla ya que el estimar una función se reduce a estimar un conjunto pequeño de coeficientes, y además esta función dependerá de unos parámetros que son conocidos. Otra ventaja que tiene es que no necesita muchos datos para poder conseguir una estimación.
Como desventajas tiene que fija un comportamiento para todo el espacio muestral, por lo que si los datos presentaban una variabilidad fuerte, no podrían ajustarse con un método paramétrico.
Otra desventaja es que hay que fijar los parámetros de los que va a depender el modelo, y hay que ser cuidadoso con eso pues una mala elección de parámetros podría estropear el modelo; además de tener

cuidado con los outlayers y high leverages por la misma razón.

4. (0.5 puntos) Si tenemos un problema de clasificación con dos variables predictoras y nos muestran las fronteras de decisión de un clasificador kNN para distintos valores de k ¿Cómo podemos saber si la frontera de decisión comienza a estar sobre-ajustada? Justificar la respuesta.

En este caso, podemos saber si la frontera de decisión comienza a estar sobre-ajustada si vemos que se pega demasiado a los datos, pues en ese caso tendríamos una frontera de decisión que no es nada flexible y se podrían estar cometiendo errores. También podemos tomar como referencia que a medida que k se incrementa la varianza va disminuyendo así como la bias va aumentando.

5. (2 puntos) Suponga que tenemos un conjunto de datos con 5 variables predictoras, X_1 ,X_2 ,X_3,X_4,X_5, de las cuales X_1 y X_2 son cuantitativas, X_3 es cualitativa con dos valores (0=hombre, 1=mujer), X_4 representa la interacción entre X_1 y X_2 , y X_5 representa la interacción entre X_1 y X_3 . La variable de salida representa el valor del salario de hombres y mujeres. Hemos ajustado un modelo por mínimo cuadrados y se han obtenido los siguientes coeficientes β_0=50,β_1=20,β_2=0.07,β_3=35,β_4=0.01,β_5=−10.

5.1. ¿Cuáles de las siguientes contestaciones es correcta y por qué?

- Para valores fijos de X_1 y X_2 los hombres ganan más en promedio que las mujeres:

  mujer: 50+X_1*20+X_2*0.07 +1*35+X_1*X_2*0.01+X_1*1*(-10)
  hombre: 50+X_1*20+X_2*0.07 +0*35+X_1*X_2*0.01+X_1*0*(-10)

  luego lo que realmente varía entre hombre y mujer:
  1*35+X_1*1*(-10)
  0*35+X_1*0*(-10)

  así que la afirmación será correcta o no dependiendo del valor que tomen X_1 y X_2, pues si, por ejemplo:
  X_1=50 , X_2=30
  la mujer gana: 50+1000+2.1+35+15-50=1052.1
  el hombre gana: 50+1000+2.1+15= 1067.1

ejemplo:
X_1=5 , X_2=45
la mujer gana: 140.4
el hombre gana: 155.4

digamos que si X_1 >=35 =>mujer gana menos. Pero si X_1 <35 y no hay mucha diferencia entre el valor de X_1 y X_2 , o son iguales =>parace que la mujer gana más. Por lo que, la afirmación es correcta pero sólo para algunas asignaciones de valores.

- Para valores fijos de X_1 y X_2 las mujeres ganan más en promedio que los hombres. Como he dicho antes, esto será verdad en algunos casos; a saber... aquellos en los que X_1 y X_2 toman el mismo valor o aquellos en los que no hay mucha diferencia entre el valor que toman X_1 y X_2.

  por ejemplo, para X_1=2 y X_2=4 obtenemos que la mujer gana 105.36 mientras que el hombre gana 90.02. O para X_1=2 y X_2=4

- Para valores fijos de X_1 y X_2 los hombres ganan más en promedio que las mujeres con tal que X_1 sea suficientemente grande. esto es cierto, tal y como dije antes. Creo que esto se debe a la penalización de multiplicar X_5 por -10, ya que en el hombre esta penalización no se da. Si es muy grande, afecta mucho. En el primer apartado puse un ejemplo.

5.2. Predecir el salario de una mujer con X_1=4.0 y X_2=110

salario mujer= 50+4*20+110*0.07 +1*35+4*110*0.01+4*1*(-10)=137.1

5.3. Dado que el coeficiente de X_4 es pequeño existe poca evidencia de un efecto de interacción entre X_1 y X_2, ¿ Verdadero o Falso? Justificar la respuesta

Esto es falso. El coeficiente de X_2 también es pequeño de por sí y no hablamos de poca evidencia de interacciones. Yo lo que entiendo al ver que el coeficiente de X_4 es pequeño es que posiblemente esta variable ya está explicada en otras variables.

6. (1.5 puntos) Tenemos un conjunto de datos de 100 observaciones con una única variable predictor y una respuesta cuantitativa. Ajustamos a dichos datos un modelo de regresión lineal $Y = \beta_0 + \beta_1 X + \varepsilon$ y un modelo de regresión cúbico $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

6.1. Supongamos que la verdadera relación entre X e Y es lineal, es decir $Y = \beta_0 + \beta_1 X + \varepsilon$. Considerar la suma de los residuos de los datos de entrenamiento (RSS) tanto para el modelo lineal como para el modelo cúbico. ¿Deberíamos esperar que en general un valor fuera menor que el otro, que fueran iguales, o no hay suficiente información para establecer una opinión por adelantado?

sabemos que el RSE es la raíz de (RSS/(n-2)) luego si el RSS es más grande, obtendremos un valor más grande de RSE y si el RSS es chico, obtendremos un valor más pequeño del RSE. Ahora bien, si la verdadera relación entre X e Y fuera, realmente lineal, entonces el modelo lineal debería ajustar bien los datos, por lo que deberíamos de obtener un RSS menor o similar al que obtendríamos con el modelo cúbico, tal que el RSE también fuese menor o igual. No obstante, esto dependerá de la información que tengamos para el problema.

6.2. Contestar lo mismo del punto anterior pero considerando las sumas RSS de los datos de test.

En este caso ya no partiremos de los datos de entrenamiento, sino que iremos con los datos test, sin embargo, si estos conjuntos son similares debería pasar algo similar con el RSS, eso sí, suponiendo que tenemos toda la información necesaria, pues en caso contrario podríamos obtener un RSS grande en el caso del modelo cúbico.

6.3. Supongamos que la verdadera relación entre X e Y es no lineal, pero no conocemos como de lejos está de ser lineal. Consideremos las sumas RSS de entrenamiento para el modelo lineal y el cúbico ¿Deberíamos esperar que en general un valor fuera menor que el otro, que fueran iguales, o no hay suficiente información para establecer nada por adelantado? Justificar la contestación.

Si la verdadera relación entre X e Y no es lineal , entonces, si obtenemos unas estimaciones de los coeficientes similares a los de verdad, obtendremos un RSS más pequeño en el modelo cúbico y por lo tanto un RSE más pequeño en el modelo cúbico lo que se entiende como que el modelo cúbico ajustaría mejor los datos que el modelo lineal.

6.4. Contestar lo mismo del punto anterior pero considerando la sumas RSS de los datos de test.

A la hora de pasar a los datos test, tendríamos un problema, pues no disponemos de suficiente información y esto podría reflejarse para mal en el modelo cúbico, obteniendo un RSS y un RSE elevados y dando a entender que ajusta los datos peor que el lineal sin tener porqué ser así.