Cuestionario 1

Aprendizaje Automático

Grado en Ingeniería Informática

Granada, 29 de Marzo de 2015.

Datos del estudiante

Fernández Bosch, Pedro
76422233-H

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):

a) 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 interesado en comprender que factores afectan al sueldo del director.

Dado el problema, posiblemente se trate de un problema de regresión, porque se pretende comprender la relación entre una determinada variable y es resto de variables predictoras para entender los factores que afectan al sueldo del director.

En este problema estaríamos especialmente interesados en la inferencia, para conocer que variables X realmente afectan al valor de Y.

N (Tamaño de la muestra): 500
P (Número de predictores): 4

b) 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.

En este caso, podríamos deducir que se trata de un problema de clasificación, puesto que nuestro objetivo es clasificar de forma binaria si el nuevo producto será un éxito o un fracaso.

En este problema estaríamos especialmente interesados en la inferencia, para evaluar la relevancia de las variables respecto a la salida.

N (Tamaño de la muestra): 20
P (Número de predictores): 4+10

c) 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.

Dado el enunciado, posiblemente se trate de un problema de regresión, porque lo realmente interesante es establecer la relación entre una determinada variable y es resto de variables predictoras para comprender los factores que afectan al % de variación del euro respecto a la variación del resto de mercados.

En este problema estaríamos especialmente interesados en la predicción, puesto que deseamos conseguir buenas predicciones del valor de 1), para los nuevos valores de X nunca antes observados.

N (Tamaño de la muestra): 53 – Numero de semanas de 2012.
P (Número e predictores): 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.

Problema de regresión:

Una compañía de seguros dispone de un departamento comercial y está interesada en comprender que factores afectan al número de clientes captados. Recopilamos un conjunto de datos de las actuaciones de 1000 comerciales de la empresa durante 2013. Por cada comercial recogemos: 1) número de clientes captados; 2) número de llamadas telefónicas realizadas; 3) número de visitas a domicilio realizadas; 4) promedio de horas trabajadas, 5) salario del comercial; además de otras variables.

En este problema estaríamos especialmente interesados en la inferencia, para conocer que variables X realmente afectan al valor de Y.

Variables más relevantes:
1) Número de clientes captados.
2) Número de llamadas telefónicas realizadas.
3) Número de visitas a domicilio realizadas.
4) Promedio de horas trabajas.
5) Salario del comercial.

Problema de clasificación:

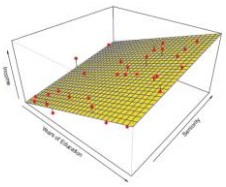
Un hotel granadino desea lanzar por primera vez una oferta vacacional para Semana Santa y desea saber si la oferta tendrá buena acogida entre los turistas. Para ello se han recogido datos de otros 50 hoteles que ya lanzaron ofertas similares el año pasado. Por cada oferta se ha medido: 1) si la oferta tuvo buena acogida o no; 2) precio de la oferta; 3) número de agencias asociadas al hotel; 4) presupuesto en publicidad; además de otras variables.

En este problema estaríamos especialmente interesados en la inferencia, para evaluar la relevancia de las variables respecto a la salida.

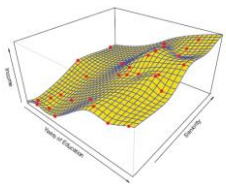
Variables más relevantes:
1) Si la oferta tuvo buena acogida o no.
2) Precio de la oferta.
3) Número de agencias asociadas al hotel.
4) Presupuesto en publicidad.

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.

- Las aproximaciones supervisadas paramétricas son aquellas en los que la estimación de f(x, β) se reduce a la estimación de β, es decir, reducen el problema de estimar f al de estimar un conjunto de parámetros.



- Las aproximaciones supervisadas no-paramétricas no asumen una única forma funcional para f, es decir, no fijan un único comportamiento para todo el espacio maestral.



Ventajas de la aproximación paramétrica en regresión y clasificación:

- Reducen el problema de estimar f al de estimar un conjunto de parámetros. (Más sencillo).
- En principio, se necesitan pocas observaciones para obtener una estimación en todo el espacio de las variables X.

Desventajas de la aproximación paramétrica en regresión y clasificación:

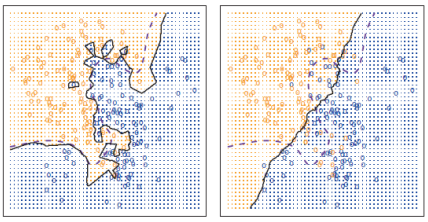
- Incluso si el nivel de ruido es bajo, obtendremos una mala estimación si el modelo no es el adecuado.
- No son tan flexibles como los métodos de regresión no-lineal y ello puede provocar predicciones más imprecisas.
- Ajustar un modelo más flexible requiere estimar un mayor número de parámetros.

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.

El sobreajuste u overfitting es el efecto de sobreentrenar un modelo con ciertos datos para los que se conoce el resultado deseado y es la peor consecuencia de un mal ajuste.

El modelo debe alcanzar un estado en el que sea capaz de predecir los resultados de otros casos a partir de lo aprendido en la fase de entrenamiento, generalizando para poder resolver situaciones distintas a las del proceso de entrenamiento.

Para el caso de un clasificador kNN con dos variables predictoras, vamos a apoyarnos en el siguiente gráfico para explicar el sobreajuste:



A primera vista puede comprobarse que el modelo de la izquierda está sobreajustado. Esto se determina comprobando que ambas variables quedan totalmente separadas unas de las otras por la frontera, hasta el punto que casos muy específicos dentro de la zona naranja han sido aislados (los outliers son también sobreajustados).

Sin embargo el modelo de la derecha representa un ejemplo de modelo vagamente ajustado. Por lo tanto, lo correcto es tomar el ajuste en la línea punteada (algo intermedio, ni mucho, ni poco). Aunque algunos datos de una variable predictora queden dentro de la frontera de la otra variable (algunos datos máximos, datos extraños, etc.), es preferible este resultado que ajustar el modelo a unas características muy específicas.

Pero si es imposible comprobarlo con una gráfica ¿qué hacemos?

Para resolver esta cuestión se utiliza el error de validación (crossvalidation). En este caso, se divide el conjunto de datos originales, apartando un pequeño subgrupo de los mismos. Una vez realizado el ajuste, se utilizan los parámetros θj obtenidos para hallar el valor estimado de la salida de estos datos.

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 $\beta_0=50, \beta_1=20, \beta_2=0.07, \beta_3=35, \beta_4=0.01, \beta_5=-10$.

a) ¿Cuáles de las siguientes contestaciones es correcta y por qué?

i) Para valores fijos de X_1 y X_2 los hombres ganan más en promedio que las mujeres.
Incorrecta.

ii) Para valores fijos de X_1 y X_2 las mujeres ganan más en promedio que los hombres
Incorrecta.

iii) 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.
Correcta.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Primero hemos comprobado el modelo para un valor de X_1 pequeño (X_1 = 0,1 y X_2=110)

Para las mujeres:
Y = 50 + 20 . 0,1 + 0,07 . 110 + 35 . 1 + 0,01 . (0,1 . 110) + (-10) . (0,1 . 1) =
50 + 2 + 7,7 + 35 + 0,11 – 1 = 93,81

Para los hombres:
Y = 50 + 20 . 0,1 + 0,07 . 110 + 35 . 0 + 0,01 . (0,1 . 110) + (-10) . (0,1 . 0) =
50 + 2 + 7,7 + 0 + 0,11 + 0 = 59,81

Ganan más las mujeres.

Ahora, hemos comprobado el modelo para un valor de X_1 más grade (X_1 = 4 y X_2=110)

Para las mujeres:
Y= 50 + 20 . 4 + 0,07 . 110 + 35 . 1 + 0,01 . (4 . 110) + (-10) . (4 . 1) =
= 50 + 80 +7,7 + 35 + 4,4 – 40 = 137,1

Para los hombres:
Y= 50 + 20 . 4 + 0,07 . 110 + 35 . 0 + 0,01 . (4 . 110) + (-10) . (4 . 0) =
= 50 + 80 +7,7 + 0 + 4,4 + 0 = 142,1

Ganan más los hombres.

Respuesta, los hombres ganan más en promedio que las mujeres con tal que X_1 sea suficientemente grande.

b) Predecir el salario de una mujer con X_1= 4.0 y X_2=110

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

Y= 50 + 20 . 4 + 0,07 . 110 + 35 . 1 + 0,01 . (4 . 110) + (-10) . (4 . 1) =
= 50 + 80 +7,7 + 35 + 4,4 – 40 = 137,1

El salario de una mujer con X_1= 4.0 y X_2=110 es de 137,1.

c) 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

Falso, aunque el coeficiente sea pequeño, puede que exista una asociación estadísticamente significativa entre predictor y que la interacción sea bastante buena.

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 $Y=\beta_0+\beta_1 X+\beta_2 X^2+\beta_3 X^3+\varepsilon$

Los residuos de los datos, "ε", son la estimación de los errores. En regresión lineal, los residuos observados y los esperados bajo hipótesis deben ser parecidos e independientes. En consecuencia, el

análisis de los residuos permite profundizar en la relación que se produce entre variables y ajustar adecuadamente la regresión.

a) Supongamos que la verdadera relación entre X e Y es lineal, es decir $Y=\beta_0+\beta_1 X+$. 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?

Para dar una respuesta acertada a esta pregunta, deberíamos de conocer los datos del problema, pero lo general sería que en la fase de entrenamiento los residuos de los datos de entrenamiento de un modelo lineal fuesen mayores que para un modelo cubico.

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

En el caso del test, la hipótesis anterior puede variar. Porque si el modelo cubico ha supuesto un sobreajuste en los datos de entrenamiento, la suma de los residuos de los datos de test con un modelo lineal podría ser menor que uno cubico. Igual que en el apartado anterior, dependemos de los datos proporcionados.

c) 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.

De nuevo se plantea otro problema, porque en este caso no sabemos cómo de lejos está la relación de ser lineal. Pero por regla general, podríamos esperar las sumas de los residuos de los datos de entrenamiento de un modelo lineal fuesen mayores que para un modelo cubico.

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

De nuevo en el caso del test, la hipótesis anterior puede variar. Porque si el modelo cubico ha supuesto un sobreajuste en los datos de entrenamiento, la suma de los residuos de los datos de test con un modelo lineal podría ser menor que uno cubico. De nuevo dependemos de los datos proporcionados.