Tarea 1.

Introducción al aprendizaje estadístico.

Profesor:

José Ángel Anguiano.

Alumnos:

Enrique Donizell Cervantes Espinoza.

Carlos Geovany Triana Barragán.

Iván Gerardo Sánchez Burgueño.

1- Para cada una de las partes (a) a (d), indique si generalmente se espera que el rendimiento de un método flexible de aprendizaje estadístico sea mejor o peor que un método inflexible. Justifica tu respuesta.

(a) El tamaño de la muestra n es extremadamente grande, y el número de predictores p es pequeño.

(b) El número de predictores p es extremadamente grande, y el número de observaciones n es pequeño.

(c) La relación entre los predictores y la respuesta es altamente no lineal.

(d) La varianza de los términos de error, i.e., $\sigma^2 = Var(\varepsilon)$, es extremadamente alta.

Solución:

(a) Si el tamaño de la muestra n es extremadamente grande, y el número de predictores p es pequeño, entonces se esperaría que el rendimiento de un método flexible de aprendizaje estadístico sea mejor que un método inflexible, puesto que la adaptación de punto a punto del método flexible se vuelve mejor para la aproximación de la función entre las variables predictoras y la de respuesta, esto no ocurre en el caso del método inflexible debido a su rigurosidad si la cantidad de datos es muy grande y la cantidad de predictoras es pequeña la separación entre los datos provocaría que la aproximación a la función entre las predictoras y la respuesta no fuese muy buena, incluso en un caso extremo, dicha función no sería un buen modelo para el análisis, puesto que las inferencias y predicciones estarían muy alejadas de la realidad.

(b) Si el número de predictores p es extremadamente grande, y el número de observaciones n es pequeño, entonces se esperaría que el rendimiento de un método flexible de aprendizaje estadístico sea peor que un método inflexible, esto debido a que el modelo inflexible tendría una mejor aproximación a la función entre las predictoras y la respuesta, i.e., las aproximaciones, las predicciones y las inferencias obtenidas con un modelo inflexible en el que el número de predictoras

es extremadamente grande y el tamaño de la muestra es muy pequeño serían muy buenas, una función definida rigurosamente sería una mejor aproximación que un modelo flexible ya que el modelo flexible podría sobrepasar el pequeño número de observaciones.

(c) Si la relación entre los predictores y la respuesta es altamente no lineal, esto significa que un modelo de aprendizaje estadístico inflexible no sería de utilidad, esto se debe a la forma de los datos, si la función tiene una forma estricta y definida la aproximación a la función entre las predictoras y la respuesta no sería buena, incluso no estaría ni cerca de aproximarse a dicha función, mientras que el método flexible se adapta punto a punto y la aproximación a la función entre los predictores y la respuesta sería mejor, además con más grados de libertad, un modelo flexible obtendría un mejor ajuste.

(d) Un método flexible sería peor que uno inflexible, puesto que los métodos flexibles se ajustan al ruido en los términos de error y aumentan la varianza.

2- Explique si cada escenario es un problema de clasificación o regresión, e indique si estamos más interesados en la inferencia o la predicción. Finalmente, proporcione n y p.

(a) Recopilamos un conjunto de datos sobre las 500 principales empresas de los EE. UU. Para cada firma registramos ganancias, número de empleados, industria y Salario CEO. Estamos interesados en entender qué factores afectan el salario del CEO.

(b) Estamos considerando lanzar un nuevo producto y deseamos saber si será un éxito o un fracaso. Recopilamos datos sobre 20 productos similares que se lanzaron previamente. Para cada producto hemos registrado si fue un éxito o un fracaso, el precio cobrado por el producto, presupuesto de marketing, precio de competencia, y otras diez variables.

(c) Estamos interesados en predecir el cambio porcentual en el USD/Euro tipo de cambio en relación con los cambios semanales en el mundo los mercados de valores. Por lo tanto, recopilamos datos semanales para todo 2012. Para cada semana registramos el% de cambio en el USD / Euro, el% cambio en el mercado estadounidense, el cambio porcentual en el mercado británico, y el cambio porcentual en el mercado alemán.

Solución:

(a) Este escenario es un problema de regresión, puesto que los datos y las variables están intrinsecamente relacionadas esto nos daría un modelo mejor aproximado del salario del CEO utilizando un modelo de regresión lineal.

Como el interés es sobre los factores que afectan el salario del CEO, entonces estariamos más interesados en hacer inferencia sobre el modelo del salario del CEO, i.e., establecer un modelo con el cual las variables predictoras nos puedan decir como es afectado el salario del CEO.

En este caso $n = 500$ empresas en los EE. UU.

p : Beneficio, número de empleados e industria.

(b) Este escenario es un problema de clasificación, ya que podemos hacer una separaciuón y clasificación de las variables, pero no una relación estricta entre ellas. Nos interesa hacer predicción, ya que queremos predecir el éxito o el fracaso del nuevo producto.

n - 20 productos similares lanzados previamente.

p - precio cobrado, presupuesto de marketing, precio de la competencia y otras diez variables.

(c) Este escenario es un problema de regresión, puesto que la variable de respuesta o de salida es cualitativa y es el porcentaje de cambio, nos interesa hacer predicciones sobre el porcentaje de cambio.

n - 52 semanas de datos semanales 2012.

p- El porcentaje de cambio en el mercado estadounidense, el porcentaje de cambio en el mercado británico y el porcentaje de cambio en el mercado alemán.

3. Ahora volvemos a visitar la descomposición de sesgo-varianza.

(a) Proporcione un bosquejo de sesgo típico (cuadrado), varianza, error de entrenamiento, error de prueba y curvas de error de Bayes (o irreductibles), en una sola gráfica, a medida que pasamos de métodos de aprendizaje estadístico menos flexibles hacia enfoques más flexibles. El eje x debe representar la cantidad de flexibilidad en el método, y el eje y debe representan los valores para cada curva. Debería haber cinco curvas. Asegúrese de etiquetar cada una.

(b) Explique por qué cada una de las cinco curvas tiene la forma mostrada en parte (a).

Solución:

4. Ahora piense en algunas aplicaciones de la vida real para el aprendizaje estadístico.

(a) Describa tres aplicaciones de la vida real en las que la clasificación podría sé útil. Describa la respuesta, así como los predictores. Es el objetivo de cada aplicación de inferencia o predicción? Explica tu responder.

(b) Describa tres aplicaciones de la vida real en las que la regresión podría sé útil. Describa la respuesta, así como los predictores. Es el objetivo de cada aplicación de inferencia o predicción? Explica tu responder.

(c) Describa tres aplicaciones de la vida real en las que el análisis de conglomerados podría ser útil.

Solución:

(a) (i). La dirección del precio del mercado de valores,

Objetivo: predicción.

Variable de respuesta: Arriba, abajo.

Variables de entrada: Cambio de porcentaje del movimiento del precio de ayer, el porcentaje del movimiento del precio del día anterior, el porcentaje de cambio, etc.

(ii). Clasificación de enfermedad.

Objetivo: inferencia.

Variable de respuesta: enfermo, saludable.

Variables de entrada: Descanso frecuencia cardíaca, frecuencia respiratoria en reposo, tiempo de ejecución

(iii) El reemplazo de piezas de automóvil.

Objetivo: Predicción.

Variable de respuesta: Necesita ser reemplazado o es bueno.

Variables de entrada: Edad de la parte, kilometraje utilizado, amperaje actual.

(b) (i) El Salario CEO.

Objetivo: inferencia.

Variables predictoras: Edad, experiencia en la industria, industria, Años de educación.

Variable de respuesta: salario.

(ii). El Reemplazo de autopartes.

Objetivo: inferencia.

Variable de respuesta: vida de la pieza del automóvil.

Variables predictoras: Edad de parte, kilometraje utilizado, amperaje actual.

(iii) clasificación de enfermedad.

Objetivo: Predicción.

Variable de respuesta: edad de muerte.

Variables de entrada: Edad actual, sexo, frecuencia cardíaca en reposo, frecuencia respiratoria en reposo, carrera de millas hora.

(c) (i). agrupación de tipo de cáncer.

Objetivo: Diagnosticar tipos de cáncer con mayor precisión.

(ii) Recomendaciones de películas de Netflix.

Objetivo: Recomendar películas basadas en usuarios que vieron y calificaron películas similares.

(iii) Encuesta de marketing. agrupación de datos demográficos para un producto (s).

Objetivo: Ver qué grupos de consumidores compran esos productos.

5. ¿Cuáles son las ventajas y desventajas de uno muy flexible (versus un enfoque menos flexible) para la regresión o clasificación? Bajo que circunstancias podrían preferirse un enfoque más flexible a un enfoque menos ¿acercamiento flexible? ¿Cuándo podría preferirse un enfoque menos flexible?

Solución:

La ventaja de un enfoque muy flexible para la regresión o clasificación es la obtención de un mejor ajuste para los modelos no lineales, disminuyendo el sesgo.

Las desventajas de un enfoque muy flexible para la regresión o clasificación es el hecho de que se requiere estimar un mayor número de parámetros siguiendo los puntos demasiado de cerca (overfit), aumentando la varianza.

Se preferiría un enfoque más flexible que uno menos flexible cuando estamos interesados en la predicción y no en la interpretabilidad de los resultados.

Se preferiría un enfoque menos flexible a un enfoque más flexible cuando están interesados en la inferencia y la interpretabilidad de los resultados.

6. Describa las diferencias entre un enfoque de aprendizaje estadístico paramétrico y uno no paramétrico. ¿Cuáles son las ventajas de un enfoque paramétrico de regresión o clasificación (en oposición a un enfoque no paramétrico)? ¿Cuáles son sus desventajas?

Solución:

Un enfoque paramétrico reduce el problema de estimar f a uno de estimar un conjunto de parámetros porque asume una forma para f.

Un enfoque no paramétrico no asume una forma funcional para f y así requiere una gran cantidad de observaciones para estimar con precisión f.

Las ventajas de un enfoque paramétrico de regresión o clasificación son las simplificación del modelado f a unos pocos parámetros y no tantas observaciones son requerido en comparación con un enfoque no paramétrico.

Las desventajas de un enfoque paramétrico de regresión o clasificación son un potencial para estimar incorrectamente f si se supone que la forma de f es incorrecta o para sobreajustar las observaciones si se utilizan modelos más flexibles.

7. La siguiente tabla proporciona un conjunto de datos de entrenamiento que contiene seis observaciones, tres predictores y una variable de respuesta cualitativa.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Rojo

Supongamos que deseamos usar este conjunto de datos para hacer una predicción para Y cuando $X_1 = X_2 = X_3 = 0$ usando K-vecinos más cercanos.

(a) Calcule la distancia euclidiana entre cada observación y el punto de prueba, $X_1 = X_2 = X_3 = 0$.

(b) ¿Cuál es nuestra predicción con $K = 1$? ¿Por qué?

(c) ¿Cuál es nuestra predicción con $K = 3$? ¿Por qué?

(d) Si el límite de decisión de Bayes en este problema es altamente no lineal, entonces ¿esperaríamos que el mejor valor para K sea grande o pequeño? ¿Por qué?

Solución:
(a)

Obs.	X_1	X_2	X_3	Distancia euclidiana	Y
1	0	3	0	$\sqrt{(0-0)^2+(3-0)^2+(0-0)^2}=3$	Rojo
2	2	0	0	$\sqrt{(2-0)^2+(0-0)^2+(0-0)^2}=2$	Rojo
3	0	1	3	$\sqrt{(0-0)^2+(1-0)^2+(3-0)^2}=\sqrt{10}\approx 3.2$	Rojo
4	0	1	2	$\sqrt{(0-0)^2+(1-0)^2+(2-0)^2}=\sqrt{5}\approx 2.23$	Verde
5	-1	0	1	$\sqrt{(-1-0)^2+(0-0)^2+(1-0)^2}=\sqrt{2}\approx 1.4142$	Verde
6	1	1	1	$\sqrt{(1-0)^2+(1-0)^2+(1-0)^2}=\sqrt{3}\approx 1.73$	Rojo

(b) La predicción para la respuesta Y es Verde, ya que la observación número 5 es el vecino más cercano para $K = 1$, esto se debe a que la primera distancia euclideana más pequeña es 1.41 que corresponde a la observación 5.

(c) La predicción para la respuesta Y es Rojo, ya que las observaciones número 2, 5 y 6 son los vecinos más cercanos para $K = 3$, esto se debe a que las 3 primeras distancias euclideanas más pequeñas son 1.41, 1.73 y 2.25 las cuales corresponden a las observaciones 2, 5 y 6.

2 es rojo, 5 es verde y 6 es rojo.

(d) Esperaríamos que el mejor valor para K sea pequeño, ya que una K pequeña sería flexible para un límite de decisión no lineal, mientras que una K grande intentaría ajustarse a un límite más lineal ya que requiere más puntos en consideración.

9. Este ejercicio involucra el conjunto de datos Auto estudiado en el laboratorio. Asegurarse que los valores faltantes se han eliminado de los datos.

(a) ¿Cuáles de los predictores son cuantitativos y cuáles son cualitativos?

(b) ¿Cuál es el rango de cada predictor cuantitativo? Puede responder esto usando la función range ().

(c) ¿Cuál es la media y la desviación estándar de cada predictor cuantitativo?

(d) Ahora elimine las observaciones 10 a 85. Cuál es el rango, media y desviación estándar de cada predictor en el subconjunto de los datos que quedan?

(e) Utilizando el conjunto de datos completo, investigue los predictores gráficamente, usando diagramas de dispersión u otras herramientas de su elección. Crea algunas gráficas destacando las relaciones entre los predictores. Comente sus hallazgos.

(f) Supongamos que deseamos predecir el consumo de combustible (mpg) sobre la base de las otras variables. ¿Sus gráficas sugieren que alguna de las otras variables pueden ser útiles para predecir mpg? Justifica tu respuesta.

Solución:

(a) Después de ingresar los datos en R y aplicar la función *str*() podemos verificar que de las variables predictoras 7 son cuantitativas y 2 son cualitativas.

(b) Utilizando la función range () obtenemos los siguientes resultados.

Predictor	Tipo	Rango
mpg	Cuantitativa	9.0 - 46.6
cylinders	Cuantitativa	3 - 8
displacement	Cuantitativa	68 - 455
horsepower	Cualitativa	
weight	Cuantitativa	1613 5140
acceleration	Cuantitativa	8.0 - 24.8
year	Cuantitativa	70 - 82
origin	Cuantitativa	1 - 3
name	Cualitativa	

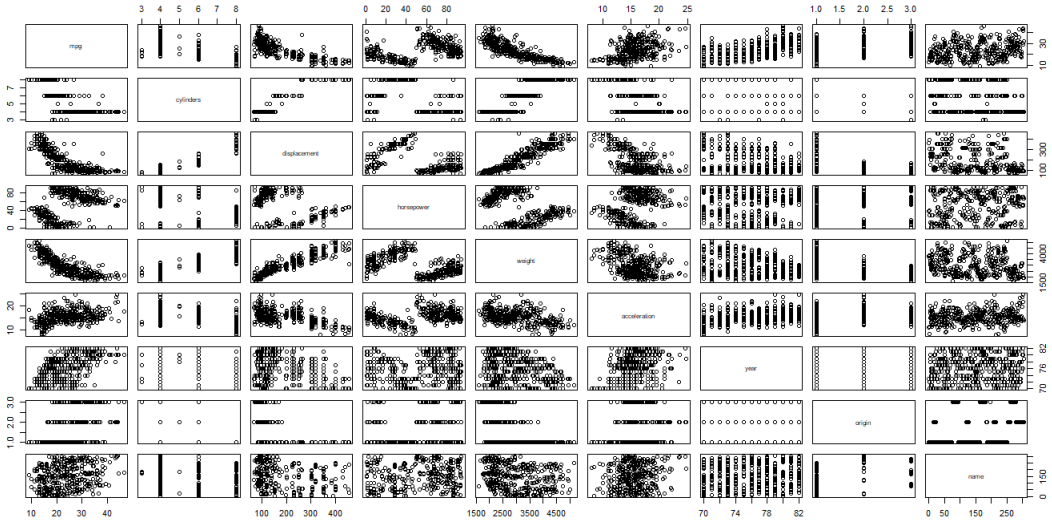
(c) Utilizando las funciones $mean()$ y $var()$ sobre los datos obtenemos los siguientes resultados.

Predictor	Tipo	Media	Varianza
mpg	Cuantitativa	23.51587	61.24321
cylinders	Cuantitativa	5.458438	2.895364
displacement	Cuantitativa	193.5327	10895.1
horsepower	Cualitativa		
weight	Cuantitativa	2970.262	718941.4
acceleration	Cuantitativa	15.55567	7.562474
year	Cuantitativa	75.99496	13.61614
origin	Cuantitativa	1.574307	0.6440857
name	Cualitativa		

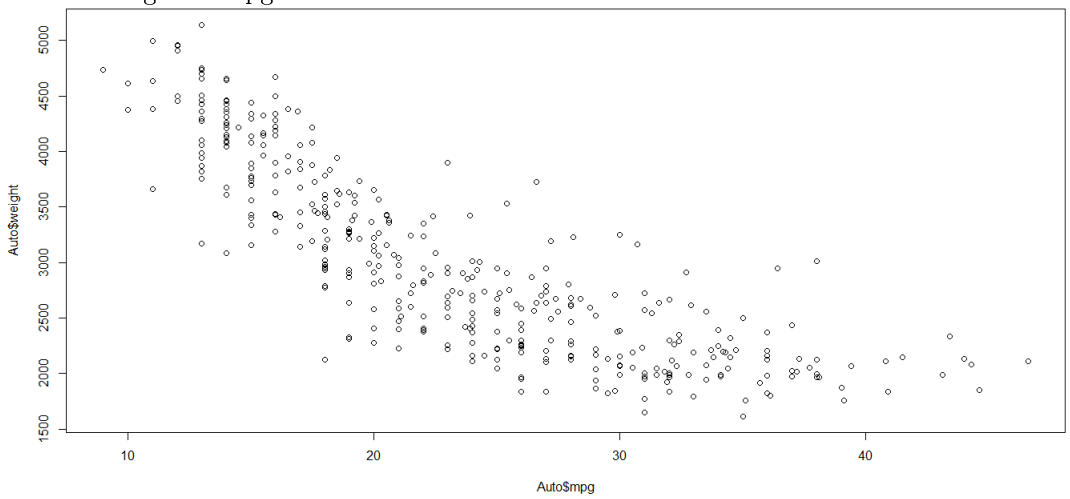
(d)

Predictor	Tipo	Rango	Media	Varianza
mpg	Cuantitativa	11 - 46.6	24.47913	62.21772
cylinders	Cuantitativa	3 - 8	5.3582	2.718127
displacement	Cuantitativa	68 - 455	185.9346	9726.936
horsepower	Cualitativa			
weight	Cuantitativa	1649 - 4997	2926.72	652315.3
acceleration	Cuantitativa	8.5 - 24.8	15.7433	7.084275
year	Cuantitativa	70 - 82	77.15888	9.602804
origin	Cuantitativa	1 - 3	1.604361	0.6711059
name	Cualitativa			

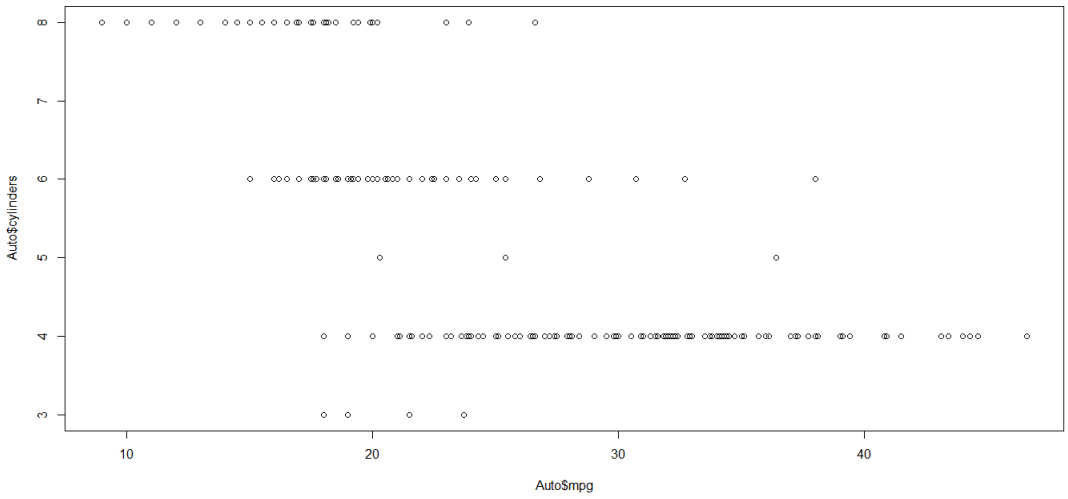
(e)
Diagrama de dispersión.



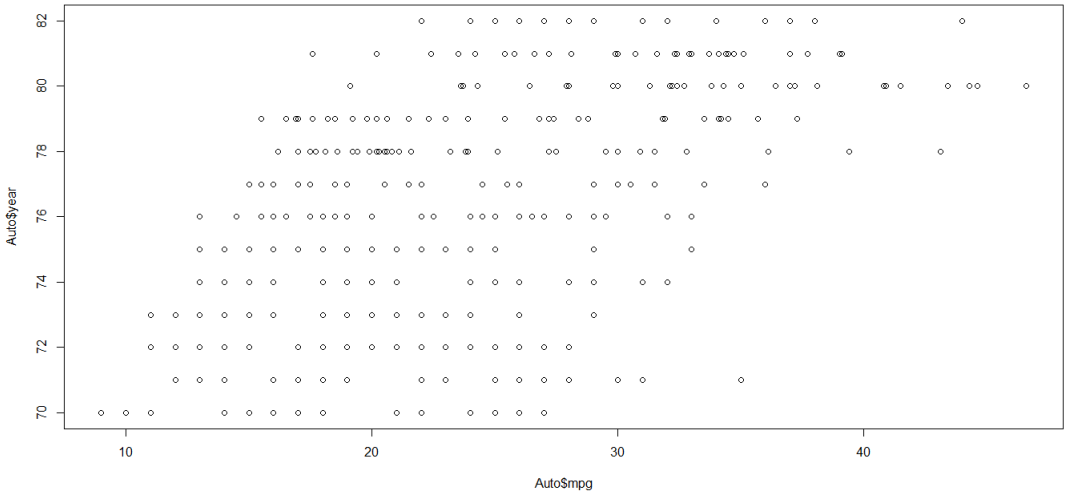
Gráfica Weight vs mpg.



Gráfica Cylinders vs mpg.



Gráfica Years vs mpg.



Después de observar las gráficas podemos concluir que los autos se vuelven más eficientes con el tiempo.

(f)

Todos los predictores muestran cierta correlación con mpg. Sin embargo, el nombre del predictor tiene muy pocas observaciones por nombre, por lo que es probable que usarlo como un predictor resulte en un overfit (sobreajuste) de los datos y no se generalice bien.