Universidad de los Andes
Facultad de Ingeniería

TÓPICOS AVANZADOS EN ESTADÍSTICA iIND-4114
Métodos de Aprendizaje Estaístico y Machine Learning
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Periodo 2015-20. Profesor: Carlos Valencia.

TAREA 1. Introducción y Modelos Lineales para Predicción.

* El informe con la solución debe ser entregado de manera individual o en parejas.
* Fecha de entrega: Jueves 3 de Septiembre del 2015 antes de clase.
* El informe debe contener procedimientos explícitos.

Para todos los problemas que se describen a continuación se asume que existe una variable de respuesta Y que se quiere predecir. Los predictores se encuentran contenidos en el vector $X = (X_1, \cdots, X_p) \in \mathbb{R}^p$. Si Y es categórica, entonces se tiene un problema de clasificación, mientras que si Y es continua, se tiene un problema de regresión. En cada caso, $f^* : \mathbb{R}^p \to \mathcal{Y}$ corresponde a la función desconocida que se quiere estimar. Para el problema de regresión, $f^*(X) = \mathbb{E}(Y|X)$, y para el de clasificación $f^*(X)$ es el clasificador de Bayes. Los datos, se suponen que son realizaciones independientes e idénticamente distribuídas de variables aleatorias $(Y_i, X_i) \sim \mathcal{P}_{YX}$, para $i = 1, \cdots, n$.

Parte A. Problemas Teóricos, Conceptuales y Experimentales.

1. Se desea predecir una respuesta continua Y, para lo cual se plantea el modelo de regresión:

$$Y = f^*(X) + \epsilon$$

donde $X = (X_1, \cdots, X_p) \in \mathbb{R}^P$ y ϵ es independiente de X. El modelo de regresión implica que $f^*(X) = \mathbb{E}(Y|X)$ es el mejor predictor posible. Para estimar f^* se plantea un modelo de regresión lineal

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon_R \,,$$

donde $\beta = (\beta_0, \cdots, \beta_p) \in \mathbb{R}^{p+1}$ se estima por mínimos cuadrados a partir de una muestra de tamaño n. El estimador resultante es:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

donde

$$\mathbf{Y}^T = (Y_1, \cdots, Y_n) \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ 1 & X_{n1} \cdots & X_{np} \end{bmatrix}$$

Para este problema, suponga que $\mathbf{X}$ es determinística, esto es, los predictores no son aleatorios sino que están dados (esto se hace por facilidad en la solución).

   a. Encuentre $\mathbb{E}(\epsilon)$ y $\mathbb{E}(\epsilon_R)$
   b. Definiendo $\text{Var}(\epsilon) = \sigma^2$ y $\text{Var}(\epsilon_R) = \sigma_R^2$, cuál de los dos es el error irreducible?

c. Demuestre que $\sigma_R^2 = \sigma^2 + \mathbb{E}(\epsilon_R)^2$

d. Es verdad que si se utilizan las primeras $p-1$ variables predictoras en el modelo, entonces $\mathbb{E}(\epsilon_R)^2$ debe disminuir?. Justifique.

e. Suponga que ahora f^* se estima como:

$$\hat{f}(X) = \hat{f}_1(X_1) + \cdots + \hat{f}_p(X_p)$$

donde cada $\hat{f}_j$ es un suavizador local, esto es, un esimador más flexible que el modelo lineal. Es $\hat{f}(x^{new})$ una mejor predicción que la del modelo lineal?. De qué depende que así sea.

2. Se busca predecir una respuesta continua Y a través de un predictor $X \in \mathbb{R}$. Suponga que las observaciones de X son determinísticas y equidistantes en el intervalo $[0,1]$, esto es, $x_i = \frac{i-1}{n-1}$, para $i = 1, \cdots, n$. En la regresión $Y = f^*(X) + \epsilon$, se sabe que el error irreducible es $\text{Var}(\epsilon) = \sigma^2$. Se define un nuevo estimador para f^* como

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in N_k(x)} Y_i \, ,$$

para $x \in [0,1]$, donde $N_k(x) \in \{1, \cdots, n\}$, es el conjunto de k índices correspondientes a los k x_i que son más próximos a x. Por ejemplo, si $x = 0.5$ y $k = 3$ y $n = 5$, entonces $N_k = \{2, 3, 4\}$.

a. Cómo se puede controlar la flexibilidad del modelo?.

b. Qué pasa si $k = n$?. Qué pasa si $k = 1$?

c. Suponga que $f^*(x) = ax$ para una constante $a \in \mathbb{R}$ y que $n = 11$. Si $k = 3$, determine

$$\mathbb{E}(MSE\hat{f}(0.5)) = \mathbb{E}\left(\hat{f}(0.5) - Y^{0.5}\right)$$

done $Y^{0.5}$ es la variable Y cuando $X = 0.5$.

d. Ahora suponga que en literal (c) se usa $k = 5$. Calcule el $\mathbb{E}(MSE\hat{f}(0.5))$. Interprete el cambio.

e. Ahora suponga que en el literal (c), $f^*(x) = ax^2$. Calcule el $\mathbb{E}(MSE\hat{f}(0.5))$ para $k = 3$ y $k = 5$. Interprete.

3. Ahora se experimenta con datos simulados. Se va simular un problema de regresión como lo hicimos en clase, con parámetros f^* (a la que se le puede cambiar su complejidad), el tamaño de muestra (n) y el error irreducible (σ). Para simular una muestra, use

```
#You can try to modify the denominator inside the cosine to
#change the complexity of the function
  f=function(x){
    y=2+x^(.2)*cos(x/.15)/x^-.45
    return(y)
  }
  plot(f,0,5)
#Points simulation: you change n and sigma
  N=300
```

```
  sigma=1.2
  x=runif(N,0,5);x=sort(x)
  y=rep(0,times=N)
  for(i in 1:N){
    y[i]=f(x[i])+rnorm(1,0,sigma)
  }
  plot(x,y)
  points(x,f(x),type="l",col=2,lwd=2)
```

El algoritmo para estimar f^* será el de k–vecinos, descrito en el punto anterior. Para encontrar $\hat{f}$ puede utilizar el código:

```
#Estimator by k-neigbor
#k es el numero de vecinos y test=TRUE si se estima sonbre los puntos x
kn=function(k,test){
  if(test=="FALSE"){z=seq(0,5,by=0.01);ll=length(z)}
  if(test=="TRUE"){z=x;ll=length(z)}
  nk=rep(0,times=ll)
  for(j in 1:ll){
    veci=which(abs(z[j]-x) %in% sort(abs(z[j]-x))[1:k])
    nk[j]=sum(y[veci])/k
  }
  return(nk)
}
```

Y para graficr el resultado, use las siguientes líneas, donde la curva roja es f^* (que en problemas reales es desconocida) y la curva azul es $\hat{f}$. Pruebe a cambiar el parámetro k qu controla qué tan suave es la función:

```
#Graficando el estimador
  k=14  #Puede cambiarlo
  plot(x,y)
  points(x,f(x),type="l",col=2,lwd=2)
  points(z,kn(k,F),type="l",col=4,lwd=2)
```

a. Relice la simulación usando un tamaño de muestra $n = 400$, donde 300 datos serán usandos para entrenamiento y 100 para prueba (*test*). Usted es libre de cambiar la función f^* haciéndola más o menos compleja, asi como de cambiar el error irreducible. Usando la muestra de *test*, calcule el MSE para cada valor de $k \in \{1, \cdots, 200\}$. Seleccione el nivel de k (que mide la flexibilidad de $\hat{f}$) que minimiza el MSE. Grafique el MSE en función de k.

b. Repita el experimento, pero usando $n = 500$, 300 para entrenamiento y 200 para prueba. Debe cambiar la respuesta?. Cuál es la diferencia con el experimento del literal (a)?. Cuál es mejor?

c. Ahora resuelva el experimento del literal (a) usando "10-fold Crossvalidation" y "Leave-one-out Crossvalidation" con los 300 datos de la muestra. Grafique el MSE estimado para cada k. Compare los resultados y comente.

4. Entendiento el efecto de la multicolinealidad para la selección de modelos (predictivos): Se simulan predictores X_1 y X_2 correlacionados entre si, para predecir Y:

```
#Generacion de variables
  x1=rnorm(100,0,4)
  x2=-1.5*x1+rnorm(100,0,3)
  y=10+2*x1+rnorm(100,0,3)
  plot(x1,y)
  plot(x2,y)
  plot(x1,x2)
```

Note que cada una de las variables predictoras se relaciona linealmente con la respuesta Y.

a. Corra el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ usando:

```
x=cbind(x1,x2)
fit=lm(y~x);   summary(fit)
```

Tienen sentido los coeficientes estimados $\hat{\beta}_1$ y $\hat{\beta}_2$. Explique lo que está pasando.

b. Ahora corra un modelo simple para cada variable predictora usando:

```
fit1=lm(y~x1);  summary(fit1)
fit2=lm(y~x2);  summary(fit2)
```

Comparando los betas estimados, cuál es la diferencia con el modelo del literal anterior?. Comente porqué pasa esto.

c. Calcule el Cp de Mallows y el BIC para cada uno de los tres modelos. Cuál de los tres utilizaría para predecir Y?. Cuál de los tres utilizaría para explicar Y?

Para entender el efecto de la colinealidad entre los predictores, se van a rotar usando la siguiente matriz de rotación parametrizada en términos del ángulo de rotación en redianes (theta):

```
hh=function(theta){
mm=matrix(c(cos(theta),-sin(theta),sin(theta),cos(theta)),ncol=2)
return(mm)}
```

Ahora se rotan los puntos (x_1, x_2)para buscar que no haya correlación entre ellos:

```
theta=1
  zz=x%*%hh(theta)
  plot(zz,xlim=c(-20,20),ylim=c(-20,20)) #puntos rotados
  plot(x,xlim=c(-20,20),ylim=c(-20,20))  #puntos originales
```

d. Pruebe diferentes valores de `theta` hasta encontrar (z_1, z_2) que tenga correlación cercana a cero. Puede ayudarse midiendo la correlación como `cor(zz)`.

e. Repita los procedimientos de los literales a-c, pero utilizando como predictores las nuevas variables. En lugar de `x1` puede usar `zz[,1]` y `zz[,2]` en lugar de `x2` en el código. Qué diferencias encuentra?

f. Construya intervalos de predicción para la respuesta Y en cada uno de los puntos muestrales $(x_1, x_2)_i$, usando:

```
predict(lm(y~x),interval="predict")
predict(lm(y~zz),interval="predict")
```

Comente sobre este resultado. Qué significa?

g. Ahora simule tres nuevas variables predictivas correlacionadas y regreselas contra Y:

```
x3=1*y+rnorm(100,0,4)
x4=-6*x2+rnorm(100,0,3)
x5=5*x1+rnorm(100,0,5)
x=cbind(x1,x2,x3,x4,x5)
fit=lm(y~x);  summary(fit)
```

Para entender cómo están las variables relacionadas entre si, genere gráficos de dispersiones por pares con:

```
pairs(cbind(y,x))  #Se pued apreciar la redundancia de informacion entre predictores
```

Ahora se van a transformar las $X's$ en una nueva matriz de datos no correlacionados Z usando componentes principales:

```
ppp=princomp(x)
z=ppp$scores
fitpca=lm(y~z);   summary(fitpca)
```

Encuentre el mejor modelo lineal predictivo para Y usando las variables originales y el mejor modelo predictivo usando los componentes principales. Compare los dos modelos resultantes. Cuál es mejor?. Comente.

h. Ahora simule otra muestra de predictores $(X_1, \cdots, X_5)$ de tamaño 100 para usar como test (tenga cuidado en camiar el nombre de las variables para no confundir los modelos). Usando el MSE calculado en en la muestra de test, escoga el mejor modelo para predecir Y entre:

i Selección secuencial de variables originales tipo *forward*
ii Selección exhaustiva de submodelos (combinatorio)
iii Rregresión por Componentes Principales
iv Regersión por *Partial Least Squares*
v Modelo pealizado tipo *Ridge*
vi Modelo pealizado tipo *Lasso*

i. Si además de predecir, se quiere interpretar el efecto (pendiente) que tiene cada X_j sobre la respuesta Y, cuál de los seis modelos en (h) usaría?. Justifique.

5. Realice los siguientes ejercicios del libro *"An Introduction to Statistical Learning"*:

Cap.2: Ejercicios 1, 2, 3. Se recomienda hacer el ejercicio 8 (no se evalua) porque le sirve para otras partes de la tarea.

Cap.3: Ejercicios 3, 4 y 15.

Cap.6: Ejercicios 2, 3 y 4.

Parte B. Problemas Prácticos (con datos).

1. Resuelva el ejercicio 9 del capítulo 6 del libro *"An Introduction to Statistical Learning"*. Para entender un poco más los datos, refiérase all ejercicio 8 del capítulo 2.

2. Consiguiendo datos reales (un caso sencillo). En el repositorio para MAchine Learning de la Universidad de California Irvine, se ecuentran una serie de bases de datos que sirven para probar algoritmos de predicción. En este caso, trabajaremos con uno muy sencillo. Vaya a la página `http://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE#` y obtenga la base de datos. El problema es predecir el valor del *Istanbul Stock Exchange* con base en los demás índices financieros. Los datos están en Excel, así que es recomendado primero, eliminar las columnas y filas redundantes. Una opción es copiar los datos a un archivo de texto (ojo con el separador decimal) con la primera fila para los nombres. Por ejemplo, usando:

```
> stock=read.table("stocks.txt",header=T)
> head(stock)
           ISE           SP          DAX         FTSE       NIKKEI     BOVESPA
1  0.038376187 -0.004679315  0.002193419  0.003894376  0.000000000  0.03119023
2  0.031812743  0.007786738  0.008455341  0.012865611  0.004162452  0.01891958
3 -0.026352966 -0.030469134 -0.017833062 -0.028734593  0.017292932 -0.03589858
4 -0.084715902  0.003391364 -0.011726277 -0.000465999 -0.040061309  0.02828315
5  0.009658112 -0.021533208 -0.019872754 -0.012709717 -0.004473502 -0.00976388
6 -0.042361155 -0.022822626 -0.013525735 -0.005025533 -0.049038532 -0.05384947
            EU           EM
1  0.012698039  0.028524462
2  0.011340652  0.008772644
3 -0.017072795 -0.020015412
4 -0.005560959 -0.019423778
5 -0.010988634 -0.007802212
6 -0.012451259 -0.022629745
```

   Encuentre el mejor modelo lineal para predecir el ISE, usando lo que hemos visto en clase.