Taller de Aprendizaje de Máquina: Reducción de dimensión

Julián D. Arias Londoño
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Medellín, Colombia
jdarias@udea.edu.co

May 16, 2017

1 Ejercicios

1. Utilice el coeficiente de correlación de Pearson a través de la función `corrcoef` y calcule la correlación entre las 18 variables del problema de predicción del resultado de la simulación de un modelo de clima (satisfactorio o falla) , disponible en: `http://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes#`. Los datos están incluidos en el archivo `DatosSeleccion.mat`. Calcule también el índice de Fisher entre cada una de las características con respecto a las variables a predecir. Analice los resultados y concluya si Ud considera que hay variables que podrían eliminarse.

   Recuerde que el índice de Fisher está dado por [1]:

$$F = \sum_{i}^{C} \sum_{j \neq i}^{C} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}$$

   donde C es el número de clases, μ_i y σ_i son la media y la desviación estándar de la clase i respectivamente.

   Respuesta:

2. Realice la selección de características utilizando el método de búsqueda secuencial hacia adelante (SFS) implementado en la función de Matlab `sequentialfs`. Utilice una función criterio tipo Wrapper a partir de un modelo Random Forest. Seleccione el número óptimo de árboles, antes de llevar a cabo la selección de variables. La ayuda de la función `sequentialfs` especifica los requerimientos de la función criterio. Sin embargo como ayuda, la siguiente es una función válida:

```
1
2 function Error=Criterio(Xtrain,Ytrain,Xtest,Ytest)
3
4 Yest = classify(Xtest,Xtrain,Ytrain);% Se usa una
    Funcion Discriminante Gausiana como criterio
5 Error = sum(Ytest ~= Yest)/length(Yest);
```

En el caso anterior se está usando un modelo de Función Discriminante Gausiana.

Responda:

- ¿Cuál es el nivel de eficiencia alcanzado con el total de variables?

- ¿Cuál es el porcentaje de reducción alcanzado?

- ¿Cuál es el nivel de eficiencia alcanzado con el subconjunto de características seleccionado?

- ¿Cuales variables (por índice), coinciden con las candidatas a eliminarse según los resultados del punto anterior?

3. Realice la selección de características utilizando el método de búsqueda por algoritmos genéticos. Use la función de GA que acompaña esta guía. Realice los cambios necesarios sobre la función criterio del punto anterior, para que pueda ser usada en este caso. **Ayuda**: La función de fitness basada en la función anterior sería:

```matlab

function fitnessVals=FitnessSelection(pop,X,Y)

CVO = cvpartition(size(X,1),'k',4);
Ncromosomas = size(pop,1);
Costos = zeros(1,Ncromosomas);
for i = 1: Ncromosomas % Se debe evaluar cada individuo de la poblacion
    X2 = X(:,pop(i,:)); % Se usa el subcojunto de variables descritas por el individuo i
    Error = zeros(1,4);
    for j = 1:4 %Numero de Folds
        Xtrain = X2(CVO.training(j),:);
        Xtest = X2(CVO.test(j),:);
        Ytrain = Y(CVO.training(j));
        Ytest = Y(CVO.test(j));
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        Yest = classify(Xtest,Xtrain,Ytrain);% Se usa una Funcion Discriminante Gausiana como criterio
        Error(j) = sum(Ytest ~= Yest)/length(Yest);
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    end
    Costos(i) = mean(Error);
end

%Calculo del fitness a partir de la funcion de evaluacion
[~,indice] = sort(Costos);
fitnessVals = 0.9.^(0:1:size(pop,1)-1);
fitnessVals(indice)=fitnessVals;
```

En este caso es necesario implementar la estrategia de validación cruzada porque el algoritmo genético, a diferencia del método de búsqueda secuencial, no lo tiene implementado.

Responda:

- ¿Cuál es el porcentaje de reducción alcanzado?

- ¿Cuál es el nivel de eficiencia alcanzado con el subconjunto de características seleccionado?

- ¿Cuales variables (por indice), coinciden con las candidatas a eliminarse según los resultados del punto 1?

- ¿Cuales variables (por indice), coinciden con las candidatas a eliminarse según los resultados del punto 2?

4. Aplique la transformación PCA (función de Matlab `pca`) al problema de clasificación. Defina los componentes principales que deberían incluirse para un 85%, 90% y un 95% de la varianza acumulada. Llene la siguiente tabla:

% de varianza	% de reducción	Número de arboles	Eficiencia
85%			
90%			
95%			

5. * Use las funciones `lassoglm y lassoPlot` para llevar a cabo la selección de variables utilizando la técnica LASSO. Un ejemplo de uso en un problema de clasificacion puede ser consultado en el siguiente enlace `http://www.mathworks.com/help/stats/regularize-logistic-regression.html`

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* New Jersey, NY, USA: Wiley-Interscience, 2nd ed., 2000.