Universidad Técnica Federico Santa María
Departamento de Informática

Máquinas de Aprendizaje Computacional (INF-393)

Semestre II 2015 - Casa Central

Tarea 1

Profesor: Carlos Valle

Instrucciones: Para realizar cada punto, primero se deben generar 20 datasets para cada problema (basados en los archivos "*cereales.data*" y "*credit.data*". A continuación se debe reportar el boxplot de los resultados sobre los 20 datasets, o en su defecto, la media y la varianza de dichas experimentaciones. Considere la función MATLAB *generating_datasets(dataname)* para generar los conjuntos de datos respectivos.
Para la estimación de parámetros, se debe elegir 5 posibles valores para el parámetro a estimar, y usando 5- cross validation se debe estimar el óptimo para cada data set (de un total de 20).
Para el informe basta que incluyan 2 de los 20 casos, y junto a ello un histograma con el número de veces que fue seleccionado cada valor posible del parámetro (Por supuesto que la suma de las frecuencias absolutas de cada valor deben sumar 20).
Para realizar las comparaciones usando training/ test sets. Si el algoritmo tiene parámetros, debe usar el alfa óptimo de cada data set.
I Regresión Lineal: Un laboratorio de análisis de alimentos asignó puntajes a 100 cereales que se venden en el mercado. Sin embargo, dicho laboratorio no revela los criterios utilizados. Para tratar de emular la fórmula que asigna los puntajes y poder predecir el puntaje de nuevos cereales, consideraremos un modelo lineal dependiente de las siguientes variables:

1. Calorías por porción.
2. Gramos de proteína.
3. Gramos de grasa.
4. Miligramos de sodio.
5. Gramos de fibra dietética.
6. Gramos de carbohdratos complejos.
7. Gramos de azúcares.
8. Miligramos de potasio.
9. Porcentaje de Vitaminas y minerales.
10. Estante de exhibición (1, 2, or 3, contando desde el suelo).
11. Peso en onzas de una porción.
12. Número de tazas de una porción.
13. **Puntaje del cereal** (variable de salida).

En el archivo "*cereales.data*" se encuentra la información de 100 productos de este tipo. Utilice los primeros 75 para entrenar los modelos y los 25 restantes como datos de prueba.

1. Estime los parámetros de la regresión lineal usando:

   a) Gradiente descendente batch. [5 %]
   b) Gradiente descendente online. [5 %] **(Solo postgrado)**
   c) Newton-Raphson. [15 %]

2. Obtenga el E.C.M. en el training set y test set para cada algoritmo del ejercicio anterior. Compare los resultados y comente. [10 %]
3. Prediga el puntaje de los cereales que se encuentran en el test set usando *Locally weighted linear regression*. Estudie el comportamiento del parámetro τ usando 5-fold cross validation sobre el training set. [10 %]
4. Compare y comente los resultados en el punto anterior con los obtenidos en el punto 2. [10 %]
5. Repita los pasos 1, 2, 3 y 4 normalizando o estandarizando los datos. (Si normalizó la salida, desnormalícela para calcular el E.C.M.) ¿Las transformaciones de los datos afectan la calidad de la predicción? [10 %] **(Postgrado debe hacer ambas)**

II Clasificación binaria: Una entidad bancaria desea evaluar a los clientes que solicitan un crédito. En el archivo "*credit.data*" se encuentra la información bancaria de 120 personas donde cada columa representa:

1. Salario Anual (Millones $).
2. Edad.
3. Dinero adeudado (Millones $).
4. Número de Hijos.
5. Casa propia.
6. Avalúo casa.
7. **Crédito** (bueno:1 / malo:0)(variable de salida).

Utilice los primeros 90 para entrenar los modelos y los 30 restantes como datos de prueba.

1. Estime los parámetros del modelo *logistic regression* usando:

   a) Gradiente descendente online. [5 %]
   b) Newton-Raphson. [10 %]

2. Obtenga el *error-rate* en el training set y test set para cada algoritmo del ejercicio anterior. Compare los resultados y comente. [10 %]
3. Repita los pasos 1 y 2 normalizando y estandarizando los datos. (Acá no es necesario desnormalizar la salida) ¿Las transformaciones de los datos tienen efecto en la clasificación? [10 %] **(Postgrado debe hacer ambas)**

Fecha de Entrega: 3 de Noviembre

Recuerde el **Código de Honor** establecido en esta asignatura.