APRENDIZAJE AUTOMÁTICO

Trabajo.2: Programación

Fecha límite de entrega: 26 de Abril 2017

Valoración máxima: 12 puntos + BONUS

NORMAS DE DESARROLLO Y ENTREGA DE TRABAJOS

Para este trabajo como para los demás es obligatorio presentar un informe escrito con sus valoraciones y decisiones adoptadas en el desarrollo de cada uno de los apartados. Incluir en el informe los gráficos generados. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. (obligatorio en pdf). **Sin este informe se considera que el trabajo NO ha sido presentado.**

Normas para el desarrollo de los Trabajos: EL INCUMPLIMIENTO DE ESTAS NORMAS SIGNIFICA PERDIDA DE 2 PUNTOS POR CADA INCUMPLIMIENTO.

- El código se debe estructurar en un único script R con distintas funciones o apartados, uno por cada ejercicio/apartado de la práctica.
- Todos los resultados numéricos o gráficas serán mostrados por pantalla, parando la ejecución después de cada apartado. No escribir nada en el disco.
- El path que se use en la lectura de cualquier fichero auxiliar de datos debe ser siempre "datos/nombre_fichero". Es decir, crear un directorio llamado "datos"dentro del directorio donde se desarrolla y se ejecuta la práctica.
- Un código es apto para ser corregido si se puede ejecutar de principio a fin sin errores.
- NO ES VÁLIDO usar opciones en las entradas. Para ello fijar al comienzo los parámetros por defecto que considere que son los óptimos.
- El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
- Poner puntos de parada para mostrar imágenes o datos por consola.
- Todos los ficheros (*.R *.pdf) se entregan juntos dentro de un único fichero zip, sin ningún directorio que los contenga.
- ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.
- **Forma de entrega**: Subir el zip al Tablón docente de CCIA.

1. Modelos Lineales

1. (3 puntos) **Gradiente Descendente**. Implementar el algoritmo de gradiente descendente.
   a) Considerar la función no lineal $E(u,v) = (u^2e^v - 2v^2e^{-u})^2$ Usar gradiente descendente y para encontrar un mínimo de esta función, comenzando desde el punto $(u,v) = (1,1)$ y usando una tasa de aprendizaje $\eta = 0{,}1$.
      1) Calcular analíticamente y mostrar la expresión del gradiente de la función $E(u,v)$
      2) ¿Cuántas iteraciones tarda el algoritmo en obtener por primera vez un valor de $E(u,v)$ inferior a 10^{-14} . (Usar flotantes de 64 bits)
      3) ¿Qué valores de (u,v) obtuvo en el apartado anterior cuando alcanzó el error de 10^{-14}.

   b) Considerar ahora la función $f(x,y) = (x-2)^2 + 2(y-2)^2 + 2\sin(2\pi x)\sin(2\pi y)$
      1) Usar gradiente descendente para minimizar esta función. Usar como punto inicial $(x_0 = 1, y_0 = 1)$, tasa de aprendizaje $\eta = 0{,}01$ y un máximo de 50 iteraciones. Generar un gráfico de cómo desciende el valor de la función con las iteraciones. Repetir el experimento pero usando $\eta = 0{,}1$, comentar las diferencias.
      2) Obtener el valor mínimo y los valores de las variables que lo alcanzan cuando el punto de inicio se fija: (2,1, 2,1), $(3,3)$,(1,5, 1,5),$(1,1)$. Generar una tabla con los valores obtenidos

   c) ¿Cuál sería su conclusión sobre la verdadera dificultad de encontrar el mínimo global de una función arbitraria?

2. (3 puntos) **Regresión Logística**: En este ejercicio crearemos nuestra propia función objetivo f (una probabilidad en este caso) y nuestro conjunto de datos $\mathcal{D}$ para ver cómo funciona regresión logística. Supondremos por simplicidad que f es una probabilidad con valores 0/1 y por tanto que la etiqueta y es una función determinista de $\mathbf{x}$.

   Consideremos $d = 2$ para que los datos sean visualizables, y sea $\mathcal{X} = [0,2] \times [0,2]$ con probabilidad uniforme de elegir cada $\mathbf{x} \in \mathcal{X}$. Elegir una línea en el plano que pase por $\mathcal{X}$ como la frontera entre $f(\mathbf{x}) = 1$ (donde y toma valores $+1$) y $f(\mathbf{x}) = 0$ (donde y toma valores -1), para ello seleccionar dos puntos aleatorios del plano y calcular la línea que pasa por ambos. Seleccionar $N = 100$ puntos aleatorios $\{\mathbf{x}_n\}$ de $\mathcal{X}$ y evaluar las respuestas $\{y_n\}$ de todos ellos respecto de la frontera elegida.
   a) Implementar Regresión Logística(RL) con Gradiente Descendente Estocástico (SGD) bajo las siguientes condiciones:
      - Inicializar el vector de pesos con valores 0.
      - Parar el algoritmo cuando $||\mathbf{w}^{(t-1)} - \mathbf{w}^{(t)}|| < 0{,}01$, donde $\mathbf{w}^{(t)}$ denota el vector de pesos al final de la época t. Una época es un pase completo a través de los N datos.
      - Aplicar una permutación aleatoria, $1, 2, \ldots, N$, en el orden de los datos antes de usarlos en cada época del algoritmo.
      - Usar una tasa de aprendizaje de $\eta = 0{,}01$

   b) Usar la muestra de datos etiquetada para encontrar nuestra solución g y estimar E_{out} usando para ello un número suficientemente grande de nuevas muestras (>999).

3. (3 puntos) **Clasificación de Dígitos**. Considerar el conjunto de datos de los dígitos manuscritos y seleccionar las muestras de los dígitos 4 y 8. Usar los ficheros de entrenamiento (training) y test que se proporcionan. Extraer las características de **intensidad promedio** y **simetría** en la manera que se indicó en el ejercicio 3 del trabajo 1.

a) Plantear un problema de clasificación binaria que considere el conjunto de entrenamiento como datos de entrada para aprender la función g.

b) Usar un modelo de Regresión Lineal y aplicar PLA-Pocket como mejora. Responder a las siguientes cuestiones.

1) Generar gráficos separados (en color) de los datos de entrenamiento y test junto con la función estimada.

2) Calcular E_{in} y E_{test} (error sobre los datos de test).

3) Obtener cotas sobre el verdadero valor de E_{out}. Pueden calcularse dos cotas una basada en E_{in} y otra basada en E_{test}. Usar una tolerancia $\delta = 0{,}05$. ¿Que cota es mejor?

4. (3 puntos). En este ejercicio evaluamos el papel de la **regularización en la selección de modelos**. Para $d = 3$ (dimensión del vector de características) generar un conjunto de N datos aleatorios $\{\mathbf{x}_n, y_n\}$ de la siguiente forma:

   - Las coordenadas de los puntos $\mathbf{x}_n$ se generarán como valores aleatorios extraidos de una Gaussiana de media 1 y desviación típica 1.
   - Para definir el vector de pesos $\mathbf{w}_f$ de la función f generamos $d + 1$ valores de una Gaussiana de media 0 y desviación típica 1. Al último valor le sumaremos 1.
   - Usando los valores anteriores generamos la etiqueta asociada a cada punto $\mathbf{x}_n$ a partir del valor $y_n = \mathbf{w}_f^T \mathbf{x}_n + \sigma\epsilon_n$, donde ϵ_n es un ruido que sigue también una Gaussiana de media 0 y desviación típica 1 y σ^2 es la varianza del ruido; fijar $\sigma = 0{,}5$

   Ahora vamos a estimar el valor de $\mathbf{w}_f$ usando $\mathbf{w}_{reg}$, es decir los pesos de un modelo de regresión lineal con regularización "weight decay". Fijar el parámetro de regularización a $0{,}05/N$.

   a) Para $N \in \{d + 10, d + 20, \ldots, d + 110\}$ calcular los errores de validación cruzada $\text{e}_1, \ldots, \text{e}_\text{N}$ y E_{cv}. Repetir el experimento 1000 veces. Anotamos el promedio y la varianza de e_1, e_2 y E_{cv} en los experimentos.

   b) ¿Cuál debería de ser la relación entre el valor promedio de e_1 y el de E_{cv} ? ¿y entre el valor promedio de e_1 y el de e_2? Argumentar la respuesta en base a los resultados de los experimentos.

   c) ¿Qué es lo que más contribuye a la varianza de los valores de e_1?

   d) Diga que conclusiones sobre regularización y selección de modelos ha sido capaz de extraer de esta experimentación.

1.1. BONUS

El BONUS solo se tendrá en cuenta si se ha obtenido al menos el 75 % de los puntos de la parte obligatoria. La máxima puntuación añadida por los BONUS es de 3 puntos.

1. (1 punto) **Coordenada descendente**. En este ejercicio comparamos la eficiencia de la técnica de optimización de "coordenada descendente" usando la misma función del ejercicio 1.1*a*. En cada iteración, minimizamos a lo largo de cada una de las coordenadas individualmente. En el Paso-1 nos movemos a lo largo de la coordenada u para reducir el error (suponer que se verifica una aproximación de primer orden como en gradiente descendente), y el Paso-2 es para reevaluar y movernos a lo largo de la coordenada v para reducir el error (hacer la misma hipótesis que en el paso-1). Usar una tasa de aprendizaje $\eta = 0{,}1$.

a) ¿Qué valor de la función $E(u, v)$ se obtiene después de 15 iteraciones completas (i.e. 30 pasos) ?

b) Establezca una comparación entre esta técnica y la técnica de gradiente descendente.

2. (1.5 puntos) **Método de Newton** Implementar el algoritmo de minimización de Newton y aplicarlo a la función $f(x, y)$ dada en el ejercicio.1*b*. Desarrolle los mismos experimentos usando los mismos puntos de inicio.
   - Generar un gráfico de como desciende el valor de la función con las iteraciones.
   - Extraer conclusiones sobre las conductas de los algoritmos comparando la curva de decrecimiento de la función calculada en el apartado anterior y la correspondiente obtenida con gradiente descendente.
3. (0.5 punto) Repetir el experimento de RL (punto.2) 100 veces con diferentes funciones frontera y calcule el promedio.
   a) ¿Cuál es el valor de E_{out} para $N = 100$?
   b) ¿Cuántas épocas tarda en promedio RL en converger para $N = 100$, usando todas las condiciones anteriormente especificadas?
4. (1 punto) Considere el ejercicio 3 de la sección de Modelos Lineales de los Ejercicios de Apoyo. Repetir los puntos del mismo pero usando una transformación polinómica de tercer orden($\Phi_3(\mathbf{x})$ en las transparencias de teoría). Si tuviera que usar los resultados para dárselos a un potencial cliente ¿usaría la transformación polinómica? Explicar la decisión.
5. (1.5 puntos) Volver al ejercicio 4 de la sección Regularización y Selección de Modelos de los Ejercicios de Apoyo.
   - Una medida del número efectivo de muestras "frescas" usadas en el cálculo de E_{cv} es el cociente entre la varianza de e_1 y la varianza de E_{cv}. Explicar por qué, y dibujar, respecto de N, el número efectivo de nuevos ejemplos(N_{eff}) como un porcentaje de N. (NOTA: Debería de encontrarse que N_{eff} está cercano a N.)
   - Si se incrementa la cantidad de regularización, ¿debería N_{eff} subir o bajar?. Argumentar la respuesta. Ejecutar el mismo experimento con $\lambda = 2{,}5/N$ y comparar los resultados del punto anterior para verificar la conjetura.