Tarea 9

Inteligencia Artificial

1. Entregar un documento PDF con todas tus respuestas teoricas. No se aceptan otro tipo de formato.
2. En esta tarea se evaluará todas las preguntas y no se asignará puntaje a preguntas incompletas.
3. Todo acto de COPIA implica la nota de $0A$. Evita copiar!.

---

1 Preguntas

1. Seleccione todas las afirmaciones que sean verdaderas. Explica tus respuestas.
   (a) Un tamaño de paso más grande mejora la estabilidad de SGD.
   (b) Un tamaño de paso más grande mejora la velocidad de convergencia de SGD bajo ciertas condiciones.
   (c) Es común aumentar el tamaño del paso periódicamente al realizar SGD.
   (d) Un tamaño de paso de 0 hace que el aprendizaje sea imposible.

2. En esta pregunta, recorreremos dos pasos de descenso de gradiente (no SGD) para la clasificación lineal utilizando la pérdida `hinge`. La pérdida `hinge` se define como:

$$Loss_{Hinge}(x, y, \mathbf{w}) = \max\{1 - (\mathbf{w} \cdot \phi(x))y, 0\}$$

   Realiza dos pasos del algoritmo de descenso de gradiente con el conjunto de datos de entrenamiento $D_{train} = \{(9, -1), (-1, 1)\}$, el vector de características $\phi(x) = [x]$, con un tamaño de paso $\lambda = 0.1$ y un vector de peso inicial $\mathbf{w} = [0]$. Recuerda que cada actualización de descenso de gradiente requiere una normalización de $1/|D_{train}|$.

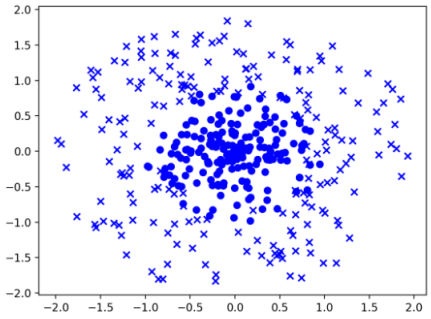
   ¿Cuál es el vector de peso actualizado después de completar estos dos pasos?.

3. Sean $x_1, x_2, \ldots, x_n$ muestras independientes desde la siguiente distribución :

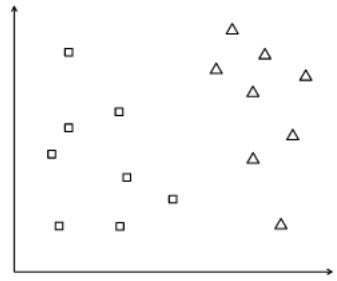
$$P(x|\theta) = \theta x^{-\theta-1}$$

   donde $\theta \geq 1, x \geq 1$. Encuentra el estimador de máxima verosimilitud de θ.

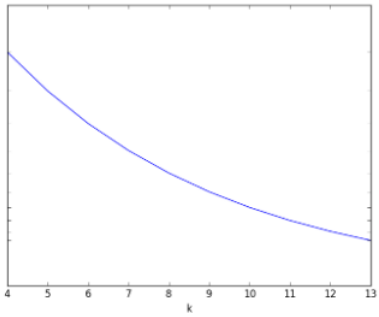
4. Considera un conjunto de datos simple de puntos $(x_i, y_i) \in R^2$, cada uno asociado con una etiqueta b_i que es -1 o $+1$. El conjunto de datos se generó muestreando puntos de datos con etiqueta -1 de un disco de radio 1.0 (mostrado como círculos rellenos en la figura) y puntos de datos con etiqueta $+1$ de un anillo con radio interno 0.8 y radio externo 2.0 (mostrado como cruces en la figura). ¿Qué conjunto de características polinomiales sería mejor para realizar una regresión lineal, asumiendo al menos tantos datos como se muestra en la figura?



5. Dadas las siguientes muestras de datos (el cuadrado y el triángulo pertenecen a dos clases diferentes), ¿qué algoritmos vistos en clase puede producir un error de entrenamiento cero?



6. Un amigo está entrenando un modelo de aprendizaje automático para predecir la gravedad de la enfermedad según k indicadores de salud diferentes. El obtiene la siguiente gráfica, donde el valor de k está en el eje x.



Indica que podría representar el eje y. Explica tu respuesta.

7. Durante el entrenamiento de tu modelo, tanto las variables independientes en la matriz $\mathbf{X} \in R^{n \times d}$ y las variables objetivo dependientes $y \in R^n$ se corrompen por el ruido. En el momento de la prueba, los puntos de datos para los que estas calculando predicciones x_{prueba}, sin ruidos. ¿Qué método(s) debes utilizar para estimar el valor de $\hat{w}$ a partir de los datos de entrenamiento para hacer las predicciones más precisas y_{prueba} a partir de los datos de entrada de prueba sin ruido x_{prueba}?. Debes asumir que se hace predicciones usando $y_{prueba} = \mathbf{X}_{prueba}\hat{w}$.

8. Considera la posibilidad de construir un árbol de decisión sobre datos con d características y n puntos de entrenamiento donde cada característica tiene un valor real y cada etiqueta toma uno de los m valores posibles. Las divisiones son bidireccionales y se eligen para maximizar la ganancia de información. Solo consideramos las divisiones que forman un límite lineal paralelo a uno de los ejes.

   (a) Demuestra o proporciona un contraejemplo: para cada valor de $m > 3$, existe alguna distribución de probabilidad de m objetos tal que su entropía sea menor que -1.

   (b) Demuestra o proporciona un contraejemplo: en cualquier camino desde la división de la raíz hasta una hoja, la misma característica nunca se dividirá dos veces.

(c) Demuestra o proporciona un contraejemplo: la ganancia de información en la raíz es menor o igual que la ganancia de información en cualquier otro nodo.