Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

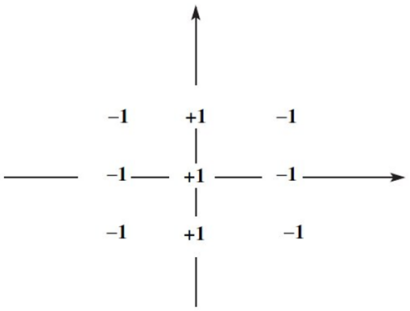
IIC2613 – Inteligencia Artificial (I/2019)

Interrogación 1

Pregunta 1

a) ¿Qué ventajas y desventajas podría tener el usar un *learning rate* alto, al minimizar una función no convexa con el algoritmo de descenso del gradiente? Entregue respuestas detalladas para cada caso. **(1,5 ptos.)**
Solución: la ventaja es que es posible evitar mínimos locales, al ejecutar pasos largos de descenso. La desventaja es que no hay garantía de que la actualización de los parámetros luego de descender genere un descenso en la pérdida, ya que la naturaleza no convexa del problema hace que la superficie a optimizar sea cambiante.

b) Considere el problema de clasificación binaria de la siguiente figura. ¿Es posible resolverlo utilizando una regresión logística? Justifique su respuesta **de manera detallada**, tanto si se puede, como si no. **(1,5 ptos.)**



Solución: sí, es posible utilizar una regresión logística para este problema. Basta con utilizar *features* de grados mayores $(x^2, x^3, x^4, \dots)$ construidas a partir de las *features* originales, que permitan que la superficie de decisión sea un polinomio que se ajuste a el set de entrenamiento.

c) Considere un problema de regresión, para el cual se utilizará un árbol de regresión. Debido a que se tiene una muestra pequeña para entrenar el árbol, se recomienda aumentar el conjunto de entrenamiento, duplicando cada uno de los ejemplos. Comente sobre la utilidad de esta estrategia. ¿Cambia la utilidad de esta estrategia si ahora nos enfrentamos a un problema de clasificación? **(1,5 ptos.)**
Solución: esta estrategia no tiene ninguna utilidad en este contexto, ya que no entrega mayor varianza a la muestra, sólo mantiene sus estadísticas con una mayor cantidad de datos.

d) Considere la siguiente expresión, que describe mediante tres términos la composición del error total en la estimación de una variable, a través de un modelo de aprendizaje:

$$Err(x) = \underbrace{\left(E\left[\hat{f}(x)\right] - f(x)\right)^2}_{\text{Bias}^2} + \underbrace{E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]}_{\text{Var}} + \underbrace{\sigma_e^2}_{\text{Error datos}} \quad (1)$$

¿Cuáles de estos términos intentan minimizar generalmente los algoritmos de aprendizaje? ¿Cómo lo hacen en cada caso? ¿Si hay términos que no se usan, cómo podrían incluirse? **(1,5 ptos.)**
Solución: en general, lás técnicas de aprendizaje buscan minimizar de manera explícita *proxies* del $Bias^2$, ya que

reducen el error en un set de datos particular, asumiendo que el tamaño del set de datos y/o su representatividad permitirá mantener la varianza controlada. Una manera de incorporar la varianza, es utilizar mecanismos rigurosos para evitar el sobreentrenamiento, como el uso de conjunto de entrenamiento aleatorios, o penalizar la complejidad del modelo. El error en los datos es algo que generalmente no se enfrenta de manera explícita en estos modelos.

Pregunta 2

Una empresa dedicada a la fabricación de maquinaria agrícola está desarrollando un sistema de regadío automático, que a diferencia de los sistemas tradicionales, le entregará a cada planta la cantidad de agua justa para maximizar su crecimiento. Para el desarrollo del sistema, la empresa ha contratado a una serie de especialistas en áreas agrícolas, de robótica y de ciencia de la computación. En base a esto, conteste las siguientes preguntas:

a) La etapa inicial de desarrollo implica la construcción de un prototipo de software que permita estimar la cantidad de agua, utilizando una base de datos que contiene K atributos de las plantas y el entorno, medidos por los especialistas agrícolas, además de una recomendación de la cantidad de agua que necesita la planta. Los atributos fueron generados por los especialistas a partir de cámaras y sensores de distinto tipo. Describa en detalle y analíticamente el problema de aprendizaje asociado a esta situación, y cómo lo resolvería. **(2.0 ptos)**
Solución: Dado que se conocen los K atributos y el valor de la función a predecir (la cantidad de agua), el problema puede ser planteado como una regresión lineal con función de pérdida dada por el error cuadrático medio de las estimaciones:

$$\underset{w}{\text{argmin}} \quad \frac{1}{2}\sum_{i=1}^{N}\left(\langle w, x_i\rangle + w_0 - y_i\right)^2, \tag{2}$$

donde $w \in \mathcal{R}^K$ y $w_0 \in \mathcal{R}$ son los parámetros y *bias* de la regresión, respectivamente, $x_i \in \mathcal{R}^K \ \forall i \in [1, N]$ son vectores que contienen para cada planta i los K atributos medidos, e $y_i \in \mathcal{R} \ \forall i \in [1, N]$ es la cantidad de agua estimada para la planta i. Dado que es un problema cuadrático, puede resolverse eficientemente utilizando un esquema de descenso del gradiente, asegurando la optimalidad de la solución.

b) Debido a un lamentable "hecho fortuito", en el que aparentemente estaría involucrada la competencia, los especialistas agrícolas desaparecieron, por lo que no es posible generar nuevas mediciones para los atributos. Utilizando los datos previamente recolectados, describa en detalle y analíticamente un problema de aprendizaje que permita estimar los K atributos, indicando que datos utilizaría y como lo resolvería. **(2.0 ptos)**
Solución: Dado que lo "único" que falta son los especialistas, es posible plantear el problema como K problemas independientes, donde cada uno de ellos estima el valor del k-ésimo atributo a través de una regresión lineal, utilizando como entrada los datos utilizados por los especialistas para estimar originalmente los valores de los atributos:

$$\underset{\{v^k\}_1^K}{\text{argmin}} \quad \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\left\langle v^k, z_i^k\right\rangle + v_0^k - x_i^k\right)^2, \tag{3}$$

donde $v^k \in \mathcal{R}^{M_k} \ \forall k \in [1, K]$ y $v_0^k \in \mathcal{R} \ \forall k \in [1, K]$ son los parámetros y *bias* de cada regresión, respectivamente, $z_i^k \in \mathcal{R}^{M_k} \ \forall i \in [1, N] \wedge \forall k \in [1, K]$ son vectores que contienen para cada atributo k las M_k *features* usadas para estimarlos, y $x_i^k \in \mathcal{R} \ \forall i \in [1, N] \wedge \forall k \in [1, K]$ es el valor del atributo k para la planta i. Dado que el problema corresponde a la suma de K problemas cuadráticos independientes, cada uno puede resolverse eficientemente utilizando un esquema de descenso del gradiente, asegurando la optimalidad de la solución.

c) Debido a un nuevo "hecho fortuito", que en esta oportunidad involucra fuego, gran parte de las bases de datos de la empresa han quedado inutilizadas, sólo pudiendo rescatar la cantidad de agua que requiere cada planta evaluada, así como las imágenes y mediciones utilizadas para calcular sus atributos. Sabiendo que la estimación del agua que necesita cada planta se debe calcular utilizando K atributos numéricos, describa analíticamente un problema de aprendizaje **conjunto**, que permita obtener modelos para estimar los atributos y la cantidad de agua necesaria para cada planta. Indique como resolver este problema y si este es lineal, cuadrático, convexo o no convexo. **(2.0 ptos)**
Solución: Dado que lo único que se conoce para estimar la cantidad de agua para cada planta son las *features* utilizadas por los expertos, y que es necesario calcular los K atributos para realizar la estimación, es necesario plantear el problema como una regresión no lineal, que involucre productos entre los parámetros de los regresores para estimador los atributos y los del regresor para estimar la cantidad de agua.

$$\underset{w,V}{\text{argmin}} \quad \frac{1}{2}\sum_{i=1}^{N}\left(\langle w, V\hat{z}_i + V_0\rangle + w_0 - y_i\right)^2, \tag{4}$$

donde $w \in \mathcal{R}^K$ y $w_0 \in \mathcal{R}$ son los pesos y *bias* de la regresión, respectivamente, $V \in \mathcal{R}^{K \times M}$ es una matriz que contiene los parámetros de K regresores (uno por fila), que estiman cada uno de los K atributos, $V_0 \in \mathcal{R}^K$ es un vector con los bias de cada uno de los K regresores en V, $\hat{z}_i \in \mathcal{R}^M \ \forall i \in [1, N]$ son vectores que contienen para cada planta i las M *features* medidas, e $y_i \in \mathcal{R} \ \forall i \in [1, N]$ es la cantidad de agua estimada para la planta i. Es importante notar que la dimensionalidad de $\hat{z}_i$ es la misma para todo i, lo que implica que cada atributo puede eventualmente ser estimado utilizando todas las *features* disponibles. Finalmente, dado que el problema no es convexo, aún es posible utilizar descenso del gradiente para resolverlo, pero sin asegurar la optimalidad de la solución.

Pregunta 3

Al igual que los árboles de decisión, un *random forest* busca que los tests realizados en cada nodo, separen de la mejor manera posible a las categorías a predecir. A pesar de que en general se utilizan umbrales para los tests con variables numéricas, nada impide que los tests se realicen de otra manera. Asumiendo que se tiene un set de datos de clasificación, donde todos los atributos son numéricos, conteste las siguiente preguntas:

a) Describa una estrategia para variar la complejidad de la clasificación en cada nodo, y que permita disminuir la correlación de los distintos árboles de un *random forest*. **(2.0 ptos.)**
Solución: Dado que no hay restricción con respecto a la técnica de clasificación a utilizar en cada nodo, se puede asumir, sin pérdida de generalidad, que se utilizarán regresiones logísticas. Luego, dado que lo fundamental es reducir la correlación entre los árboles, para cada nodo se muestreará aleatoriamente cuántos y cuáles atributos se utilizarán en la regresión. Esta estrategia es extendible de manera sencilla para más de dos clases, utilizando múltiples regresiones en un esquema *one-vs-all*.

b) Describa detalladamente como entrenar este nuevo clasificador, ya sea indicando el proceso o el problema de optimización asociado. **(2.0 ptos.)**
Solución: La construcción de cada árbol sigue el mismo esquema que en el caso de los *random forest* tradicionales, con la salvedad de que dado que la regresión logística estima la probabilidad de pertenecer a una clase, se utilizará el valor 0.5 como umbral, para construir las dos ramas del árbol por cada nodo. La ganancia de información sigue siendo válida para la construcción del árbol.

c) ¿Cuál es el espacio de hipótesis de este nuevo clasificador? **(1.0 ptos.)**
Solución: Lo único que cambia con respecto al espacio de hipótesis de un *random forest* (orden de selección de atributos y umbrales por cada árbol), es que ahora en vez de umbrales se tendrán los parámetros de cada una de las regresiones.

d) ¿Como controlaría el *overfitting* de cada uno de los árboles? **(1.0 ptos.)**
Solución: Dado que la complejidad de una regresión puede relacionarse con la cantidad de parámetros que utilice, es posible controlar el overfitting de cada árbol, penalizando la cantidad de parámetros que utiliza la regresión, al mismo tiempo que se premia el rendimiento. En otras palabras, al calcular la ganancia de información para cada uno de los conjuntos de atributos evaluados, se debe penalizar esta por una función de la cantidad de atributos utilizados en la regresión.