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

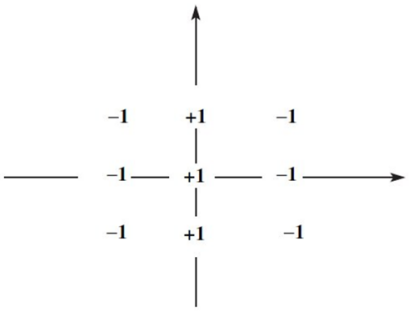
IIC2613 – Inteligencia Artificial (II/2023)

Ejercicios Control

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:

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

¿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.

Pregunta 4

a) Un fabricante de hardware se encuentra en el proceso de diseño de una nueva generación de CPUs para dispositivos móviles, cuya principal característica es un mecanismo adaptativo de activación/desactivación de los núcleos (cores), en base a al comportamiento de los usuarios. Con el fin de ponerse a tono con el mercado mundial, el equipo de desarrollo planea utilizar técnicas de aprendizaje de máquina para construir el sistema de activación de núcleos.

   i) Indique como podría enfrentar el equipo desarrollador este problema. Especifique claramente que tipo de aprendizaje utilizaría, y qué ventaja tendría esto por sobre un sistema basado en reglas. **(2 ptos.)**
   Solución: El equipo podría enfrentar el problema usando aprendizaje supervisado, recolectando datos de uso de usuarios, por ejemplo carga de CPU y temperatura de esta al ejecutar distintos programas, y la cantidad de núcleos de la CPU. De esta manera, luego de filtrar y preprocesar los datos de entrada, se podría construir un modelo para predecir la cantidad de núcleos óptima dada la carga y temperatura. La ventaja de esto es que no es necesario poner reglas explícitas de comportamiento, sólo indicar cuál sería el comportamiento deseado para los ejemplos.

   ii) Con el fin de entrenar los modelos, el equipo de desarrollo recolectó un gran volumen de datos de uso de las nuevas CPUs, desde los computadores de sus propios integrantes. Comente sobre las ventajas y desventajas de esta decisión. **(2 ptos.)**
   Solución: Al recolectar un gran volumen de datos, es posible capturar de mejor manera las particularidades del problema. Por otro lado, el hecho de que sólo se utilizaran datos del equipo de desarrollo, implica que los datos están fuertemente sesgados, lo que seguramente generará sobreentrenamiento.

b) Con el fin de categorizar a los alumnos de la Escuela de Ingeniería en uno de 10 posibles perfiles profesionales, un investigador decide utilizar una regresión lineal con polinomio de grado 1, sobre un conjunto de datos con distintas mediciones hechas a los estudiantes. Los registros de este conjunto se ubican en un espacio de características de 37 dimensiones, cada una con dominio en $\mathbb{R}$. Para calibrar los parámetros del modelo de regresión, el investigador decide utilizar un método distinto al del descenso del gradiente, debido a la posibilidad de caer en mínimos locales. Comente sobre las decisiones que tomó el investigador para llevar a cabo el estudio. **(2 ptos.)**
   Solución:

   i. Dado que el problema es de clasificación y no de regresión, el uso de un modelo de regresión lineal es incorrecto. Lo correcto sería usar una regresión logística, o algún algoritmo de clasificación.

   ii. El problema de optimización asociado a una regresión lineal es convexo, por lo que sólo tiene un mínimo local, que coincide con el global. Dado que el descenso del gradiente converge a un mínimo local (o punto crítico), la suposición del investigador es incorrecta (el descenso del gradiente no tiene problemas en una regresión lineal).

Pregunta 5

a) ¿Como se podría realizar clasificación con un árbol de decisión, si falta el valor de alguna de las dimensiones del vector de entrada? **(1 pto.)**
Solución: existen al menos dos opciones: i) rellenar el valor faltante en base a los existente en el set de entrenamiento y ii) calculando la moda de las predicciones, si se toman todos los caminos del test que no se puede realizar.

b) Explique por qué la cantidad de parámetros activos (distintos de cero) en una red neuronal, puede dar una noción de la complejidad del modelo. **(1 pto.)**
Solución: los parámetros iguales a cero implican menos parámetros en uso (menos conexiones entre neuronas), luego, un modelo con más parámetros iguales a cero, es menos complejo.

c) Indique como utilizaría validación cruzada para estimar la mejor profundidad de un árbol de decisión. **(1 pto.)**
Solución: para cada ronda de validación, se limita la profundidad máxima del árbol. Luego, después de K rondas, se puede elegir la mejor profundidad (hasta K) en base al rendimiento promedio.

d) ¿De qué manera el método del momentum disminuye el riesgo de caer en mínimos locales en redes neuronales con capas ocultas? **(1 pto.)**
Solución: al combinar linealmente el valor del nuevo gradiente con la dirección de descenso anterior (una combinación de gradientes previos), los puntos con derivada cero pueden ser evitados, ya que no disminuye la *velocidad* del descenso.

e) ¿En qué situaciones es preferible utilizar el radio de ganancia por sobre la ganancia de información? **(1 pto.)**
Solución: cuando la cantidad de valores de la variable a testear es muy grande (ganancia de información tiende a elegir siempre a estas variables).

f) ¿Como podría utilizarse una red neuronal para aumentar la resolución de imágenes (manteniendo una buena calidad)? **Hint:** enfóquese en el entrenamiento de una red para esta tarea. **(1 pto.)**
Solución: utilizando un set de entrenamiento donde se tiene cada imagen en dos versiones (pequeña y grande) y una red que toma la versión pequeña y genera como salida una imagen de tamaño grande. La función de pérdida sería la suma de las diferencias cuadrática entre los valores de cada pixel predicho y el real (versión grande de la imagen).

Pregunta 6

a) Considere un problema de clasificación sobre variables categóricas. Extienda el algoritmo de construcción de los árboles de decisión basado en la ganancia de información, para que se puedan realizar tests sobre dos variables de manera simultánea **(2 ptos.)**.
Solución: Basta con modificar la métrica de ganancia de información, tomando ahora las proporciones de ejemplos (probabilidades) en el producto cartesiano de las variables (atributos) elegidas.

b) En general, al momento de decidir el valor a testear en un nodo de un árbol de decisión, se toma la ganancia de información como métrica. Una de las desventajas de esta, es que no considera la nueva estructura del árbol en el cálculo (la resultante de seleccionar ese test, con distinta profundidad y número de nodos), lo que puede derivar en problemas de sobreentrenamiento. Extienda la métrica de la ganancia de información, agregando un nuevo término aditivo, de manera que ahora, para tomar la decisión, se considere información sobre la posible nueva estructura del árbol. **Hint:** considere la decisión en un nodo como la minimización del riesgo estructural empírico. **(2 ptos.)**
Solución: La idea es agregar un término que penalice tanto la nueva profundidad del árbol y/o la nueva cantidad de nodos. Es importante notar además que un incremento en la profundidad es más importante desde el punto de vista del sobreentrenamiento, cuando se hace a mayor profundidad, *i.e.*, a mayor profundidad, mayor riesgo de sobreentrenamiento. Una posible solución podría ser la siguiente:

$$F(S, A) = Gain(S, A) - \alpha \cdot 2^d \cdot |S_A|$$

donde α es una constante predefinida, d es la nueva profundidad del árbol y $|S_A|$ es la cantidad de valores que toma el atributo A (cantidad de nodos que se agregarán).

c) Considere una competencia, donde se debe resolver un problema de regresión en base a variables categóricas. Dado que en este problema el riesgo de sobreentrenamiento es alto, sólo se permite utilizar árboles de regresión sobre una (1) de las variables disponibles, con el fin de limitar la profundidad del árbol. Utilizando múltiples árboles de **manera secuencial** (cada árbol puede utilizar la variable que quiera), indique como es posible construir un sistema de regresión que estime de mejor manera la función buscada. **(2 ptos.)**
Solución: Una posible solución es estimar de manera secuencial el valor de la regresión, corrigiendo las estimación en base al residuo de esta última, *i.e.*, si $R(x)$ es el valor a estimar para un ejemplo x, la regresión se realizaría usando $K + 1$ árboles de regresión F_k de profundidad 1: $R(x) = F_0(x) + F_1(x) + \cdots + F_K(x)$. En esta modalidad, el árbol F_0 realizaría la regresión tradicional sobre x, mientras que F_1 estimaría el valor $R(x) - F_0(x)$. De manera análoga, el árbol F_k estimaría $R(x) - \sum_i = 0^k F_i(x)$.

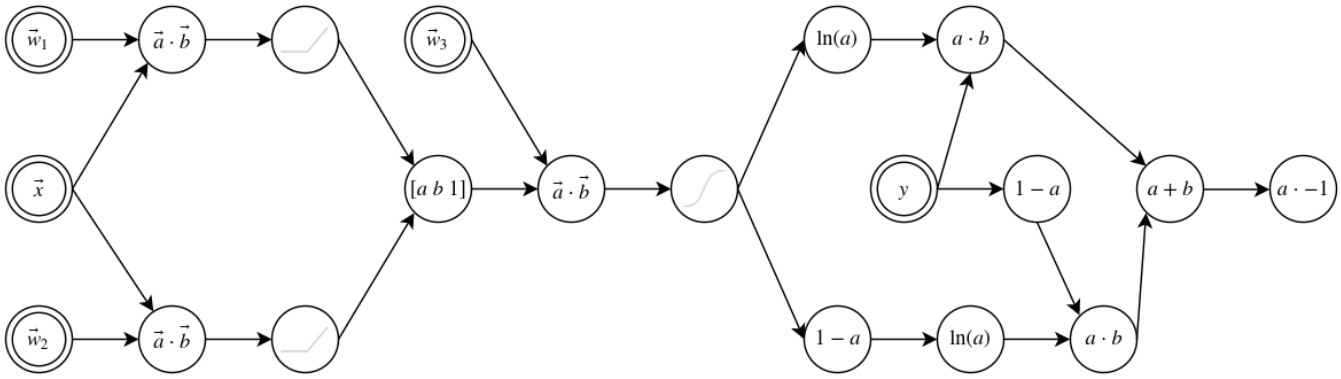
Pregunta 7

Dado un conjunto de ejemplos $x = \bigcup_i^N x_i$, con $x_i \in \Re^L$ y sus etiquetas asociadas $y = \bigcup_i^N y_i$, con $y_i \in [0, 1]$, considere la función de pérdida *cross-entropy* definida a continuación:

$$E(x, y; w) = -\sum_i y_i \ln o^i + (1 - y_i) \ln(1 - o^i)$$

donde o^i es la salida de la red (perceptrón sigmoidal) para el ejemplo i y w es el vector de parámetros de la red. En base a esto, conteste las siguientes preguntas:

a) Construya el grafo de cómputo para una red con una capa oculta de 2 neuronas con función de activación ReLU, que realice la clasificación de x, utilizando como pérdida la función $E(x, y; w)$. No combine múltiples operaciones en un sólo nodo del grafo. **(3 ptos.)**
Solución:



b) Extienda la definición de $E(x, y; w)$ para el escenario de clasificación multiclase (más de dos categorías), *i.e.*, $y_i \in [0, K]$. **(3 ptos.)**
Solución: Una posible solución es asumir que la salida de la red tiene K dimensiones (una por categoría), y que cada una de estas tiene como no linealidad una sigmoide. Luego, la siguiente función tiene como objetivo penalizar: i) valores distintos de 0 para cualquier dimensión que no sea la correspondiente a la clase correcta, y ii) cualquier valor distinto de 1 para la clase correcta:

$$\hat{E}(x, y; w) = -\sum_i \left(ln(o^i_{y_i}) + \sum_{k \neq y_i}^{K} ln(1 - o^i_k) \right)$$

Pregunta 8

Un *Double Soft-margin SVM* es un algoritmo de clasificación similar a un *Soft-margin SVM*, pero que impone dos restricciones sobre el margen funcional: una sobre su valor mínimo (*Soft-margin SVM*) y una sobre el máximo. Tomando esto en consideración, responda las siguientes preguntas:

a) Defina el problema de aprendizaje para un *Double Soft-margin SVM*, como un problema de minimización cuadrático con restricciones lineales. **(3 ptos.)**
Solución: Por cada ejemplo, se agregan dos nuevas restricciones, que imponen un límite superior constante al margen (τ), que si es violado se penaliza en una cantidad μ_i.

$$\begin{aligned} \underset{w,b,\{\xi_i\},\{\mu_i\}}{\text{argmin}} \quad & \frac{\lambda}{2}\|w\|^2 + \sum_i^N \xi_i + \sum_i^N \mu_i \\ \text{s.a} \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in (1, N) \\ & \xi_i \geq 0, \forall i \in (1, N) \\ & y_i(w^\top x_i + b) \leq \tau + \mu_i, \forall i \in (1, N) \\ & \mu_i \geq 0, \forall i \in (1, N) \end{aligned}$$

b) Modifique el problema de aprendizaje definido en el ítem anterior, planteándolo esta vez como un problema de minimización cuadrático sin restricciones. **(3 ptos.)**
Solución:

$$\underset{w,b}{\text{argmin}} \quad \frac{\lambda}{2}\|w\|^2 + \sum_i^N \max\{0, 1 - y_i(w^\top x_i + b)\} + \max\{0, y_i(w^\top x_i + b) - \tau\}$$