Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Segundo Semestre 2024
Fundamentos Matemáticos para Ciencia de Datos IMT3120

Tarea 2

Instrucciones. Puede usar las conclusiones de una pregunta para responder las preguntas que siguen.

En clase se criticó el modelo de predicción binaria en el caso en que existe ruido en las etiquetas. En esta tarea exploraremos un mejor modelo para este caso, generalizándolo además al caso de etiquetas sobre un conjunto finito.

Pregunta 1. Sea $\mathcal{X} = \mathbb{R}^d$ e $\mathcal{Y} = [K] := \{1, 2, \ldots, K\}$. Consideramos una distribución $\mathcal{D}$ soportada en $\mathcal{X} \times \mathcal{Y}$ tal que

$$\mathbb{P}[Y = k | X = x] = \eta_k(x), \tag{1}$$

donde $\eta : \mathbb{R}^d \to {\Delta_K}^*$ es una función desconocida. Para el problema de aprender η, consideramos $\mathcal{Y}' = \Delta_K$ y una clase $\mathcal{F} \subseteq (\mathcal{Y}')^{\mathcal{X}}$.

Sea $S = \{(X_i, Y_i)\}_{i \in [n]} \overset{iid}{\sim} \mathcal{D}$ una muestra aleatoria.

(a) Se propone una heurística para definir $\mathcal{F}$. Vamos a proponer funciones $\pi : \mathbb{R}^d \mapsto \Delta_K$ que satisfacen

$$\sum_{i=1}^{n} \pi_k(X_i) X_{ij} = \sum_{i=1}^{n} \mathbb{1}(k = Y_i) X_{ij} \qquad (\forall k \in [K])(\forall j \in [d]). \tag{2}$$

De una justificación probabilística para la heurística (2).
Indicación. Use la regla de Bayes.

(b) Otra heurística que utilizaremos es la siguiente: dentro de todas las posibles funciones $\pi : \mathbb{R}^d \mapsto \Delta_K$, escogemos aquella que maximice la entropía:

$$\pi^* = \arg\max \Big\{ \int -\pi(x) \ln(\pi(x)) \, d\mathbb{P}_X(x) : \ \forall x \in \mathcal{X}, \pi(x) \geq 0, \ \sum_{k=1}^{K} \pi_k(x) = 1, \ \pi \text{ satisface (2)} \Big\}.$$

Dado que las integrales de arriba no se pueden calcular, las estimamos a través de los datos. Esto lleva al problema

$$\begin{aligned} \max_{\pi} \quad & -\sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k(X_i) \ln(\pi_k(X_i)) \\ \text{s.a.} \quad & \pi_k(X_i) \geq 0 \quad \forall i \in [n], \forall k \in [K], \\ & \textstyle\sum_{k=1}^{K} \pi_k(X_i) = 1 \quad \forall i \in [n], \\ & \pi \text{ satisface (2)}. \end{aligned}$$

Notar que para este problema de optimización sólo importa cómo está definida π sobre la muestra, S. Pruebe que la solución óptima de este problema satisface que existen vectores $w_1^*, \ldots, w_K^* \in \mathbb{R}^d$ tales que

$$\pi_k^*(X_i) = \frac{\exp\{\langle w_k^*, X_i \rangle\}}{\sum_{h=1}^{K} \exp\{\langle w_h^*, X_i \rangle\}}.$$

*Se define $\Delta_K = \{\lambda \in \mathbb{R}_+^K : \sum_{k \in [K]} \lambda_k = 1\}$. Este conjunto representa a todas las distribuciones de probabilidad sobre $[K]$.

La última identidad de arriba la extrapolaremos heuristicamente a todo el espacio (no solamente para los datos observados). Consideramos entonces la familia parametrica

$$\mathcal{F} = \Big\{ \pi(W,\cdot) : \mathbb{R}^d \mapsto \Delta_K \text{ donde } \pi_k(x) = \frac{\exp\{\langle w_k, x\rangle\}}{\sum_{h=1}^K \exp\{\langle w_h, x\rangle\}} : W \in \mathbb{R}^{K\times d} \Big\}.$$

Pregunta 2. (a) Pruebe que la log-verosimilitud de los datos dada una hipótesis $\pi \in \mathcal{F}$ es

$$\mathcal{L}(S|\pi) \propto -\mathcal{R}_S(W) := \frac{1}{n}\sum_{i\in[n]} \Big[\langle w_{Y_i}, X_i\rangle - \ln\Big(\sum_{k=1}^K \exp\{\langle w_k, X_i\rangle\}\Big)\Big],$$

y que el problema de maximización de la verosimilitud

$$(P) \min_{W\in\Omega} \mathcal{R}_S(W)$$

es un problema convexo, si Ω es convexo y cerrado. Este modelo se conoce como *regresión logística*.
Indicación. Puede usar (no necesita probar) el hecho de que la composición de una función lineal $T : \mathbb{R}^{K\times d} \mapsto \mathbb{R}^K$ con una función convexa $g : \mathbb{R}^K \mapsto \mathbb{R}\cup\{+\infty\}$ es tal que la composición $f(W) = g(TW)$ es convexa.

(b) Pruebe que la función objetivo del problema (P) es Lipschitz, y determine su constante para la norma ℓ_2.

Opcional +0,3pts. Pruebe que $\mathcal{R}_S$ es suave (y determine la constante correspondiente C), en el sentido de que su Hessiano es uniformemente acotado

$$\langle \nabla^2 \mathcal{R}_S(W)H, H\rangle \le C\|H\|^2 \quad (\forall W, H \in \mathbb{R}^{K\times D}),$$

donde $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$ es un producto interno (no necesita probarlo). Puede escoger la norma para la cual establece esta suavidad.

Pregunta 3. Consideramos $(\mathbb{R}^m, \|\cdot\|_2)$ y $\mathcal{B}(x,R)$, la bola centrada en x de radio $R \ge 0$.

Sea $\Omega \subseteq \mathbb{R}^m$. Decimos que un conjunto finito $\mathcal{C} \subseteq \Omega$ es un ϵ-*cubrimiento* de Ω si para todo $x \in \Omega$ existe $c \in \mathcal{C}$ tal que $\|x - c\|_2 \le \epsilon$. Por otra parte, decimos que un conjunto $\mathcal{P} \subseteq \Omega$ es un ϵ-*empaquetamiento* de Ω si para todo $p, q \in \mathcal{P}$, $\mathcal{B}_2(p,\epsilon) \cap \mathcal{B}_2(q,\epsilon) = \emptyset$. Denotamos como $N(\epsilon)$ al número de ϵ-cubrimiento de Ω; es decir, la mínima cardinalidad de un ϵ-cubrimiento. Similarmente, denotamos $P(\epsilon)$ al número de ϵ-empaquetamiento, que es la máxima cardinalidad de un ϵ-empaquetamiento de Ω.

(a) Pruebe que

$$N(2\epsilon) \le P(\epsilon) \le N(\epsilon).$$

(b) Concluya que si $\Omega = \mathcal{B}(0, R)$

$$\left(\frac{R}{\epsilon}\right)^m \leq N(\epsilon) \leq \left(1 + \frac{2R}{\epsilon}\right)^m.$$

(c) Sea ahora $\Omega \subseteq \mathbb{R}^{K\times d}$ y $\mathcal{X} \subseteq \mathbb{R}^d$, ambos compactos. Sea además

$$\mathcal{G} := \Big\{\phi(W, \cdot) : \mathcal{X} \mapsto \Delta_K \,\Big|\, \phi(W, (X, Y)) = -\langle w_Y, X\rangle + \ln\Big(\sum_{k=1}^{K} \exp\{\langle w_k, X\rangle\}\Big), \quad W \in \Omega\Big\}.$$

Pruebe que si $\mathcal{C}$ es un ϵ-cubrimiento de Ω y

$$\mathcal{N} = \{\phi(W, \cdot) : \mathcal{X} \mapsto \Delta_K : \ W \in \mathcal{C}\},$$

entonces $N_{LB}(2\epsilon L, \mathcal{G}, L_1(\mathcal{D}) \leq |\mathcal{N}|$, donde $L = \sup_{x\in\mathcal{X}} \|x\|_2$. Concluya que el problema de regresión logística con conjunto de parámetros $\Omega \subseteq \mathbb{R}^{K\times d}$ compacto y atributos $\mathcal{X} \subseteq \mathbb{R}^d$ en un compacto es aprendible, y determine una cota sobre su complejidad muestral, en función de $L, R, \epsilon, \delta, K, d$.

Parte Computacional.

En esta parte probaremos el modelo de regresión logística en un conjunto de datos reales:

MNIST. `https://www.tensorflow.org/datasets/catalog/mnist`. Este conjunto de datos contiene 70,000 imágenes (60,000 de entrenamiento y 10,000 de validación) de 28x28 pixeles de números del 0 al 9 escritos a mano por diferentes personas. El objetivo es construir un clasificador que dada la matriz de pixeles prediga el número escrito.

Para resolver el problema de regresión logística, consideramos un caso irrestricto, $\Omega = \mathbb{R}^{10\times(28^2+1)}$; es decir, para cada categoría parametrizamos un vector que representa una transformación afín, incluyendo el parámetro afín (notar que esta parametrización requiere que los datos se extiendan a un atributo adicional, que se le asigna el valor 1). Para resolver el problema de optimización se sugiere usar un método de gradiente con paso adaptativo

Algoritmo: Descenso de Gradiente con Backtracking
Input: Modelo inicial $W^0 = 0$, Tiempo T
Output: Modelo final W

Inicializar $W \leftarrow W^0$
$\eta \leftarrow 1$
Repetir
  Calcular gradiente $\nabla f(W)$
  while $f(W - \eta\nabla f(W)) > f(W) - \frac{\eta}{2}\|\nabla f(W)\|_2^2$ do
    $\eta \leftarrow \eta/2$
  end while
  Actualizar modelo $W \leftarrow W - \eta\nabla f(W)$
Hasta que hayan transcurrido T segundos
Entregar W

(a) Es posible que su computador no logre calcular una sóla iteración del algoritmo (trabajando con los 60,000 datos de entrenamiento a la vez). Para probar su código, intente primero clasificar con 100, 1,000, 10,000 y 60,000 datos (escogidos al azar). Pruebe su código y reporte cuando tome más de 5 minutos en terminar (en este caso, diremos que el algoritmo falló).

(b) Queremos evaluar ahora el error de validación del algoritmo. Una forma de hacerlo es, para cada dato de validación x, predecir con la categoría k que maximiza la probabilidad en π:

$$\hat{k}(x) = \arg\max_{k\in[K]} \pi_k(x).$$

Para los modelos obtenidos con los subconjuntos de datos de la parte (a), evalúe su error de clasificación en los 10,000 datos destinados para este propósito. Mejora el error de clasificación usando más datos? Muestre 2-3 ejemplos mal clasificados y discuta qué ocurrió: son los datos difíciles de clasificar a la vista? Cuál es el grado de confianza (valor de $\pi_{\hat{k}(x)}(x)$) de la clasificación para la regresión logística en estos ejemplos?

(c) Ahora exploramos algunas heurísticas para escalar el algoritmo. La primera es utilizar cada dato de manera secuencial. Es decir, en cada paso del método del gradiente sólo calcule el gradiente de la función de prérdida con respecto a un dato (no utilizado anteriormente), y utilice este gradiente para la actualización del modelo.

```
Algoritmo: Metodo de Gradiente Estocastico (1 pasada)
    Input: Modelo inicial W^0 = 0, Tiempo T, paso η > 0
    Output: Modelo promediado Ave

    1:   Inicializar W ← W^0
    2:   Promedio Ave ← W^0
    3:   Iteracion t ← 1
    4:   Repetir
    5:       Calcular gradiente del dato t-esimo: ∇f_t(W)
    6:       Calcular nuevo modelo W' ← W − η∇f_t(W)
    7:       Actualizar promedio Ave ← ((t−1)/t)Ave + (1/t)W'
    8:       Actualizar modelo W ← W'
    9:   Hasta que hayan transcurrido T segundos,
          o cuando los n datos hayan sido usados
   10:  Entregar Ave
```

Escoja un paso fijo $\eta = 10^{-l}$ con $l = 1, 2, \ldots, 4$ y corra su algoritmo hasta que pasen 5 minutos. Reporte cuantos datos logra utilizar el algoritmo. Escoja el modelo con el paso que mejor predice sobre el conjunto de validación, y reporte su error de clasificación sobre este conjunto. Grafique algunos de los ejemplos donde el algoritmo clasifica erroneamente, y discuta por qué podría estar ocurriendo este problema.

(d) Finalmente, usamos la misma idea del algoritmo, anterior, pero reemplazando la línea 5 por

```
5': Escoger un subconjunto aleatorio de M datos
     y calcular ∇f_M(W).
```

Notar que a diferencia del algoritmo anterior, este método podría reutilizar los datos de la muestra en distintas iteraciones. En vista de esto, llamamos a este algoritmo `Metodo de Gradiente Estocastico de Multiples Pasadas`. Este algoritmo posee otro hiperparámetro: $M \in [60,000]$. Pruebe todas las combinaciones entre pasos $\eta = 10^{-l}$ con $l = 1, 2, 3, 4$ y $M = 10^p$ con $p = 1, 2, 3, 4$. Realice el mismo mecanismo para elegir su modelo usando

el conjunto de validación. Grafique en una tabla de 4×4 el error sobre el conjunto de validación de todos sus modelos.

- Es el algoritmo sensible a estos parámetros?
- Existe una diferencia significativa entre el error de entrenamiento y el de validación?

Compare el error de validación de este algoritmo con el obtenido en la parte (c).