Departamento de Ingeniería Matemática
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Curso: Aprendizaje de Máquinas, MA5204
Profesor: Felipe Tobar
Profesores Auxiliares: V. Faraggi, F. Fêtis, B. Moreno, F. Vásquez, A. Wortsman
Fecha de publicación: 24/03/21

Tarea Teórica #1

Dedicación recomendada: 10 horas

Fecha de entrega: 4 de abril

Instrucciones: La tarea es individual. El formato de entrega es libre (hoja escaneada, Latex, hoja digital, etc), solo se pide que sea legible. Sea riguroso con sus demostraciones. Dudas y consultas por el foro de Ucursos.

P1. Recuerdo de Probabilidades [6 pts. (35 %)]

Se quiere estudiar el comportamiento de un vector bidimensional que tiene sus dos componentes ortogonales, independientes y que siguen una distribución normal. Al realizar las mediciones respectivas de cada componente, se obtiene una Muestra Aleatoria Simple (MAS, donde cada dato es generado desde una misma distribución y son independientes entre sí (i.i.d)) $U = (U_1, ..., U_n)$ de n observaciones con $U_n \sim \mathcal{N}(0, \sigma^2)$ y una MAS $W = (W_1, ..., W_n)$ de n observaciones con $W_n \sim \mathcal{N}(0, \sigma^2)$. En específico, se busca estudiar el comportamiento de los módulos de los vectores obtenidos. Se obtiene una nueva MAS $X = (X_1, ..., X_n)$ dada por:

$$X_i = \sqrt{U_i^2 + W_i^2}.$$

a) **[2 pt.]** Encuentre la función de densidad de X_i, donde $i = 1, ..., n$.
Indicación: Plantee la función de distribución acumulada $F_{X_i}(x_i; \sigma)$, utilice cambio de variables y luego aplique TFC.

b) **[2 pt.]** Considere ahora la MAS dada por $X = (X_1, ..., X_n)$. Calcule el Estimador de Máxima Verosimilitud ($\hat{\sigma}^2_{EMV}$) de σ^2.

c) **[2 pt.]** Demuestre que $\hat{\sigma}^2_{EMV}$ es insesgado para σ^2 y calcule la varianza de $\hat{\sigma}^2_{EMV}$.
Indicación: Le puede ser útil encontrar la distribución de X_i^2, asumir la esperanza y la varianza conocida, y recordar que:
$$\mathbb{V}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2.$$

Una propiedad importante de los Estimadores de Máxima Verosimilitud tiene que ver con su comportamiento asintótico, el cual nos dice que cuando la cantidad de observaciones tiende a infinito, el EMV tiene un comportamiento normal. Formalmente, consideramos una colección de n observaciones $Y_i \sim \mathcal{P}_\theta$ con $i = 1, ..., n$, donde $\mathcal{P}_\theta$ es una distribución con parámetro a estimar θ (parámetro real). Se tiene que la secuencia de EMV's ($\hat{\theta}^{(n)}_{EMV}$) cumplen con lo siguiente:

$$\sqrt{n}(\hat{\theta}^{(n)}_{EMV} - \theta) \to \mathcal{N}(0, I(\theta)^{-1}),$$

donde $I(\theta)$ denota la Información de Fisher (IF) y la convergencia es en distribución. La expresión de la IF, considera a la v.a. Y que genera los datos con parámetro real θ y viene dada por:

$$I(\theta) = \mathbb{E}_\theta \left(\left(\frac{\partial \log p_\theta(Y)}{\partial \theta} \right)^2 \right) = -\mathbb{E}_\theta \left(\frac{\partial^2 \log p_\theta(Y)}{\partial \theta^2} \right) \quad (1)$$

(Bonus) Calcule la distribución asintótica de $\hat{\sigma}^2_{EMV}$.

P2. Regresión Lineal [6 pts. (45 %)]

Para este problema, además de las partes a) y b), usted puede elegir entre realizar la parte c) o la parte d), ambas tienen el mismo puntaje. Si lo desea, también puede realizar ambas partes. En caso de realizar ambas, la mitad del puntaje de la parte con menor puntaje, entre la c) y la d), se agregará como bonus a este problema. [1]

a) **[1 pt.]** Suponga que $y = a_0 + a_1 x_1 + a_2 x_2 + \epsilon$, con $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

Escriba una expresión para $\mathbb{P}(y|x_1, x_2)$. Teniendo en cuenta esta expresión, sea $\mathcal{D} = \{x_1^{(i)}, x_2^{(i)}, y^{(i)}\}_{i=1}^n$ el conjunto de datos de entrenamiento. Desarrolle la log-verosimilitud de este conjunto.

b) **[2 pts.]** Demuestre que encontrar el Estimador de Máxima Verosimilitud de la log verosimilitud anterior es equivalente a minimizar el error cuadrático.

c) **[3 pts.]** Ahora considere el modelo visto en clases:

$$Y = \tilde{X}\theta + \epsilon,$$

donde $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ tal que $\mathbf{0} \in R^{(M+1)\times 1} \wedge \mathbf{\Sigma} \in R^{(M+1)\times(M+1)}$ con $\Sigma_{i,j} = \sigma^2$ para $i = j$ y $\Sigma_{i,j} = 0$ para $i \neq j$ con $i, j \in [1, ..., M+1]$.

Vuelva a desarrollar los pasos anteriores para este caso.

Observación: Es altamente recomendado que repase qué forma tiene cada una de las matrices en la expresión anterior (lo puede encontrar en el apunte del curso).

d) **[3 pts.] Regresión de Ridge con descenso del gradiente** o *Recuerdo de Cálculo Diferencial*

En clase, se estudiaron las soluciones directas de mínimos cuadrados, llamadas *ecuaciones normales*, para mínimos cuadrados y mínimos cuadrados regularizados que permiten obtener de manera directa parámetro óptimo θ.

Otra forma de estimar el parámetro óptimo θ es través del método del descenso del gradiente. Este también consiste en minimizar el error cuadrático en todo el conjunto de datos disponible, $\mathcal{J}(\mathbf{y}, \theta) = \frac{1}{2}||\mathbf{y} - \tilde{X}\theta||_2^2$ tal que $\tilde{X} \in R^{N\times(M+1)}$, pero de manera iterativa. No es necesario que usted entienda este método ahora: solamente debe saber que una parte esencial del descenso de gradiente es calcular el gradiente de $\mathcal{J}$ con respecto a θ, $\frac{\partial \mathcal{J}}{\partial \theta}$.

En este problema, usted debe derivar el gradiente $\frac{\partial \mathcal{J}}{\partial \theta}$ para el siguiente funcional:

$$\mathcal{J}_{reg}(\mathbf{y}, \theta) = \frac{1}{2}||\mathbf{y} - \tilde{X}\theta||_2^2 + \frac{1}{2}\rho||\theta||_2^2,$$

que corresponde a la regularización de Tikhonov (con el cuál se obtiene la regresión de Ridge).

[1] Ejemplo: Si hace las dos partes (c) y d)) y en la c) obtiene un 1.5/3 y en la d) obtiene un 3. Entonces, usted tiene 0.75 puntos de bonus en total.

P3. Máxima Verosimilitud [6 pts. (20 %)]

Siete científicos con habilidades experimentales dispares reportan distintas estimaciones de un parámetro μ:

Científico	A	B	C	D	E	F	G
Estimación	-27.020	3.57	8.191	9.898	9.603	9.945	10.056

Discuta cómo encontrar el parámetro óptimo y cuán confiable es cada científico. Para este fin asuma que la estimación de cada científico puede ser considerada como una muestra de distribuciones normales de igual media (μ) pero distintas varianzas ($\sigma_1^2, ..., \sigma_7^2$). Observe de los datos que las estimaciones de los científicos A y B son poco confiables y que el parámetro buscado debería estar entre 9 y 10 ¿es posible abordar este problema usando máxima verosimilitud?