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: 26/04/21

Tarea Teórica #2

Dedicación recomendada: 10 horas

Fecha de entrega: 09 de mayo

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. Clasificación lineal multiclase (30 %)

En clases se vio que es posible extender la formulación de la clasificación lineal binaria para $K > 2$ clases utilizando K funciones lineales de la forma

$$y_k = a^\top x + b$$

y luego asignar la clase de una muestra x mediante $k^* = \text{argmax}_k y_k$. Muestre que este clasificador multiclase es coherente a diferencia de los métodos *one-versus-all* y *one-versus-one*. En particular, muestre que usando este clasificador propuesto:

- No quedan regiones del espacio de entrada sin clasificar
- No existen regiones del espacio de entrada asignadas a más de una clase
- Eligiendo $k = 2$, este criterio colapsa al presentado en el caso binario
- Muestre cómo es la región de decisión y el subconjunto correspondiente a cada clase

P2. Convergencia del algoritmo del perceptrón (35 %)

Considere un conjunto de datos $\{x_i\}_{i=1}^N \subset \mathbb{R}^M$ con etiquetas de clase $t_i \in \{-1, 1\}$ $\forall i \in \{1, \dots N\}$ separable. Pruebe que el algoritmo de entrenamiento del perceptrón converge a un hiperplano de separación en una cantidad finita de pasos, para ello

1. Denote el hiperplano por $f(x) = \theta_1^\top x + \theta_0 = 0$ o de una forma más compacta $\theta^\top \phi(x) = 0$. Sea $z_i = \frac{\phi(x_i)}{||\phi(x_i)||}$. Muestre que la separabilidad implica la existencia de un θ_{sep} tal que $\theta_{\text{sep}}^T z_i t_i \geq 1 \quad \forall i \in \{1, \dots, N\}$

2. Dado un θ_{old} el algoritmo identifica un punto z_i que está mal clasificado, y hace el update $\theta_{\text{new}} \leftarrow \theta_{\text{old}} + z_i t_i$ Muestre que $||\theta_{\text{new}} - \theta_{\text{sep}}||^2 \leq ||\theta_{\text{old}} - \theta_{\text{sep}}||^2 - 1$ y que por lo tanto el algoritmo converge en no más de $||\theta_{\text{start}} - \theta_{\text{sep}}||^2$ pasos.

P3. Modelo generativo de clasificación para K - clases (35 %)

Considere un modelo generativo de clasificación para K clases con priors $p(C_k) = \pi_k$ y probabilidades condicionales dado la clase $p(x|C_k)$. Suponga que tenemos un conjunto de entrenamiento $\{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$. Con $t^{(i)} = (t_k^{(i)})_{k=1}^{K}$ codificado con un *one hot encoding*, es decir, si $x^{(i)}$ es de la clase k, $t_k^{(i)} = 1$ y $t_j^{(i)} = 0 \ \forall j \neq k$. Asuma que todos los datos son tomados independientemente.

1. Encuentre una expresión para la log-verosimilitud del modelo.
2. Muestre que la solución mediante máxima verosimilitud para los prior π_k viene dada por

$$\pi_k = \frac{N_k}{N}$$

   Donde N_k es la cantidad de puntos asignados a la clase k

   Indicación: Utilice multiplicadores de Lagrange

3. Considere ahora que $x|C_k \sim \mathcal{N}(x|u_k, \Sigma)$, muestre que la solución mediante máxima verosimilitud para la media u_k viene dada por

$$u_k = \frac{1}{N_k} \sum_{i=1}^{N} t_k^{(i)} x^{(i)}$$

4. Muestre que la solución mediante máxima verosimilitud para la matriz de covarianza viene dada por

$$\Sigma = \sum_{k=1}^{K} \frac{N_k}{N} S_k$$

   Donde

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N} t_k^{(i)} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^{\top}$$

Anexo: Le pueden ser útiles las siguientes propiedades (bonus de 1 pto en cualquier pregunta por demostrarlas). Para cualquier matriz A invertible

$$\frac{\partial |A|}{\partial A} = |A| A^{-\top}$$

donde $A^{-\top} = (A^{-1})^{\top}$, además

$$\frac{\partial x^{\top} A^{-1} x}{\partial A} = -A^{-\top} x x^{\top} A^{-\top}$$

y finalmente para cualquier matriz A simétrica

$$\frac{\partial (x-s)^{\top} A (x-s)}{\partial s} = -2A(x-s)$$