Prueba de Aprendizaje Estadística EMALCA 2025

26 de Junio, 2025 **Nombre:** Soluciones

Ejercicio 1. (15 puntos)

Sea la función verdadera $f(x) = x^2$, y suponga que las observaciones siguen el modelo

$$Y_i \;=\; f(x_i) + \varepsilon_i \;=\; x_i^2 + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Disponemos siempre de los mismos tres puntos de entrenamiento $x_1 = 1, \quad x_2 = 2, \quad x_3 = 4$, pero cada re-muestreo genera nuevos ruidos ε_i. Deseamos predecir en el punto de consulta

$$x_* = 3 \quad \Longrightarrow \quad f(x_*) = 3^2 = 9.$$

a) Para cada modelo k-NN $k = 2, 3$, calcule el error cuadrático medio esperado MSE = Sesgo^2 + Varianza.

b) Determine qué valor de k minimiza la MSE esperada y explique el resultado en términos del compromiso sesgo–varianza.

a) $K=2$

$$\hat{f}_{(2)} = \frac{Y_2 + Y_3}{2} = \frac{(4+\epsilon_2) + (16+\epsilon_3)}{2} = 10 + \frac{(\epsilon_2+\epsilon_3)}{2}$$

$k=3$ ó $f_{(3)} = \frac{Y_1+Y_2+Y_3}{3} = \frac{(1+\epsilon_1)+(4+\epsilon_2)+(16+\epsilon_3)}{3} = 7 + \frac{(\epsilon_1+\epsilon_2+\epsilon_3)}{3}$

Error Cuadratico Medio: CECM)

2-NN Cesgo: $(9-10)^2 = 1$

Varianza: $\text{Var}\left(\frac{\epsilon_2+\epsilon_3}{2}\right) = \frac{1}{2}$

ECM: $1 + \frac{1}{2} = 1.5$

3-NN Cesgo: $(9-7)^2 = 4$

Varianza: $\text{var}\left(\frac{\epsilon_1+\epsilon_2+\epsilon_3}{3}\right) = \frac{1}{3}$

ECM: $4 + 1/3 = 13/3$

El cesgo de 2-NN es menos que 3-NN y la varianza es mas.
En combinacion el ECM de 2-NN es mejor

Ejercicio 2. (5 puntos)

Considere una situación en que tiene predecir si un paciente tiene cáncer de pulmón. En la población, solo el 1 % de los pacientes tienen cáncer de pulmón. La red neuronal que usted entreno determina la categoría (cáncer, no cáncer) correctamente 99 % de los casos del conjunto de datos de prueba. Explique de manera concisa si tiene suficiente información para determinar si su método funciona bien basado en esos porcentajes.

no hay sufiente informacion. Por ejemplo, el metodo podria clasificar a todos los pacientes "no cancer" y obtener el 99% de los casos correctos pero cometeria un error para todos los pacientes que si tienen cancer!