5. Sesgo y Varianza

Ejercicio 5.1. Queremos demostrar la descomposición **Sesgo-Varianza** para problemas de regresión:

$$Error\_esperado(x^{(i)}; L) = \mathrm{E}_{D_n}\left[error\big(y^{(i)}, \hat{h}_{(L,D_n)}(x^{(i)})\big)\right]$$
$$\overset{\text{reg}}{=} \Big(\text{Sesgo}\left[\hat{h}_{(L,D_n)}(x^{(i)}); f(x^{(i)})\right]\Big)^2 + \text{Var}\left[\hat{h}_{(L,D_n)}(x^{(i)})\right] + \text{Var}(\varepsilon)$$

I) Según lo visto en clase, ¿cuál es la métrica de error que utilizamos en regresión para la descomposición de sesgo-varianza?

II) Demostrar la descomposición de sesgo y varianza haciendo las suposiciones hechas en clase.

III) Si ahora definimos error como $error(y^{(i)}, pred^{(i)}) = y - pred$, ¿Se obtiene la misma descomposición? (en caso de no poder llegar a una fórmula aclarar dónde encuentran un problema).

Ejercicio 5.2. Verdadero o Falso:

(a) Aumentar la cantidad de datos suele ayudar a contrarrestar problemas de varianza.

(b) Aumentar la cantidad de datos suele ayudar a contrarrestar problemas de sesgo.

(c) Un modelo muy complejo suele producir sesgo alto.

(d) Un modelo muy complejo suele producir varianza alta.

(e) Sesgo alto está asociado a problemas de undefitting.

(f) Varianza alta está asociado a problemas de overfitting.

Ejercicio 5.3. Supongamos que se construyen cuatro clasificadores para discriminar grabaciones de conversaciones en inglés contra grabaciones de conversaciones en español. La siguiente tabla muestra los resultados obtenidos según cuatro algoritmos (para ser más específicos, 4 configuraciones, es decir algoritmos junto a sus hiperparámetros).

(a) ¿Cuál de estos modelos parecen estar sobreajustando y cuáles subajustando?

(b) ¿Qué intentarías hacer si pensás que un modelo sufre de subajuste o sobreajuste?

(c) ¿Cuáles de las configuraciones dirían que sufren de alto sesgo?

(d) ¿Cuáles de las configuraciones dirían que sufren de alta varianza?

(e) ¿Dirías que alta varianza implica sobreajuste? ¿Dirías que sobreajuste implica alta varianza?

(f) Si C1 y C2 fueran árboles, ¿qué características pensás que tendrían?

(g) Imaginen ahora que las grabaciones tienen mucho ruido de fondo y hasta un humano tiene problemas para detectar su origen, ¿dirían que la configuración C2 tiene sesgo alto? (suponer que los demás resultados no existen)

Configuración:	**C1**	**C2**	**C3**	**C4**
Accuracy (sobre entrenamiento)	.99	.90	.90	.99
Accuracy (sobre validación)	.89	.89	.75	.98

Tabla 2: Sesgo y Varianza