Examen final de Aprendizaje Estadístico

PREGUNTAS DE LA PARTE DE MANUEL MUCIENTES

---

1. **Si tenemos un alto error de entrenamiento y de test. ¿Sirve de algo añadir más muestras?**
   Este es un caso de subaprendizaje. Agregar más muestras no ayudará, ya que el problema está en la capacidad del modelo para encontrar patrones relevantes en los datos. Esto se da porque el modelo es demasiado simple y/o por usar características poco significativas.

2. **Si tenemos un alto error de entrenamiento y de test. ¿Sirve de algo añadir más variables predictoras?**
   Este es un caso de subaprendizaje. Agregar más variables predictoras puede ayudar si estas son significativas y aportan información relevante al modelo. Hay que evitar añadir variables no relevantes o correlacionadas con otras para evitar complejidad innecesaria.

3. **Si tenemos un alto error de test, pero bajo de entrenamiento. ¿Sirve de algo añadir más muestras?**
   Este es un caso de sobreaprendizaje. Agregar más muestras puede ayudar a mejorar el modelo, ya que el problema está en que el modelo aprendió los datos de entrenamiento demasiado bien y no generaliza bien a nuevos datos.

4. **¿Hay que normalizar las entradas para knn?**
   Sí. KNN usa la distancia entre puntos para obtener los vecinos más cercanos. Esta medida, típicamente la distancia euclidiana, es sensible a la escala de las variables, por lo que el cálculo se verá dominado por las variables con mayor escala. Por tanto, hay que normalizar las entradas para que todas tengan la misma importancia.

5. **¿Hay que normalizar las salidas cuando es knn para regresión?**
   No es necesario. KNN para regresión da como salida la media de los valores de los vecinos más cercanos, por lo que el cálculo no depende de la escala de las salidas. Sin embargo, es recomendable normalizar las salidas cuando queremos comparar modelos y usamos métricas de evaluación que dependen de la escala de las salidas, como puede ser el MSE, para que la comparación sea justa.

6. **¿Como procedemos si al llegar a una intersección de un árbol, falta el valor de la variable divisora?**
   Se puede proceder de varias formas. Una forma simple es que el ejemplo que carece del valor en dicha variable, vaya por las ambas ramas. Al final, ese ejemplo estará en, al menos, dos hojas, por lo que se le asignará la media de los valores de las hojas a las que llegó, en caso de regresión. En un caso de clasificación, sumamos los vectores de probabilidad de las hojas a las que llegó y dividimos por el número de hojas.

7. **¿Que relación hay entre el "mínimo número de ejemplos" y el sobre aprendizaje en un árbol?**
   Un mínimo número de ejemplos muy bajo puede llevar a sobreaprendizaje. LLevado al caso extremo, $n_{min} = 1$, se harán divisiones hasta que cada hoja tenga un único ejemplo, lo que llevará a un árbol muy profundo y complejo, ajustándose demasiado bien a los datos de entrenamiento y perdiendo capacidad de generalización.

8. **Una estructura como la del ejercicio 1 de neuronas. Decir como quedaría $z^{(2)}$ dejando indicadas las operaciones con matrices sin resolverlas.**
   $$z^{(2)} = W^{(1)} \cdot a^{(1)} + b^{(1)}, \quad \text{con } a^{(1)} \equiv x$$

9. **Una estructura como la del ejercicio 1 de neuronas. Decir como quedaría $a^{(3)}$ dejando indicadas las operaciones con matrices sin resolverlas.**
   $$a^{(2)} = f(z^{(2)}), \qquad z^{(3)} = W^{(2)} \cdot a^{(2)} + b^{(2)}, \qquad a^{(3)} = f(z^{(3)}) \underset{f=\mathbb{1}}{=} z^{(3)}$$

10. **Una estructura como la del ejercicio 1 de neuronas. Decir como quedaría $a^{(2)}$, el error de la capa de salida y el error de la capa 2, dejando indicadas las operaciones con matrices sin resolverlas.**
    $$a^{(2)} = f(z^{(2)}), \qquad \delta^{(3)} = (a^{(3)} - y) \circ \cancel{f'(z^{(3)})}\ ^{1}, \qquad \delta^{(2)} = [(W^{(2)})^T \cdot \delta^{(3)}] \circ f'(z^{(2)})$$

11. **Diferencia entre *full connected* y *local connected* en la estructura de una neurona.**
Si hablamos de una única neurona, decimos que esta densamente conectada (*full connected*) si está conectada a todas las neuronas de la capa anterior, mientras que si está conectada solo a un subconjunto de neuronas, decimos que está localmente conectada (*local connected*), lo que reduce el número de parámetros y la complejidad del modelo. Este último tipo de conexión puede ir acompañada de compartición de pesos.

12. **Dada una lista de observaciones con coeficientes α_i, y un valor de C, decir cuáles de ellos están en el margen.**
Dentro del margen están los vectores de soporte, es decir, los que tienen un $\alpha_i > 0$. Si $\xi_i = 0$ (y por tanto $0 < \alpha_i < C$), el vector de soporte está en el límite del margen. El resto de vectores de soporte tienen $\xi_i > 0$ y $\alpha_i = C$.

13. **Al aumentar "C" en SVC, ¿tendemos al sobreaprendizaje o al subaprendizaje?**
Aumentar C en SVC tiende al sobreaprendizaje. Al hacer esto, estamos buscando una mayor precisión al separar las clases, lo que nos lleva a reducir el margen. Al reducir el margen, reducimos el número de vectores de soporte y tendemos a un modelo sobreaprendido.

14. **¿Qué relación hay entre el número de vectores de soporte y "C"?**
Al aumentar C estamos buscando una mayor precisión al separar las clases, lo que nos lleva a reducir el margen. Al reducir el margen, reducimos el número de vectores de soporte.

15. **¿Como afecta gamma al kernel radial?**
El kernel radial presenta la forma funcional de una gaussiana, por lo que podemos hacer la analogía de γ como la inversa de la varianza de una gaussiana. Un valor alto de γ implica una gaussiana con varianza baja (estrecha), permitiendo pocos vectores de soporte, lo que puede llevar al sobreaprendizaje. Por otro lado, un valor bajo de γ implica una gaussiana con varianza alta (ancha), lo que hace que el modelo recaiga en más vectores de soporte.

16. **¿Que es el α_m en ada boosting?**
El α_m en AdaBoost es el peso que se le asigna al clasificador débil $G_m(x)$ en la combinación lineal de clasificadores débiles que forman el clasificador fuerte $G(x)$. Este peso depende del error de clasificación del clasificador débil.

17. **¿Como afecta α_m a lo que contribuye el árbol m a la salida?**
α_m es directamente proporcional a la contribución del árbol m a la salida. Cuanto menor sea el error de clasificación de ese árbol, mayor será el peso que se le asigne en la combinación y, por tanto, mayor será su contribución a la salida final.

18. **¿Qué relación hay en random forest entre m y una baja calidad de los predictores?**
Si tenemos un gran número de predictores de baja calidad, un valor pequeño de m (número de predictores que cogemos) puede dar un mal rendimiento ya que hay una probabilidad baja de seleccionar un predictor relevante. Si la calidad de los predictores es baja es mejor usar un valor alto de m para que haya más probabilidad de seleccionar predictores relevantes de cara a la construcción del árbol.

19. **¿Qué relación hay en random forest entre m y el sobreaprendizaje?**
Random Forest introduce una aleatoriedad en la selección de predictores para cada árbol para disminuir la correlación entre los árboles y evitar el sobreaprendizaje. Si m es muy grande, los árboles serán muy similares y por tanto estarán altamente correlacionados, dando una varianza alta y tendiendo al sobreaprendizaje. Por otro lado, si m es pequeño, los árboles tenderán a ser diferentes y se mantendrá la varianza baja.