Trabajo 3: Cuestiones de Teoría

Ignacio Vellido Expósito

May 24, 2019

1. ¿Podría considerarse Bagging como una técnica para estimar el error de predicción de un modelo de aprendizaje?. Diga si o no con argumentos. En caso afirmativo compárela con validación cruzada.

   Sí, ya que bagging se basa en bootstrapping, y este remuestrea los datos quedando una parte de ellos sin participar en el entrenamiento. Estos datos permiten estimar el error conocido como "out-of-bar error".

   En relación con cross-validation, ambas técnicas remuestrean los datos, pero mientras que bagging selecciona de la muestra con reemplazamiento CV no.

2. Considere que dispone de un conjunto de datos linealmente separable. Recuerde que una vez establecido un orden sobre los datos, el algoritmo perceptron encuentra un hiperplano separador interando sobre los datos y adaptando los pesos de acuerdo al algoritmo

Algorithm 1 Perceptron

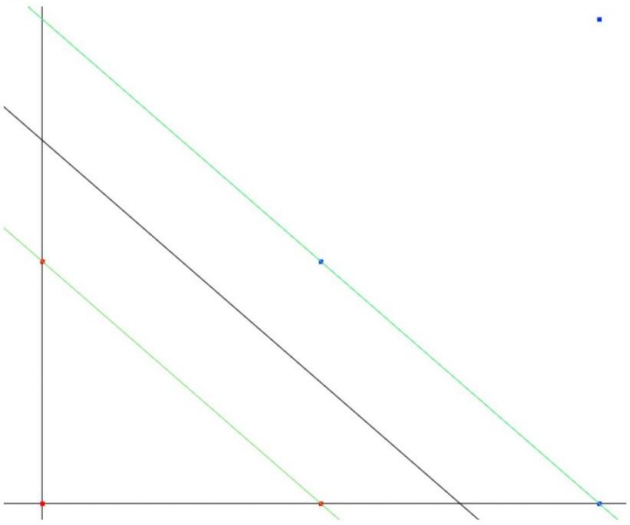
1: **Entradas**: $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, $\mathbf{w} = 0$, $k = 0$
2: **repeat**
3: $k \leftarrow (k + 1) \bmod n$
4: **if** $\text{sign}(y_i) \neq \text{sign}(\mathbf{w}^T \mathbf{x}_i)$ **then**
5: $\mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$
6: **end if**
7: **until** todos los puntos bien clasificados

   Modificar este pseudo-código para adaptarlo a un algoritmo simple de SVM, considerando que en cada iteración adaptamos los pesos de acuerdo al caso peor clasificado de toda la muestra. Justificar adecuadamente/matematicamente el resultado, mostrando que al final del entrenamiento solo estaremos adaptando los vectores soporte.

   Pregunta no contestada

3. Considerar un modelo SVM y los siguientes datos de entrenamiento: Clase-1:{(1,1),(2,2),(2,0)}, Clase-2:{(0,0),(1,0),(0,1)}
   a) Dibujar los puntos y construir por inspección el vector de pesos para el hiperplano óptimo y el margen óptimo.
   b) ¿Cuáles son los vectores soporte?
   c) Construir la solución en el espacio dual. Comparar la solución con la del apartado (a)

a) Tenemos:



Sabemos que se debe cumplir la ecuación:

$$w^T x + b = 0 \tag{1}$$

Por inspección obtenemos como parámetros:

$$w = [-1, -1] \quad b = 1,5 \tag{2}$$

Generalizando:

$$w = [-c, -c] \quad b = 1,5c \tag{3}$$

Gráficamente vemos que el margen óptimo es 1. Así que, por la fórmula del margen tenemos que:

$$\frac{2}{\sqrt{2}c} = 1 \tag{4}$$

Resolviendo:

$$c = \frac{2}{\sqrt{2}} \tag{5}$$

Por tanto, tenemos como vector de pesos:

$$w = [-\frac{2}{\sqrt{2}}, -\frac{2}{\sqrt{2}}] \tag{6}$$

b) Son: $(0,1), (1,0), (2,0)$ y $(1,1)$

c) Apartado no contestado

4. ¿Cúal es el criterio de optimalidad en la construcción de un árbol? Analice un clasificador en árbol en términos de sesgo y varianza. ¿Que estrategia de mejora propondría?

   En términos de optimalidad, se pretende obtener el árbol más simple que explique los datos.

   Los árboles tienen poco sesgo puesto que no se están imponiendo restricciones sobre la función objetivo. A su vez, puesto que tienden al sobreajuste, cuentan con una gran varianza ya que la mínima alteración en los datos puede dar con una construcción del árbol totalmente diferente.

   Para evitar este alto grado de varianza se pueden utilizar técnicas vistas como el promediar con bagging.

5. ¿Cómo influye la dimensión del vector de entrada en los modelos: SVM, RF, Boosting and NN?.

   En general, una mayor dimensionalidad implica un mayor riesgo de sobreajuste ya que necesitamos una muestra más grande para estimar el mayor número de parámetros. En los diferentes modelos:

   - SVM: Como cuenta con un alto grado de regularización puede evitar el sobreajuste.

   - Boosting: Puesto que los árboles de decisión sufren de por sí de sobreajuste boosting es afectado aún más.

   - RF: Como utiliza distintas características y datos para diferentes árboles, no le afecta en gran manera.

6. El método de Boosting representa una forma alternativa en la búsqueda del mejor clasificador respecto del enfoque tradicional implementado por los algoritmos PLA, SVM, NN, etc. a) Identifique de forma clara y concisa las novedades del enfoque; b) Diga las razones profundas por las que la técnica funciona produciendo buenos ajustes (no ponga el algoritmo); c) Identifique sus principales debilidades; d) ¿Cuál es su capacidad de generalización comparado con SVM?

   La novedad se sitúa en el hecho de realizar cambios sobre la probabilidad de la muestra, haciendo que algunos datos influyan en mayor o menor medida en el error.

   Produce buenos resultados ya que, al igual que SVM, no solo reduce el error en la muestra sino que hace la búsqueda del mejor margen. De esta manera podemos seguir reduciendo el error fuera de la muestra tras minimizar el de dentro, obteniendo gran generalización.

   Este método, al ser dependiente de los clasificadores que se utilice, puede producir sobreajuste. También falla cuando trabajamos con pequeñas muestras o poco representativas de la población, pues supone que los datos son representativos

7. Discuta pros y contras de los clasificadores SVM y Random Forest (RF). Considera que SVM por su construcción a través de un problema de optimización debería ser un mejor clasificador que RF. Justificar las respuestas.

   Respecto a SVM, tiene una gran flexibilidad gracias a la introducción del kernel e incluyendo el parámetro C cuenta con mayor potencial de generalización, lo que le permite menor grado de sobreajuste que los árboles. Como contra, no tienen interpretabilidad.

   Por otra parte Random Forest permite obtener gran precisión con los datos partiendo de los árboles de decisión mientras se reduce su alta variabilidad. Esto es así puesto que al estar construido sobre la técnica de bagging se construyen árboles no correlados.

   La mayor desventaja que tienen es la pérdida de interpretabilidad que se buscaba con los árboles de decisión. También cuentan con mayor gasto computacional al tener que construir múltiples árboles.

   Comparando ambos métodos, los dos pueden obtener un buen grado de generalización, pese a que RF sea más dependiente del tipo y la cantidad de datos.

8. ¿Cuál es a su criterio lo que permite a clasificadores como Random Forest basados en un conjunto de clasificadores simples aprender de forma más eficiente? ¿Cuales son las mejoras que introduce frente a los clasificadores simples? ¿Es Random Forest óptimo en algún sentido? Justifique con precisión las contestaciones.

   Pregunta no contestada

9. En un experimento para determinar la distribución del tamaño de los peces en un lago, se decide echar una red para capturar una muestra representativa. Así se hace y se obtiene una muestra suficientemente grande de la que se pueden obtener conclusiones estadísticas sobre los peces del lago. Se obtiene la distribución de peces por tamaño y se entregan las conclusiones. Discuta si las conclusiones obtenidas servirán para el objetivo que se persigue e identifique si hay algo que lo impida.

   Pregunta no contestada

10. Identifique que pasos daría y en que orden para conseguir con el menor esfuerzo posible un buen modelo de red neuronal a partir una muestra de datos. Justifique los pasos propuestos, el orden de los mismos y argumente que son adecuados para conseguir un buen óptimo. Considere que tiene suficientes datos tanto para el ajuste como para el test.

    1. Partir de un número de unidades ocultas $m = \frac{1}{d}\sqrt{N}$, siendo d la dimensión de entrada, de manera que se podamos hacer $E_{out} \longrightarrow Ein$ y $Ein \longrightarrow 0$

    2. Ajustar el número de unidades usando cross-validation de manera que podamos obtener los mejores resultados.

    3. Utilizar técnicas de regularización para evitar el sobreajuste, ya que la potencia de la red neuronal hace que tienda a hacerlo.

    4. Reducir el tiempo de cómputo con un learning rate variable.

BONUS: Los BONUS solo serán tenidos en cuenta si en el cuestionario obligatorio se ha consiguido al menos un 75 % de los puntos totales.

1. (1.5 puntos) Suponga que durante 5 semanas seguidas, recibe un correo postal que predice el resultado del partido de futbol del domingo, donde hay apuestas substanciosas. Cada lunes revisa la predicción y observa que la predicción es correcta en todas las ocasiones. El día de después del quinto partido recibe una carta diciendole que si desea conocer la predicción de la semana que viene debe pagar 50.000€. ¿Pagaría?
   a) ¿Cuantas son las posibles predicciones gana-pierde para los cinco partidos?
   b) Si el remitente desea estar seguro de que al menos una persona recibe de él la predicción correcta sobre los 5 partidos, ¿Cual es el mínimo número de cartas que deberá de enviar?
   c) Después de la primera carta prediciendo el resultado del primer partido, ¿a cuantos de los seleccionados inicialmente deberá de enviarle la segunda carta?
   d) ¿Cuantas cartas en total se habrán enviado depués de las primeras cinco semanas?
   e) Si el coste de imprimir y enviar las cartas es de 0.5€ por carta, ¿Cuanto ingresa el remitente si el receptor de las 5 predicciones acertadas decide pagar los 50.000€?
   f) ¿Puede relacionar esta situación con la función de crecimiento y la credibilidad del ajuste de los datos?

   Pregunta no contestada

2. (1.5 puntos) Considere un modelo de red neuronal con dos capas totalmente conectadas: d unidades de entrada, n_H unidades ocultas y c unidades de salida. Considere la función de error definida por $J(\mathbf{w}) \equiv \frac{1}{2}\sum_{k=1}^{c}(t_k - c_k)^2 = \frac{1}{2}||\mathbf{t} - \mathbf{z}||^2$, donde el vector $\mathbf{t}$ representa los valores de la etiqueta, $\mathbf{z}$ los valores calculados por la red y $\mathbf{w}$ los pesos de la red. Considere que las entradas a la segunda capa se calculan como $z_k = \sum_{j=0}^{N_H} y_j w_{kj} = \mathbf{w}_k^t \mathbf{y}$ donde el vector $\mathbf{y}$ representa la salida de la capa oculta.
   a) Deducir con todo detalle la regla de adaptación de los pesos entre la capa oculta y la salida.
   b) Deducir con todo detalle la regla de adaptación de los pesos entre la capa de entrada y la capa oculta.

   Usar θ para notar la función de activación.

   Pregunta no contestada