Aprendizaje Automático: Cuestionario 1

David Cabezas Berrido

1 de mayo de 2020

1. 1

Utilizaría validación cruzada (1000 épocas de tamaño 999 para cada modelo).

Para cada uno de los 5 modelos, calcularía el error de validación cruzada $E_{CV} = \frac{1}{1000} \sum_{i=1}^{1000} E_{val}(g_i^-)$ donde $E_{val}(g_i^-)$ es el error en el elemento (x_i, y_i) de la muestra de la función resultante de entrenar el modelo con los 999 datos restantes.

Este error E_{CV} proporciona un estimador insesgado de $\overline{E}_{out}(999)$, E_{out} esperado del modelo entrenado sobre conjuntos de datos de 999 elementos. Por eso considero un buen criterio escoger el modelo con el menor error de validación cruzada.

Nuestra solución final es el modelo elegido entrenado con los 1000 datos, luego tenemos un ligero beneficio en la curva de aprendizaje al pasar de 999 datos a 1000, así que podemos (en promedio) estimar la cota $E_{out} \leq E_{CV}$ del modelo elegido.

Learning From Data, 4.3.3

2. 2

Etiquetando las clases como $\{0, 1\}$, la función con el error de clasificación más bajo es:

$f_{\mathcal{P}}(x) = \begin{cases} 1 & \text{si } \mathcal{P}(y = 1|x) \geq \frac{1}{2} \\ 0 & \text{en otro caso} \end{cases}$ para cada muestra $x \in \mathcal{X}$.

Para cualquier otra función g de $\mathcal{X}$ a $\{0, 1\}$, $E_{out}(f_{\mathcal{P}}) \leq E_{out}(g)$

Regla de Bayes del tema de Modelos Lineales

3. 3

Para cada n (muestra (x_n, y_n)), $E_n = \text{máx}(0, 1 - y_n w^T x_n)^2$ es diferenciable por ser composición de diferenciables: $z \mapsto \text{máx}(0, z)^2$ es derivable con derivada $z \mapsto 2\,\text{máx}(0, z)$ y $w \mapsto 1 - y_n w^T x_n$ es diferenciable con gradiente $w \mapsto -y_n x_n$.

Usando la regla de la cadena: $\nabla E_n = -y_n x_n 2\,\text{máx}(0, 1 - y_n w^T x_n)$ Por tanto, la regla de adaptación de gradiente descendente es $w^{(t+1)} = w^{(t)} + \eta y_n x_n 2\,\text{máx}(0, 1 - y_n {w^{(t)}}^T x_n)$.

Utilizando $\nabla \frac{1}{N} \sum_{n=1}^{N} E_n = \frac{1}{N} \sum_{n=1}^{N} \nabla E_n$, concluimos que para el error promedio, la regla de adaptación es $w^{(t+1)} = w^{(t)} + \eta \sum_{n=1}^{N} y_n x_n 2\,\text{máx}(0, 1 - y_n {w^{(t)}}^T x_n)$.

(La constante $\frac{1}{N}$ no es necesaria, va incluida en el learning rate)

4. 4

La única correcta es la tercera opción.

La 2 no es correcta porque existen casos de funciones f de alta complejidad, para los cuales necesitaría utilizar una clase de funciones H también de alta complejidad para aproximar f en la muestra. Esto provocaría que hubiese alta varianza en la clase y necesite muchos datos para garantizar que esa aproximación sigue siendo buena fuera de la muestra, posiblemente más de los que poseo.

Por tanto, no puedo asegurar que se cumpla la opción 2, luego la 1 tampoco.

La opción 3 sí que puedo garantizarla. Si utilizo una clase de funciones H de baja complejidad, la varianza de la clase será baja y el error fuera de la muestra será similar al error en la muestra.
Por tanto, si obtengo un error bajo en la muestra lograré a), y en caso contrario sabré que ha ocurrido b).

5. 5

En este caso (determinístico) no se puede aprender la función f, ya que es una función elegida de forma arbitraria que no está regida (que sepamos) por ninguna distribución. Es imposible conocer f fuera de la muestra.
Learning From Data 1.3.1

6. 6

Añadiría la hipótesis de que la los valores proporcionados fuesen muestras independientes e identicamente distribuidas de una distribución de probabilidad $\mathcal{P}$. Esta hipótesis es algo restrictiva y no me permite asegurar que vaya a obtener la hipótesis óptima de H, pero al menos me permite aproximar f con cierta probabilidad.
Aun así, el teorema de No-Free-Lunch me advierte de que es posible que no dé con el algoritmo adecuado para aprender la función f, pero puedo intentar estimar la distribución $\mathcal{P}$ por la muestra y usar algún algoritmo adecuado para esa distribución.
Learning From Data 1.3.2

7. 7

Queremos que $h(z)$ minimice la función $E(x) = \sum_{i=1}^{N} |x - z_i|$. E es suma de funciones convexas y por tanto convexa, basta encontrar donde se anula la derivada (E se puede subderivar obteniendo:
$E'(x) = \sum_{i=1}^{N} \text{sign}(x - z_i)$
Como $\text{sign}(x - z_i) \in \{-1, 0, 1\}$, $E'(x)$ sólo se anula cuando hay el mismo número de +1 que de -1, esto equivale a que x deje (estrictamente) a la izquierda y a la derecha el mismo número de valores z_i, por tanto E alcanza un mínimo en la mediana.
Tomamos $h(\mathbf{z}) = \text{Mediana}(z_1, \dots z_N)$.
Cuando aparece un error aberrante (un outlier), la media (óptimo del error cuadrático) se desplaza de forma lineal con el outlier (el error aberrante influye en la media dividido por N), sin embargo la mediana como mucho se desplaza un dato, incluso aunque el error aberrante se vaya a infinito (lo que haría que la media también se vaya a infinito). Por tanto lo considero una buena decisión.

8. 8

$P_x[h(x) \neq y] = (1 - \mu)(1 - \lambda) + \mu\lambda$ Este error no depende de μ para $\lambda = \frac{1}{2}$
Llamando f_1 a la versión con ruido de f, tenemos $P_x[h(x) \neq f(x)] = P_x[h(x) \neq f_1(x)] = \mu$ y $P_x[f_1(x) = y] = \lambda$.
h predice mal y cuando h difiere de f_1 y f_1 acierta y cuando h coincide con f_1 y f_1 falla.
$P_x[h(x) \neq y] = P_x[h(x) = f_1(x)] \cdot P_x[f_1(x) \neq y] + P_x[h(x) \neq f_1(x)] \cdot P_x[f_1(x) = y] = (1-\mu)(1-\lambda) + \mu\lambda$
Si queremos que esta expresión no dependa de μ, la derivada respecto de μ tendrá que ser 0, de donde obtenemos $-(1 - \lambda) + \lambda = 0 \Leftrightarrow \lambda = \frac{1}{2}$.
Este ejercicio está en el telecurso de Learning From Data acompañado de una pequeña pista. http://work.caltech.edu/homev

9. 9

Regla de clasificación para un x: Si $g(x) \geq k$ predice +1, en caso contrario predice -1, para alguna constante k fija. Si no tenemos en cuenta los costes y queremos minimizar la probabilidad de error, tomamos $k = 0{,}5$.

Al incorporar los costes: costo(+1) $= 0 \cdot P[y = +1|x] + 10 \cdot P[y = -1|x] = 10(1 - g(x))$ costo(-1) $= 1000 \cdot P[y = +1|x] + 0 \cdot P[y = -1|x] = 1000g(x)$
Debemos hacer la predicción con menos costo:
costo(+1)$\leq$costo(-1)$\Leftrightarrow 10(1 - g(x)) \leq 1000g(x) \Leftrightarrow g(x) \geq \frac{10}{1000+10} = \frac{1}{101}$
Nueva regla de clasificación para un x: Si $g(x) \geq \frac{1}{101}$ predice +1, en otro caso predice -1.
Cuanto mayor es el coste del falso positivo respecto al del falso negativo, más seguros debemos estar de que la etiqueta es -1 para predecir -1.

10. 10

Es irrelevante.
El teorema de No-Free-Lunch nos asegura que todos los algoritmos son equivalentes en media para todas las posibles funciones objetivo, por tanto no hay ninguno mejor que otro.

11. 11

$\mathcal{H}$ tiene suficiente complejidad para explicar la función objetivo en la muestra, pero esto provoca que la clase presente una gran varianza. Por tanto la función obtenida como mejor aproximación de la función objetivo depende en gran medida de la muestra, por lo que la tendencia a sobreajustar es alta, que seamos capaces de explicar la función objetivo en la muestra no nos garantiza que podamos hacerlo fuera.
A medida que decrece la complejidad de $\mathcal{H}$, su varianza se va reduciendo y por tanto también la tendencia a sobreajustar, posiblemente reduciendo el error fuera de la muestra. Por otra parte, la clase va perdiendo capacidad para explicar la función objetivo, lo que provoca que aparezca un error por sesgo entre la función media de la clase $\mathcal{H}$ y la función objetivo.
La disminuición del error por varianza compensa al ligero aumento del error por sesgo, por lo que el error fuera de la muestra va disminuyendo. Siempre y cuando no reduzcamos demasiado la complejidad, lo que haría el ruido determinista muy significativo y empezaría a aumentar el error debido a la incapacidad de la clase $\mathcal{H}$ para aproximar la función objetivo.
También puede existir error por el ruido en la muestra, el cual seguimos manteniendo.

12. 12

Considero w fijo.
Dado un punto de la muestra (x_n, y_n), distinguimos dos casos:
Si el punto está bien clasificado, entonces $sign(w^T x_n) = y_n$. Por lo que
$0 = [[sign(w^T x_n) \neq y_n]] \leq$ máx$(0, 1 - y_n w^T x_n)$
Si el punto está mal clasificado, entonces $sign(w^T x_n) \neq y_n$ y se tendrá $y_n w^T x_n < 0$, por lo que
máx$(0, 1 - y_n w^T x_n) = 1 - y_n w^T x_n > 1 = [[sign(w^T x_n) \neq y_n]]$
Utilizando esta cota en cada uno de los sumandos concluimos que
$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} [[sign(w^T x_n) \neq y_n]] \leq \frac{1}{N} E_n(w)$