Trabajo 3 - Teoría
Aprendizaje Automático

Francisco Solano López Rodríguez

May 31, 2018

1. Tanto "bagging" como validación-cruzada cuando se aplican sobre una muestra de datos nos permiten dar una estimación del error de un modelo ajustado a partir de dicha muestra. Enuncie las diferencias y semejanzas entre ambas técnicas. Diga cual de ellas considera que nos proporcionará una mejor estimación del error en cada caso concreto y por qué.

   Solución:

   La semejanza entre ellas, es que ambas generan conjuntos de entrenamiento sobre los que crean modelos de predicción y sobre los que pueden obtener una estimación del error del modelo usando un conjunto test formado por elementos que no se encuentran en el conjunto de entrenamiento.

   La validación cruzada divide en dos conjuntos disjuntos el conjunto original de datos, uno de ellos lo usa para el train y el otro para el test. Por ejemplo k Fold Validation divide en K particiones, una de ellas se usa para el test y las K-1 para el train y se repite K veces con cada partición.

   Bagging genera B conjuntos de entrenamiento usando bootstraping, es decir remuestrea de forma aleatoria y con reemplazamiento, con lo que en un conjunto podría haber elementos repetidos. A igual que puede haber elementos repetidos, puede ser que en un conjunto haya elementos del conjunto original que no hayan sido tomados para el conjunto de entrenamiento, pues bien dichos elementos serán los que formarán el conjunto test. En promedio este conjunto contiene una tercera parte del conjunto original.

   Como vemos ambos crean conjuntos de entrenamiento y de test para validar, la diferencia es la forma de la que lo hacen.

   Una ventaja en cross validation es que podemos elegir el tamaño del conjunto train y test, por ejemplo un 10-fold validation que usará un 90% de los datos para el train y el 10% restante para el test. En cambio en bagging no podemos hacer esto y en promedio se tendrá dos tercios para el train y un tercio para el test.

   Bagging puede ser una buena opción para mejorar la generalización. Validación cruzada podría ser recomendable para elegir hiperparámetros.

---

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

```
1  Entradas: (x_i, y_i), i = 1,...,n, w = 0, k = 0
2  repeat
3     k ← (k+1) mod n
4     if sign(y_i) ≠ sign(W^T x_i) then
5        w ← w + y_i 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.

   Solución:

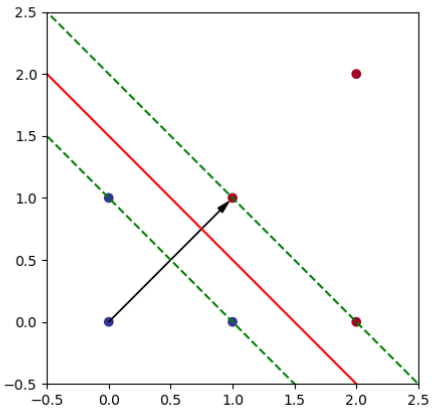
```
Entradas: (x_i, y_i), i = 1,...,n, w = 0, k = 0
repeat
    peor = 1

    for k = 1 to n do
        if y_k(W^T x_k + b) ≤ 1 then
            if |W^T x_peor + b| < |W^T x_k + b| then
                peor = k
            end if
        end if
    end for

    w ← w + y_peor x_peor
    b ← b + y_peor

until criterio de terminacion
```

Funciona igual que el perceptron solo que adapta los pesos de acuerdo al peor clasificado, luego si los datos son linealmente separables convergerá igual que lo hacía el perceptrón. Los peor clasificados serán aquellos que estén mal clasificados es decir $\text{sign}(y_i) \neq \text{sign}(W^T x_i)$ y además cuya distancia al hiperplano sea la mayor, esta se calcula como $|W^T x_k + b| / \parallel w \parallel$. Con forme el algoritmo converja a la solución la distancia del peor clasificado cada vez será menor, luego llegará un momento a partir de cual solo adaptemos los vectores soporte ya que serán los más cercanos al hiperplano.

---

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.
   Solución:



   Las lineas verdes discontinuas representan los margenes óptimos, la linea roja el hiperplano óptimo cuyo vector de pesos es (1,1).

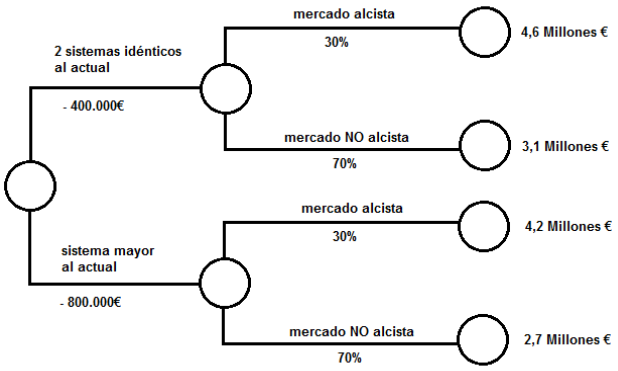
   (b) ¿Cuáles son los vectores soporte?
   Solución:
   Los vectores de soporte de la clase 1 son (1,1) y (2,0). Los vectores de soporte de la clase 2 son (1,0) y (0,1).

   (c) Construir la solución en el espacio dual. Comparar la solución con la del apartado (a)
   Solución:

---

4. Una empresas esta valorando cambiar su sistema de proceso de datos, para ello dispone de dos opciones, la primera es adquirir un nuevo sistema compuesto por dos sistemas idénticos al actual a 200.000 euros cada uno, y la segunda consiste en adquirir un nuevo sistema mucho mayor por 800.000 euros. Las ventas que la empresa estima que tendrá a lo largo de la vida útil de cualquiera de sus nuevos equipos es de 5.000.000 de euros en el caso de un mercado alcista, a lo que la empresa le asigna una probabilidad de que suceda del 30 %, en caso contrario, las ventas esperadas son de 3.500.000 euros. Construir el árbol de decisiones y decir que opción es la más ventajosa para la empresa.

   Solución:



   Evidente la opción más ventajosa es la de comprar 2 sistemas idénticos al actual, ya que cuando el mercado es alcista las ganancias son mayores que comprando un sistema mayo al actual y lo mismo ocurre cuando el mercado no es alcista.

---

5. ¿Que algoritmos de aprendizaje no se afectan por la dimensionalidad del vector de características? Diga cuáles y por qué.

   Solución:

   Aquellos algoritmos con mayor capacidad de generalización y regularización ya que estos pueden disminuir los efectos adversos que nos encontramos en alta dimensionalidad en los cuales es fácil caer en el sobreajuste. Por ejemplo SVM es un clasificador que funciona bien en altas dimensiones ya que puede presentar una gran regularización.

---

6. Considere la siguiente aproximación al aprendizaje. Mirando los datos, parece que los datos son linealmente separables, por tanto decidimos usar un simple perceptron y obtenemos un error de entrenamiento cero con los pesos óptimos encontrados. Ahora deseamos obtener algunas conclusiones sobre generalización, por tanto miramos el valor d vc de nuestro modelo y vemos que es d + 1. Usamos dicho valor de d_{vc} para obtener una cota del error de test.

   Argumente a favor o en contra de esta forma de proceder identificando los posible fallos si los hubiera y en su caso cual hubiera sido la forma correcta de actuación.

   Solución:

   Para empezar ha cometido el grave error de mirar los datos y dejarse llevar por lo que ha visto, además debido a ello ha decidido usar un perceptron dejando de lado otros modelos. Esto está provocando un sesgo en la información y debemos de evitarlo. También se ha dicho que hay un error de entrenamiento cero con los pesos óptimos encontrados, lo cual puede haber producido un sobre ajuste.

   Con todo lo que se ha hecho es muy posible que aunque el modelo sea perfecto dentro de los datos del entrenamiento, se tenga que se ajusta muy mal para los datos de fuera de la muestra.

Una forma correcta de actuar podría ser tener en cuenta varios modelos empezando con aquellos más simples como puede ser el modelo lineal. Hacer pruebas de validación separando en conjuntos de entrenamiento y test (por ejemplo validación cruzada) con lo que evitar el sobreajuste, y con ello hacer una estimación lo mas ajustada posible del Eout.

---

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.

   Solución:

   SVM: entre sus ventajas tenemos que encuentra el hiperplano de separación optimo, es efectivo en espacios de alta dimensionalidad ya que tiene una gran capacidad de generalización. En contra tiene que cuando el conjunto de dato es grande, el tiempo de entrenamiento requerido es muy elevado.

   Random forest: Una de sus mayores ventajas es la reducción de la varianza. Es un modelo de decisión muy preciso y es eficiente en grandes conjuntos de datos. Da una estimación de que variables son importantes en la clasificación. En contra tiene que son difíciles de interpretar, a diferencia de los árboles de decisión. Puede haber sobreajuste si los datos son ruidosos.

   No siempre SVM es mejor clasificador que Random forest, dependerá de cada problema cual de ellos será mejor para clasificación. Por ejemplo si hay muchas variables categóricas podría ser mejor elección Random forest en lugar de SVM. Además en casos reales si el conjunto de datos es demasiado grande SVM podría tardar demasiado en entrenar por lo que podría ser preferible Random forest. Ninguno de los dos es mejor que otro, dependerá de cada caso concreto.

---

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.

   La forma en la que construye muchos árboles de decisión y al final clasificar por 'mayoria simple'.

   La mejora que introduce es la reducción de la varianza.

---

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.

   Solución:

   El error principal de este experimento está en la forma de seleccionar la muestra, la cual se ha realizado echando una red para capturar peces. Esta manera de tomar la muestra hace que estemos excluyendo peces cuyo tamaño es inferior al de los huecos de dicha red (o disminuyendo la probabilidad de capturar peces de menor tamaño), provocando un error muestral en el experimento. Además habría que tener en cuenta también que diferentes zonas del lago pueden tener distintas distribuciones de tamaños y también por ejemplo la época del año podría infuir en el tamaño de los peces. Por todos estos motivos tenemos que no se están cumpliendo los objetivos que se persiguen.

---

10. Identifique dos razones de peso por las que el ajuste de un modelo de red neuronal a un conjunto de datos puede fallar o equivalentemente obtener resultados muy pobres. Justifique la importancia de las razones expuestas.

    Solución:

(a) Inicialización de los pesos. Si se inicializan todos los pesos a cero o al mismo valor, entonces no hay movimiento hacia el óptimo local, y si se inicializan a valores grandes se satura el sigmoide (ej: $\tanh(w^t x_n) \approx \pm 1$), entonces el gradiente será cercano a cero y el algoritmo no llegará a ninguna parte.

(b) Criterio de parada. Parar teniendo en cuenta solo el tamaño del gradiente puede ser una mala opción, ya que podríamos parar demasiado pronto, debido a que se ha llegado a una zona plana. Lo mejor es combinar varios criterios como número de iteraciones, valor de E_{in} y el tamaño del gradiente.