Universidad De Granada

Grado ingeniería informática (2017 – 2018)

Aprendizaje Automático

Trabajo 3: Cuestiones de Teoría

Trabajo realizado por Antonio Miguel Morillo Chica.

---

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 técncias. Diga cual de ellas considera que nos proporcionará una mejor estimación del error en cada caso concreto y por qué.**

Ambos son técnicas para la evaluación de modelos predictivos pero se diferencian en la forma en la que realizan las particiones. La validación cruzada realiza n particiones que incluyen n-1 pliegues para el entrenamiento y 1 para la evaluación, tipicamente se usa el error medio de 10 validaciones y se reentrena un único modelo con todos los datos. Sin embago bagging realiza n muestras con repetición de datos iniciales y he aquí la mayor diferencia entre ellos. En cada validación de VC sabemos que los ejemplos son distintos para cada entrenamiento, en el bagging no pero se estima que al mennos el 67% de los datos deberían de ser únicos.

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**

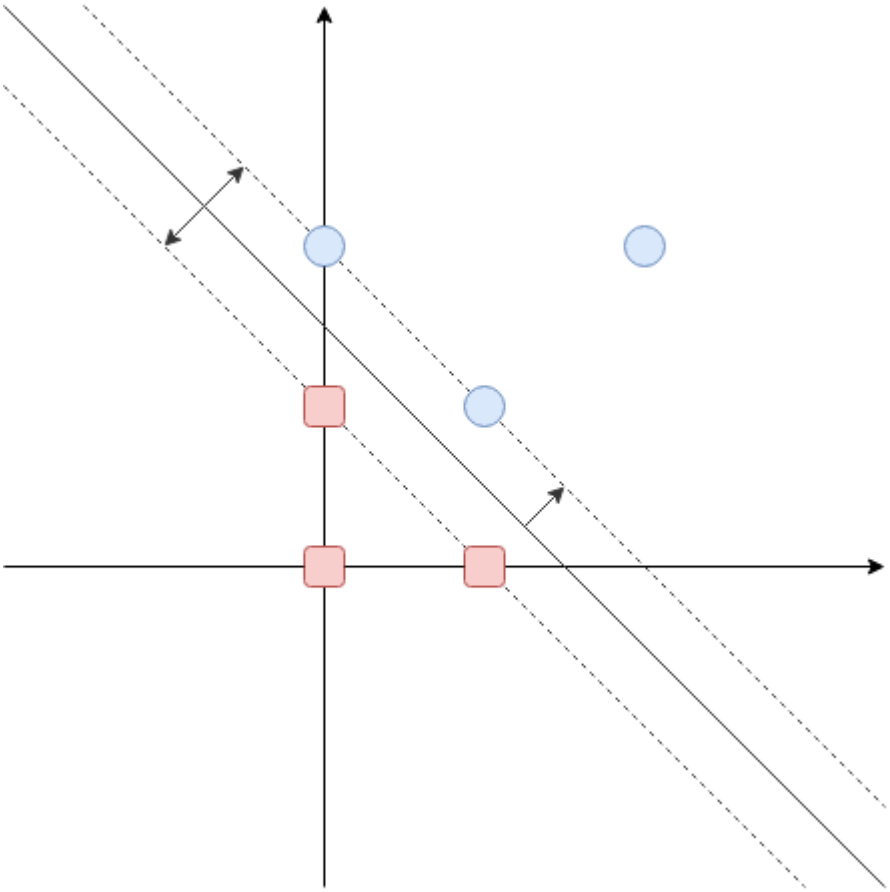
```
Entradas: (x i , y i ), i = 1, ... , n, w=0, k=0
repeat
    k ← (k + 1) mod n
    if sign(y i ) 6 = sign(w^T x_i) then
        w ← w + y_i x_i
    end if
until todos los puntos bien clasificados
```

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.

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.**

Trás el proceso que haga SVM la representación en la separación de los datos sería la sigiuente:



b) **¿Cuáles son los vectores soporte?**

Los vectores soporte serían (1,0) – (0,1) y el segundo, (2,0) – (1,1).

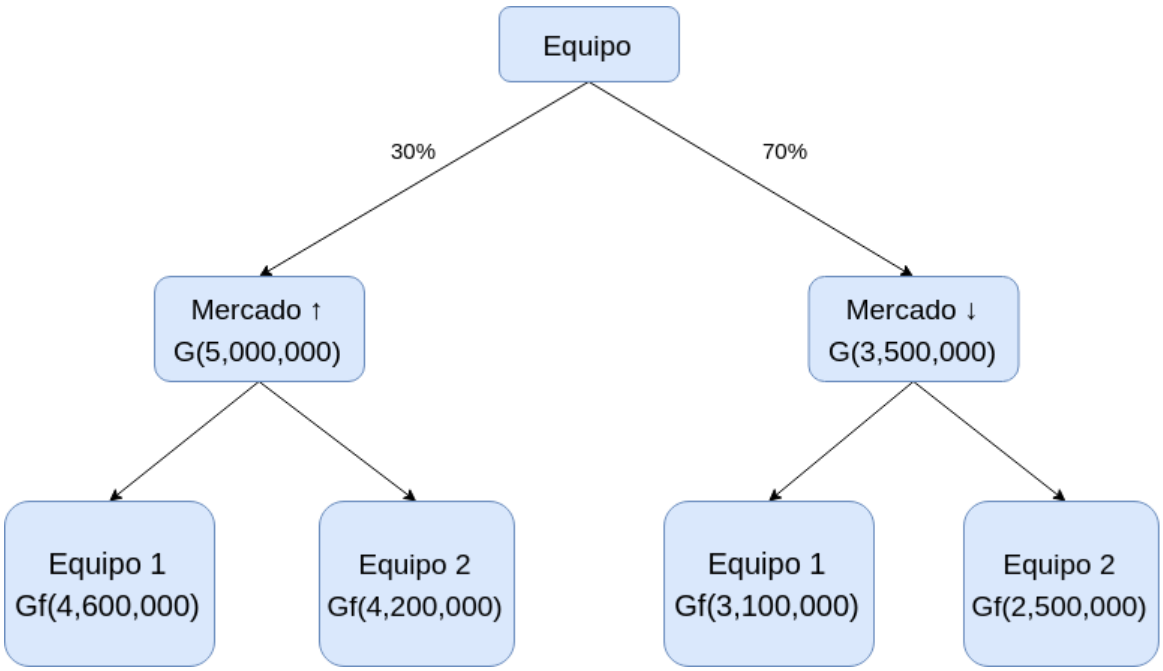
c) **Construir la solución en el espacio dual. Comparar la solución con la del apartado (a)**

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.**

Los datos aportados son los siguientes:

Equipo	Coste	Ganancia +	Ganancia -
1	400,000	5,000,000	3,500,000
2	800,000	5,000,000	3,500,000

El arbol de decisiń sería:



La mejor decisión, en cualquier escenario sería la primera opción, la que implica comprar dos equipos de 200.000 euros.

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

Como podemos ver en las transparencias de clase, Sesión4:

- Alta dimensionalidad: Mayor posibilidad de ser linealmente separable.
- Poca dimesionalidad: Posible conjunto que no es linealmente separable.

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. 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. Usamos dicho valor de d vc para obtener una cota del error de test.**

Obtener cero como error de entrenamiento no garantiza una buena generalización, dado que podría haberse sobreajustado sobre los datos de entrenamiento. También podría ser que el conjunto de datos sea pequeño, por lo que es fácil ajustarlo, pero si se añadiese una mayor cantidad de datos, podría cambiar completamente la clasificación. En mi opinión, podrían haberse implementado varios modelos y evaluarlos con validación cruzada para determinar cuál es el que mejor clasifica los datos garantizando una buena generalización.

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.**

- Support Vector Machine:
  - Pros:
    - Se realiza de forma similar a la regresión logística cuando la separación lineal
    - Se desempeña bien con un límite no lineal dependiendo del núcleo utilizado
    - Maneje bien los datos de alta dimensión
  - Contras:
    - Susceptible a sobreajuste / problemas de entrenamiento según kernel
- Random Forest:
  - Pros:

        - Es uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero.
        - Corre eficientemente en bases de datos grandes.
        - Puede manejar cientos de variables de entrada sin excluir ninguna.
        - Da estimados de qué variables son importantes en la clasificación.
        - Tiene un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.
        - Computa los prototipos que dan información sobre la relación entre las variables y la clasificación.
        - Computa las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos.
        - Ofrece un método experimental para detectar las interacciones de las variables.
    - Contras:
        - Se ha observado que Random forests sobreajusta en ciertos grupos de datos con tareas de clasificación/regresión ruidosas.
        - A diferencia de los árboles de decisión, la clasificación hecha por random forests es difícil de interpretar por el hombre.
        - Para los datos que incluyen variables categóricas con diferente número de niveles, el random forests se parcializa a favor de esos atributos con más niveles. Por consiguiente, la posición que marca la variable no es fiable para este tipo de datos. Métodos como las permutaciones parciales se han usado para resolver el problema.
        - Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes.

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.**

Bajo mi criterio lo que proporciona que aprendan de forma más eficiente es el hecho de que las decisiones son tomadas por las decisiones individuales de cada arbol que es contruido de forma aleatoria.

Las nuevas mejoras que introduce son la aplica la técnica general de agregación de bootstrap, o embolsado, a los aprendices de árboles. El empaquetamiento repetidamente (B veces) selecciona una muestra aleatoria con reemplazo del conjunto de entrenamiento y ajusta los árboles . Esto disminuye la varianza del modelo sin aumentar el sesgo, esto significa que las predicciones son altamente sensibles al ruido pero el promedio de muchos árboles no es siempre que no estén muy correlados.

(Falta parte de optimo)

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.**

Las conclusiones obtenidas no servirán si aunque la muestra sea grande esta no es representativa de la población. En este caso, la muestra no sería representativa dado que la posibilidad de coger peces más grandes es mayor que la de coger peces pequeños con la red, a no ser que la red sea muy fina. También podría darse que las zonas en las que se eche la red haya un mayor número de peces pequeños que de peces grandes o viceversa. Teniendo en cuenta estos posibles escenarios, la muestra no seguiría una distribución uniforme, por lo que no sería representativa sobre la población y las conclusiones obtenidas no servirían para determinar la distribución del tamaño de los peces. El método de selección de la muestra siempre debe asegurar que la muestra se elija de forma aleatoria y que sea uniformemente distribuida.

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.**

- Inicialización de los pesos:
  - No devemos inicializar los pesos al mismo valor o a cero sino acabaremos en un óptimo local, tampoco utilizar con valores muy altos ya que saturan la función sigmoidal el gradiante propagará ceros.
- Criterio de terminación:
  - Si cambiamos el criterio con uno menos “completo”, entendiendo por completo que el criterio de terminación venga determinado por más de un evento como: numero de iteraciones, tamaño del gradiante, error interno...