PRÁCTICA 2

Descripción

La base de datos, que se va a emplear en este trabajo, contiene mediciones hechas con un teléfono móvil inteligente. En concreto, el objetivo del problema es poder realizar el reconocimiento de la actividad que el sujeto está llevando a cabo. Esta base fue construida mediante las grabaciones recogidas de 30 sujetos con edades comprendidas entre los 19 y los 48 años en el marco de un estudio clínico. A los participantes se les solicito que realizaran actividades de la vida diaria mientras llevaban en la cintura mediante un cinturón un teléfono inteligente con sensores inerciales. El objetivo es clasificar las actividades en una de las seis actividades realizadas, es decir, WALKING, WALKINGUPSTAIRS, WALKINGDOWNSTAIRS, SITTING, STANDING, LAYING. Utilizando su acelerómetro y giroscopio integrados, se capturó la aceleración lineal en 3 ejes y la velocidad angular de esos mismos 3 ejes a una frecuencia constante de 50 Hz.

Las señales de los sensores (acelerómetro y giroscopio) se preprocesaron aplicando filtros de ruido y, a continuación, se muestrearon con la técnica de ventanas deslizantes de anchura fija de 2,56 segundos y con un solapamiento del 50% (128 lecturas/ventana). A partir de cada ventana, se obtuvo un vector de características calculando las variables tanto en el dominio del tiempo, como de la frecuencia. Cabe destacar que la señal de aceleración ha sido filtrada para eliminar la componente correspondiente a la gravedad y quedarse solo con el movimiento del cuerpo.

Atributos

El primero de los atributos, es el identificador del sujeto que realizó el experimento. A continuación, se tiene la aceleración triaxial del acelerómetro (aceleración total) y la aceleración corporal estimada así como la velocidad angular triaxial del giroscopio. Una serie de 550 medidas y medias tomadas en el dominio del tiempo y la frecuencia Etiqueta con de la actividad que se estaba realizando en ese momento. Resumen del problema:

Número de ejemplos	10299
Número de variables	561
Tipos de datos	Enteros, Reales y Categóricos
Nulos	Sí
Enlace para la descarga	Datos

Entrega

Para realizar la entrega de este trabajo se solicitarán dos elementos fundamentales:

1. Un notebook o un fichero Julia que se pueda ejecutar completamente con todas las aproximaciones.
2. Una memoria en la que se recojan los resultados de las aproximaciones así como los gráficos y las conclusiones que se quieran aportar sobre cada una de las aproximaciones realizadas.

Trabajo a realizar

Uno de los elementos que se debe de destacar es el hecho de que algunas instancias (ejemplos) cuentan con al menos uno o más de los valores continuos que no está especificado. Por ello se pide:

Preparación de los datos (20% de la calificación)

1. Cargar los datos del problema.
    - Realice una pequeña descripción de los mismo
        - número de variables
        - número de instancias
        - número de individuos
        - número de clases de salida
2. Calcule el porcentaje de nulos que hay por variable y en el conjunto total de sistema.
3. Preparar los datos para su uso en diferentes técnicas de clasificación. Es decir hacer las transformaciones necesarias en los datos y rellenar los datos faltantes cuando sea posible.
4. Haga un *holdout* del 10% de los datos estos se reservarán hasta el final, pero hágalo *individual-wise* es decir es el 10% de los individuos el que quedará fuera use el valor de semilla 172 para realizar la división. Compruebe que individuos quedan fuera.
5. Con los datos restantes, prepare un 5-Fold cross validation usando como semilla el valor 172. Tenga presente que la división debiera de hacerse teniendo en cuenta las instancias de un participante no estén en más de uno de los conjuntos.
6. Con cada uno de los conjuntos resultantes realice la normalización de los datos mediante MinMaxScaler

Creación de los modelos básicos (30% de la calificación)

7. A continuación, realice las siguientes reducciones de dimensionalidad
    - NO aplicar ningún tipo de reducción
    - Filtrado ANOVA
    - Filtrado Mutual Information
    - Filtrado RFE con el método de LogisticRegression con una eliminación del 50% de las variables en cada pasada.
8. Sobre el filtrado, aplicar alguna de las siguientes técnicas de reducción de la dimensionalidad:
    - NO aplicar nada

- PCA
- LDA
- ICA
- Isomap
- LLE

Para cada una de estas técnicas represente mediante las dos primeras características conservadas el conjunto de datos

9. Cree los siguientes clasificadores con los conjuntos resultantes del paso anterior.
   - MLP con al menos las siguientes arquitecturas: [50], [100] [100, 50]
   - KNN con valores de vecindario entre 1, 10 y 20
   - SVM con el parámetro C con valores 0.1, 0.5 y 1.0

Creación de los modelos de ensemble (30% de la calificación)

10. Adicionalmente, con los datos sólo con el tratamiento de Filtrado ANOVA, recrear las siguientes técnicas
    - BaggingClassifier con clasificador base KNN con número de vecinos 5 y número de estimadores 10 y 50
    - AdaBoosting con estimadores SVM con kernel lineal siendo el número de estimadores 5.
    - GBM (GradientBoostingClasifier), con 50 estimadores y un learning_rate de 0.2
11. Entrenar con el conjunto completo de entrenamiento (todo lo que componía el 5-fold cross-validation) y testear son el 10% reservado
    - Coger las 5 mejores combinaciones de los modelos anteriores de clasificación, (1 KNN, 1 SVM, 1 MLP, 1 Bagging y 1 AdaBoosting)
    - Crear un Random Forest con valor para los estimadores del 500 y profundidad máxima de 10
    - Crear un Hard Voting con las mejores combinaciones del KNN, SVM y MLP (uno para cada una de las técnicas)
    - Crear un Soft Voting con las mejores combinaciones del KNN, SVM y MLP (uno para cada una de las técnicas) para los pesos coger el porcentaje de acierto en test de cada una de las combinaciones en el 5-fold cross-valiadation
    - Crear un Ensemble Stacking con MLP como clasificador final, así mismo, use como base las mejores combinaciones del SVM, KNN y MLP
    - Crear un XGBoost con los valores por defecto

- Crear un LightGBM, con los valores por defecto
- Crear un Catboost, con los valores por defecto

Conclusiones (10% de la calificación)

12. Imprima la importancia de las variables seleccionadas por el Random Forest y por el XGBoost ordenadas por importancia.
13. Realice un contraste de hipótesis para determinar cuál de los modelos y extraiga las conclusiones.

Presentación (10% de la calificación)

Se tendrán en cuenta 3 elementos principalmente:

- Presentación, es decir que el notebook vaya acompañado de gráficos, que la redacción sea clara.
- Organización que el seguimiento de la memoria sea sencillo basado en las diferentes aproximaciones. Se espera que la memoria cuente con los siguientes puntos al menos:
  - Introducción, breve descripción de los datos sobre los que se efectuará el proyecto
  - Descripción de la preparación de los datos
  - Definir la estructura del código, es decir librería y descripción del pipeline de ejecución de las pruebas
  - Descripción de los resultados de las pruebas, se valorará el incluir tablas con los resultados, así como gráficos explicativos
  - Coclusiones que se extraen de las pruebas
- Conclusiones, que los argumentos esgrimidos tengan el sustento y las correspondientes referencias a los datos mostrados en tablas y gráficos.