Mentoría Data Science aplicado a BCI

Consignas TP3: Introducción al Aprendizaje Automático

En este Trabajo Práctico utilizaremos la totalidad de las señales del dataset y con todos sus canales para resolver los siguientes ejercicios.

Extracción de Características y Armado de Datasets

A) Introducción (opcional):
   a) Familiarizarse con la clase BCIDataset del notebook asociado al TP. Estudiar el código en detalle. Estudiar los atributos del objeto y su impacto. Estudiar los atributos generados internamente (“resultados”) en la clase en relación a lo charlado en la presentación del TP.
   b) Estudiar cómo varía el número de ejemplos en el dataset y la dimensión de cada dato según la variación de la ventana de tiempo seleccionada y el criterio de solapamiento.

B) Características Temporales:
   a) Usando BCIDataset junto con el extractor de features básico de fft (“naif_fft_features”), o el procesador de datos que desee, analice la influencia que tiene el tamaño de la ventana en el dominio de tiempo en la resolución en frecuencia del espectrograma de potencia.
   b) Teniendo en cuenta el inciso anterior, las frecuencias de estimulación y la frecuencia de muestreo pertinentes, ¿cuál considera que es el número adecuado de muestras temporales que puede recortar conservando la mayor cantidad de información útil en el dominio de la frecuencia? (t).
   c) En adición a la serie temporal cruda -”complete_examples_signal” de BCIDataset- (concatenada o no a lo largo de los canales, según su elección), defina una estrategia de extracción de atributos en el dominio de tiempo que opere sobre la serie cruda, ejemplo: algún criterio como la media en cada canal para el ejemplo. Sean creativos pero no dediquen mucho tiempo a este inciso, es más bien para tener un punto de comparación.
   d) Guarde los datasets generados de la forma que considere conveniente.
   e) Para cada dataset, analice la contribución de información de cada feature estudiado al propósito de clasificación. Estudie la correlación entre features. Estudie la correlación entre features y etiquetas. En conjunto con la exploración de los TP anteriores, ¿considera útil estos atributos?

C) Características Espectrales (en Frecuencia):
   a) Usando BCIDataset, con el extractor de features básico de fft, o el procesador de datos que desee, genere el dataset de ejemplos utilizado como atributos el espectrograma de potencia. Determine la estrategia que considere más pertinente, si concatenar los canales o trabajar los canales en forma individual.

b) En adición, a los vectores de atributos del inciso a), generar dos estrategias adicionales de extracción de features en el dominio de la frecuencia, al menos una de ellas tiene que implicar un número de atributos <= 8. Generar los correspondientes datasets, SEAN CREATIVOS. Si reducen la información a algunas frecuencias específicas, tengan en cuenta que la frecuencia de estimulación puede no encontrarse entre los residuos de frecuencia analizados, o aún si fuera así, que la respuesta al estímulo no sea exactamente de la misma frecuencia del estímulo en sí.
c) Guarde los datasets generados de la forma que considere conveniente.
d) Repita el inciso e) del apartado B). En el caso de los vectores de pocos features realice un pairplot para visualizar en baja dimensionalidad el problema de clasificación.
e) Analice la distribución del dataset, número de ejemplos por clase, balance entre clases, número de ejemplos “puros” y ejemplos “impuros”.

D) Particionado del dataset.
   a) A partir de los datasets generados en los apartados B) y C), definir un esquema de particionado (train, validation, test) apropiado para el entrenamiento de un algoritmo de clasificación. Tener en cuenta:
      i) Frecuencias de cada una de las clases y balanceo.
      ii) Inclusión de etiquetas puras e impuras.
      iii) Influencia del solapamiento en la independencia de los conjuntos.
      iv) Porcentaje de la base total que se considera en cada conjunto.
      v) Tener en cuenta la influencia de los diferentes pacientes y sesiones. ¿Qué estrategia prioriza? Utilizar todos los pacientes y validar/testear sobre conjuntos definidos teniendo en cuenta esa variable. O tener un conjunto de validación/testeo completamente independiente formado por otro paciente.
      vi) Fundamente todas las decisiones tomadas.

Aprendizaje automático

El objetivo de este trabajo práctico no es evaluar y comparar diferentes algoritmos de clasificación, sino entender el problema de clasificación en su conjunto y definir el esquema general para trabajar en él. En ese sentido:

A) Entendimiento del problema.
   a) Está claro que hasta aquí hablamos de un problema de clasificación supervisada, pero ¿a qué subclase dentro de ellos pertenece? (multilabel? multiclase?)
   b) En palabras, describa cómo podría plantearse un problema de regresión con los datos estudiados.
   c) De la misma forma, ¿cómo podría pensarse un problema de clasificación no supervisada a partir de los datos disponibles?

B) En el problema de clasificación supervisada:
   a) ¿Qué métricas considera que son apropiadas para evaluar el desempeño de algún algoritmo de clasificación?
   b) ¿Qué funciones de costo consideran apropiadas para entrenar un algoritmo para este problema?

C) Seleccione un algoritmo de aprendizaje supervisado estudiado en la Diplomatura.

a) Elija uno de los datasets generados previamente.
b) Determine un benchmark de performance usando una asignación aleatoria de etiquetas.
c) Entrene el algoritmo seleccionado, (no invertir mucho tiempo en el ajuste de hiperparámetros).
d) Genere curvas de la evolución de las métricas relevantes a lo largo del entrenamiento (desempeño y loss), tanto para el conjunto de entrenamiento como de validación. Analice la presencia de los fenómenos de overfitting, underfitting, y su relación con el bias y variance.
e) Sobre el conjunto de test, utilice las métricas definidas en el apartado B-a) para determinar el desempeño del algoritmo. Compárelo con el benchmark de desempeño.

D) Teniendo en cuenta lo observado en el apartado C), entrene el mismo algoritmo en todos los datasets generados previamente (5 en total). Use siempre el mismo esquema de particiones (los 5 datasets deberían ser diferentes features de los mismos ejemplos en el dominio del tiempo).
a) Compare el desempeño del algoritmo sobre el conjunto de test.
b) ¿Qué influencia tienen los features elegidos sobre el algoritmo utilizado?