Trabajo final: Cíclope

Marvik

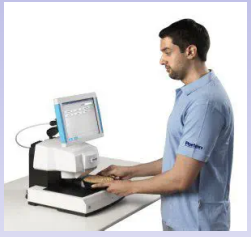
Curso Deep Learning en la práctica

Tabaré Pérez

I+D

C.I.T.M.P.S.A.

Introducción



El mito del Cíclope

En la mitología griega, los Cíclopes (en griego Kýklopes, que viene de “kyklos”, rueda, círculo y “ops”, ‘ojo’) eran los miembros de una raza de gigantes con un solo ojo en mitad de la frente.

El instrumento usado para relevar los espectros con los cuales desarrollaremos este trabajo, tiene “un sólo ojo” con el cual “observa” a las muestras en longitudes de onda del NIR (Near Infrared) entre 950nm y 1650nm. Por esta característica es que uso el nombre “Ciclope” en este trabajo.

Espectroscopía

¿Qué es la espectroscopía?

La espectroscopía agrupa a un conjunto de diferentes técnicas que usan radiación electromagnética para obtener datos de la estructura y propiedades de la materia que son usadas para la resolución de una amplia variedad de problemas analíticos.

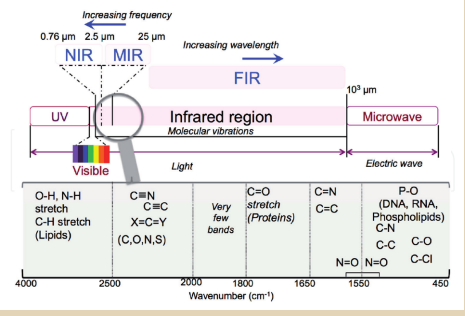
Espectro de luz visible: un ejemplo “Real”

Espectroscopía

¿Qué es la espectroscopía?

La espectroscopía agrupa a un conjunto de diferentes técnicas que usan radiación electromagnética para obtener datos de la estructura y propiedades de la materia que son usadas para la resolución de una amplia variedad de problemas analíticos.

Espectro de luz NIR - Near Infrared (Infrarrojo cercano)



Ley de Lambert y Beer: una versión simplificada

La intensidad de un haz de luz monocromática, que incide perpendicular sobre una muestra, decrece (exponencialmente) con la concentración de la muestra según esta ley:

$$A = k * C$$

En donde:

- A = Absorbancia
- k = Constante (incorpora características propias del analito)
- C = Concentración del analito

Ley de Lambert y Beer: una versión simplificada

Entonces podemos expresar que:

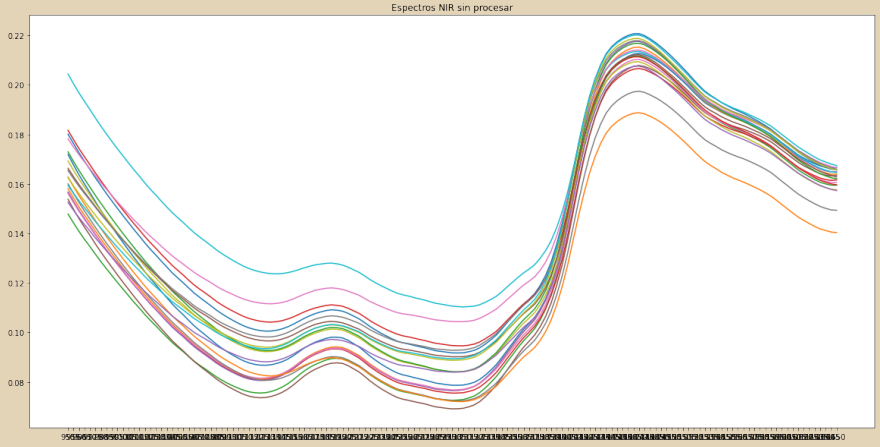
$$C = \sum_1^n k_i * A_i$$

- C = Estimación de la concentración del analito.
- n = Número de longitudes de onda del espectro NIR.
- k = Coeficiente i del modelo correspondiente a la longitud de onda i del espectro.
- A_i = Absorbancia en la longitud de onda i de espectro.

En los dos problemas de estimación (nicotina y azúcar), lo que hay que resolver es calcular los k_i de tal forma de obtener las concetraciones de los analitos de interés con un error apropiado para la tarea asignada.

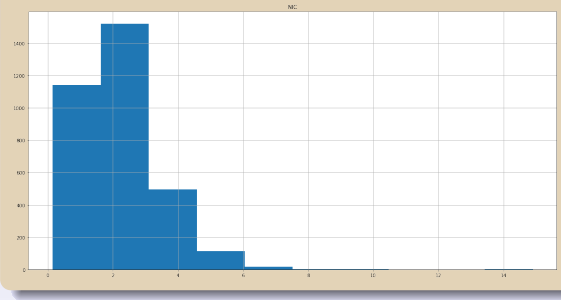
Los datos

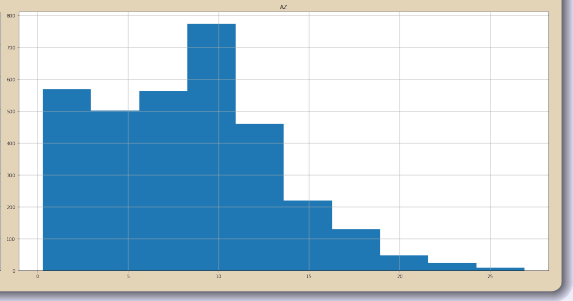
Espectros NIR de tabaco sin procesar:



Los datos

Distribuciones de nicotina y azúcar: valores de laboratorio





Un poco de historia

La “QUIMIOMETRÍA”:

- La quimiometría se desarrolló en la década de 1960. Extrae información de los sistemas químicos mediante el uso de métodos como la estadística multivariada, las matemáticas aplicadas y la informática, para abordar problemas de química, bioquímica, medicina, biología e ingeniería química.
- Svante Wold (Umeå Universitet, Suecia) inventó la palabra **“quimiometría”** para una solicitud de subvención a fines de 1971.
- En 1974, junto con Bruce Kowalski, Universidad de Washington, Seattle, EE. UU., crean la “Sociedad Internacional de Quimiometría (ICS)”. El primer artículo con la palabra quimiometría fue publicado por Wold en 1972. Sorprendentemente, solo se cita siete veces según “Web of Science”.

Un poco de historia

La “QUIMIOMETRÍA”:

- En la década de 1980, aparecen las primeras revistas dedicadas al tema como por ejemplo “Chemometrics and Intelligent Laboratory Systems” y “Journal of Chemometrics”, el primer libro con la palabra quimiometría en el título, varios simposios de ICS, la primera serie de libros sobre el tema (Research Studies Press), el primer software dedicado (ARTHUR, SIMCA, y UNSCRAMBLER), y los primeros talleres.
- El desarrollo de la quimiometría está fuertemente relacionado con la incorporación de instrumental con capacidad de digitalización de los resultados analíticos y de conectividad con computadoras.
- Actualmente, el uso de técnicas de “Machine Learning”, “Deep Learning” y “Artificial Inteligence” son cada vez más usadas en este campo que acompaña tanto el desarrollo de los nuevos métodos de cálculo como el aprovechamiento de hardware cada vez más poderoso para el tratamiento de grandes volúmenes de datos.

La propuesta

Objetivos del trabajo:

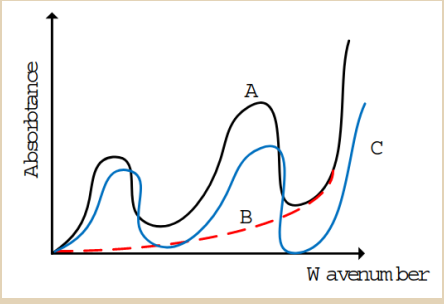
Usar los espectros NIR de muestras de diferentes tipos de tabaco obtenidos por un espectrofotómetro Perten 7250 para implementar técnicas de DL para la modelización, predicción y clasificación de dichas muestras. Se dispondrá de una base de datos de espectros con sus respectivas etiquetas correspondientes a los tipos de tabacos presentes en el set y concentraciones de nicotina y azúcar determinadas usando métodos analíticos de referencia y acreditados ISO17025.

Prueba de hipótesis:

Tanto en las estimaciones de nicotina y azúcar como en la clasificaión de tabacos, deseamos probar un posible pre-procesamiento de los espectros para mejorar los modelos: **"Corrección de la línea de base"**.

Pre-procesamiento

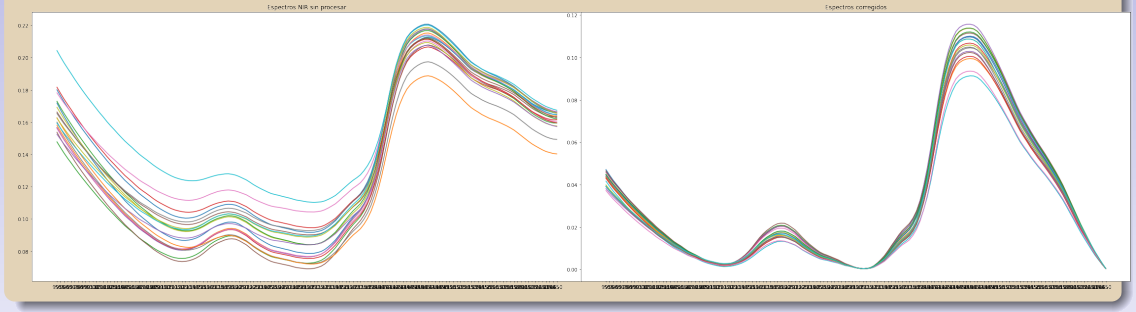
Corrección de la línea de base:



- **A** es el espectro con línea base desviada que se contradice con la condición de linealidad de la ley de Lambert - Beer.
- **B** es la línea de base estimada que,
- al restarla a **A** obtenemos **C** que es el espectro corregido que cumple con la condición de linealidad deseada.

Pre-procesamiento

Corrección de la línea de base:



Observación:

Es de destacar que, dada la naturaleza de los espectros y de la información contenida en los mismos, es importante conservar la forma o sea la relación de alturas entre todas las longitudes de onda. Cualquier procesamiento que se haga debe de respetar esto ya que el espectro de una sustancia es su huella digital. El procedimiento de corrección de línea de base respeta esta condición.

La propuesta

El set de datos:

- Para la clasificación: clasificacion-02.csv
  - Es es un archivo en el cual vamos a agregar una variable binaria que identificará si una muestra es de tabaco virginia o no. Vamos a usar todos los datos de tabaco virginia, burley, oriental y venas que disponemos.
    Tenemos en total 1140 registros:
    - 637 tabacos tipo VIRGINIA.
    - 503 tabacos tipo BURLEY, ORIENTAL y VENAS.
- Para la estimación de nicotina y azúcar: nic-az.csv
  - Para este problema, vamos a usar, como origen de datos, el archivo tabacos.csv que contiene 3294 muestras de tabaco con sus respectivos análisis de laboratorio de nicotina y azúcar. Una vez revisado, se genera el archivo de trabajo nic-az.csv.

Primer problema:Estimación de nicotina usando ANN densa

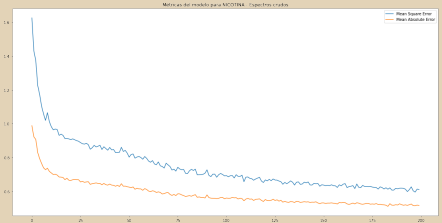
Arquitectura de la red:

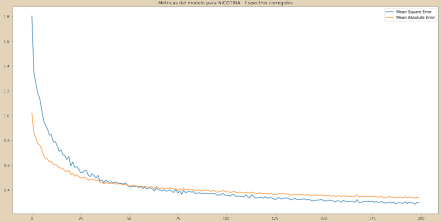
```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_4 (Dense)              (None, 64)                9088

 dense_5 (Dense)              (None, 32)                2080

 dense_6 (Dense)              (None, 16)                528

 dense_7 (Dense)              (None, 1)                 17

=================================================================
Total params: 11,713
Trainable params: 11,713
Non-trainable params: 0
_________________________________________________________________
```

- Learning rate = 0.01
- epochs: 200
- Activación: *relu*
- Optimizador: *RMSprop*

Primer problema:Estimación de nicotina usando ANN densa

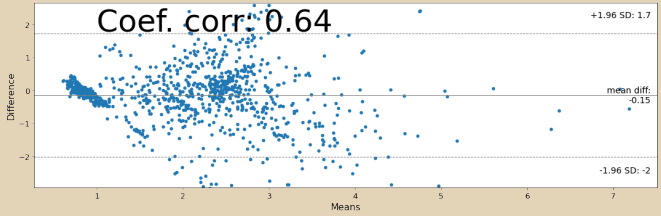
Evolución de las métricas: MSE y MAE



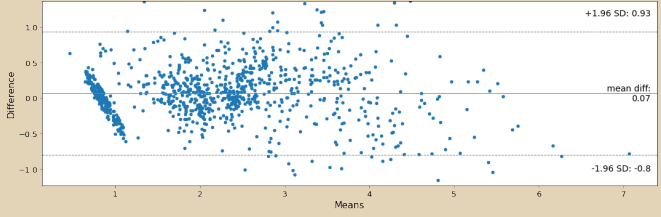


Primer problema:Estimación de nicotina usando ANN densa

Evolución de las métricas: Gráficos Bland - Altman



Coef. corr: 0.93



Segundo problema: Estimación de azúcar usando MLP Regressor

Arquitectura:

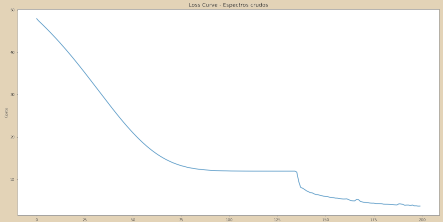
- MLPRegressor(hidden layer sizes=(141, 64, 32, 1)) con función de activación *relu* y *adam* para la optimización de los pesos.
- 200 iteraciones máximas.

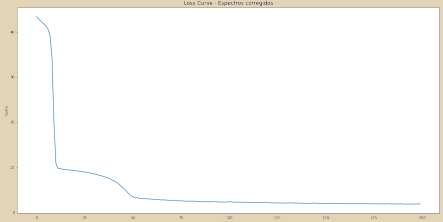
Evolución de las métricas:

ESPECTROS CRUDOS	**ESPECTROS CORREGIDOS**
MAE: 2.20	MAE: 1.40
MSE: 7.90	MSE: 3.51
RMSE: 2.81	RMSE: 1.90
R2: 0.68	R2: 0.86

Segundo problema: Estimación de azúcar usando MLP Regressor

Funciones de pérdida:





Tercer problema: Clasificación de tabacos

Clasificación de tabacos tipo Virginia

El problema que se plantea es hacer una clasificación binaria que permita diferenciar aquellos tabacos que son del tipo Virginia de aquellos que no lo son. Recordemos que el set de datos a usar contiene 1140 registros de los cuales 637 corresponden a tabacos tipo Virginia y 503 a tabacos tipo Burley, Oriental y venas (nervaduras de la hoja del tabaco).

Tercer problema: Clasificación de tabacos

Arquitectura:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 16)                2272

 dense_1 (Dense)             (None, 18)                306

 dropout (Dropout)           (None, 18)                0

 dense_2 (Dense)             (None, 20)                380

 dense_3 (Dense)             (None, 24)                504

 dense_4 (Dense)             (None, 1)                 25

=================================================================
Total params: 3,487
Trainable params: 3,487
Non-trainable params: 0
```

- Optimizador de los pesos: *Adam*
- Función de pérdida: *Binarycrossentropy* (es una clasificación)
- Métrica: *Accuracy*
- epochs: 50

Tercer problema: Clasificación de tabacos

Matrices de confusión: espectros crudos

	Test set		Full set	
	Recall: **0.83**		Recall: **0.84**	
0	162	17	568	68
1	27	136	80	423
	0	**1**	**0**	**1**

Tercer problema: Clasificación de tabacos

Matrices de confusión: espectros corregidos

	Test set		Full set	
	Recall: **0.95**		Recall: **0.94**	
0	165	14	603	33
1	8	155	28	475
	0	**1**	**0**	**1**

Conclusiones

Podemos decir que:

- Es alentador ver que con solamente corregir la línea de base del set de espectros podemos mejorar la predicción de los analitos de interés y la clasificación de los tabacos.
- Se necesita mucho más trabajo en acondicionar los sets de muestras (las distribuciones de las concentraciones de los analitos no son el todo adecuadas), el pre procesamiento de los espectros como se describió en la propuesta del trabajo (transformadas de Fourier, derivadas, PCA, algoritmos genéticos para la selección de las mejores longitudes de onda, etc) y en la arquitectura de los modelos.
- Los pasos a seguir son los de profundizar en estos temas y ampliar el alcance para poder estudiar si es viable la construcción de modelos de predicción más complejos que incluyan, por ejemplo, las características organolépticas de los diferentes tipos de tabaco y así poder predecir el comportamiento en boca y nariz del fumador o la confección de mapas de similitud que apoyen a la compra de tabacos para la producción de nuesros productos.

Stack de trabajo:

Configuración del laptop personal:

- Lenovo T490
- Distribución Linux Mint

Composición del texto del informe y la presentación asociada:

- Emacs: Herramienta de composición para la generación del informe final:
  - En modo orgmode para la composición del texto y exportación a LaTeX.
  - En modo LaTeX para los ajustes finos del texto y la generación del entregable en PDF.
  - Uso del Beamer LaTeX package, de la distribución TeXLive para la generación de las filminas de las presentación.

Stack de trabajo:

Limpieza de datos:

- Emacs: Limpieza de los archivos .csv.
- LibreOffice Calc: formateo de números y textos para la exportación del set limpio a .csv

Entorno de procesamiento:

- Google Colab

Repositorio de datos y programas:

- Github
- Cliente git de Linux para interactuar con el respositorio.

Stack de trabajo:

Librerías Python

- Keras
- Tensorflow
- Skitlearn
- Numpy
- Pandas
- Matplotlib
- StatsModels
- BaselineRemoval
- PyCompare
- Pingouin

Muchas gracias!