**1) daban un cuadrito con 4 frases y 3 decían la clase (A o B), pedían: probabilidad de clase A; prob de clase B; vocabulario, y prob de una palabra sabiendo que era de clase B. Hacerlo considerando smoothing laplace**

* Vocabulario

Chileno - Santiago - Valparaíso - Allende - Montevideo - Uruguay

| vocabulario | = 6

* Probabilidad de clase C?

“Cantidad de documentos de la clase C, sobre el total” “¿Que probabilidad tengo de que sea C (conteo)?”

4/5

* Probabilidad de clase U?

1/5

* Probabilidad de Chileno sabiendo que es de C

P(“Chileno” | C) = (8 + 1) / (13 + 6) = 0.47

* Probabilidad de Chileno sabiendo que es de U

P(“Chileno” | U) = (1 + 1) / (3 + 6) = 0.22

* Probabilidad de Uruguay sabiendo que es de U

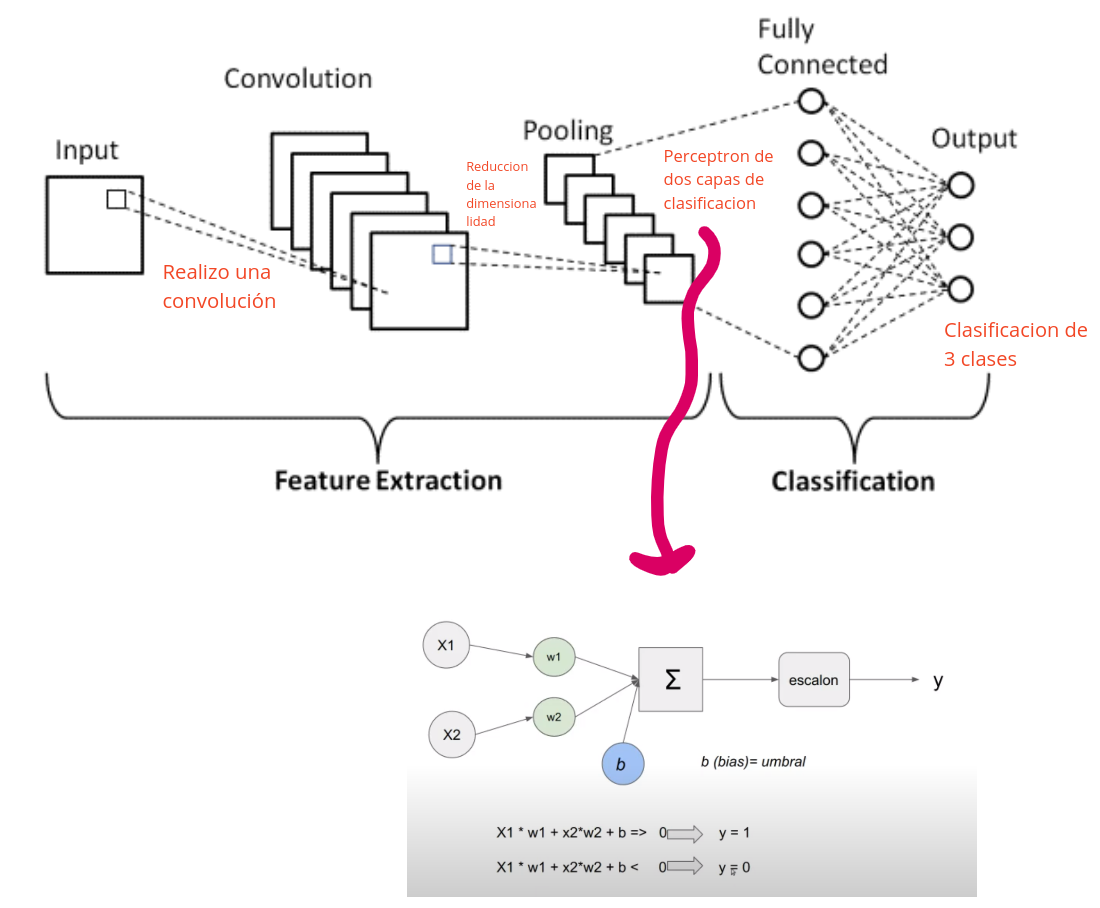
P(“Uruguay” | U) = (1 + 1) / (3 + 6) = 0.22

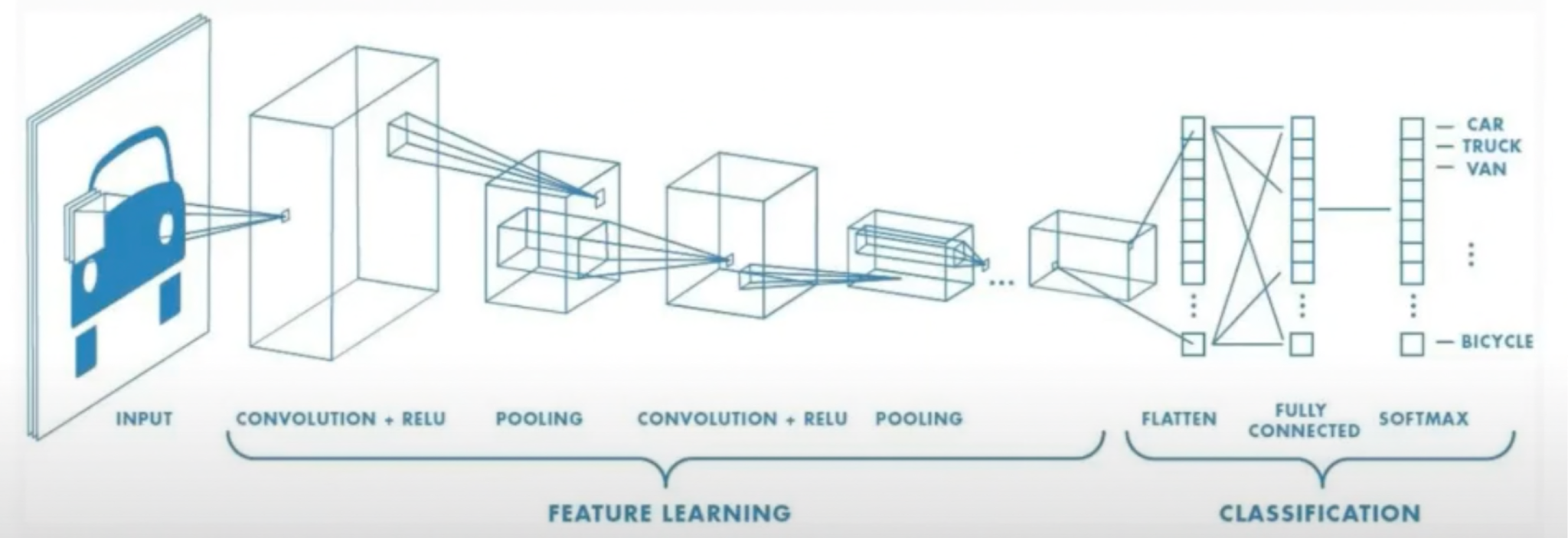
* Probabilidad de Uruguay sabiendo que es de C

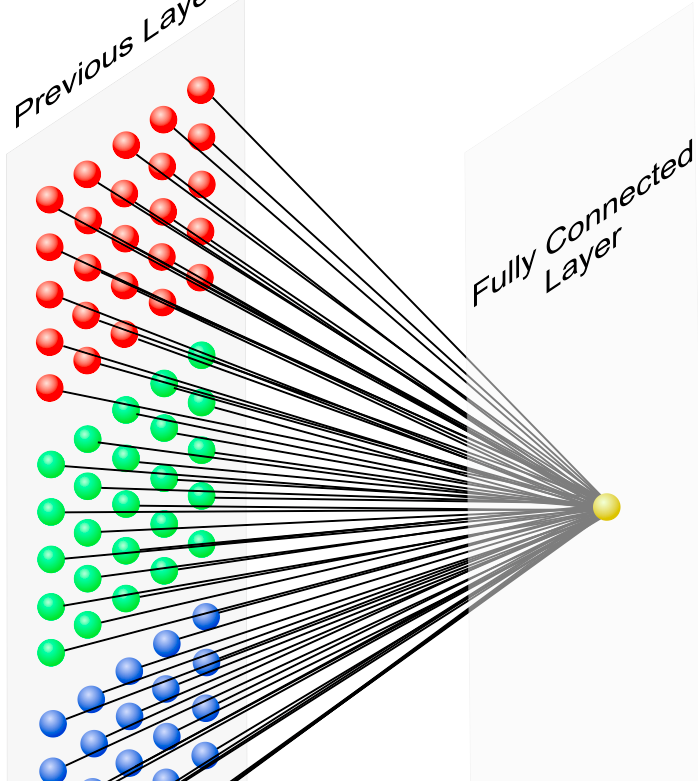
P(“Uruguay” | C) = (1 + 1) / (13 + 6) = 0.11

| 1 | Chileno Santiago Chileno | C |
| --- | --- | --- |
| 2 | Chileno Chileno Valparaiso | C |
| 3 | Chileno Allende | C |
| 4 | Montevideo Uruguay Chileno | U |
| 5 | Chileno Chileno Chileno Montevideo Uruguay | C |

**2) dibujar esquema sintético de una red convolucional, explicar las capas, el orden y funciones de activación**





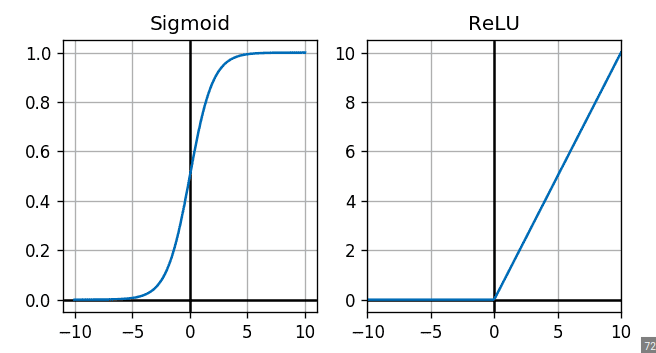


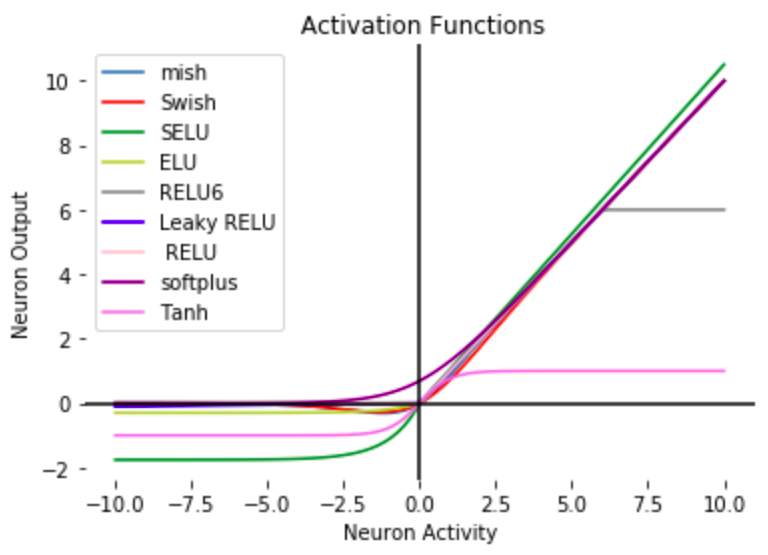
Capas:

1. Capa input: El dominio de entrada (imagen).
2. Capa Pooling: Reducción de la dimensionalidad. Hace uso de la función de activación.
3. Capa Fully Connected: Capa de clasificación.
4. Capa output: Resultado final del proceso de la red.

Funciones de Activación (más usadas):

* Relu: Se usa en capas ocultas para evitar el desvanecimiento del gradiente. Tiene una mejor performance en tiempo computacional.
* SoftMax: Suele usarse en la última capa.
* ...





**3) Del tp1 explicar qué técnica se usó para reducción de dimensión y comentar como fue la performance en el conjunto de test.**

¿Qué técnicas usamos?

* Eliminamos del dataset las variables irrelevantes.
* En caso de tener dos o más variables con un alto nivel de correlación entre si, eliminamos estas, ya que tenerlas a todas no nos va a aportar más información de la que conseguimos con una sola.
* Estudiamos también si nos son relevantes las variables con un valor bajo de varianza (no lo son).
* Estudiamos los distintos candidatos a outliers (univariados y multivariados), y en caso de determinar que lo sean procedemos a eliminarlos.

¿Cómo fue la performance en test?

Dado que el conjunto de técnicas realizadas durante el proceso de la reducción de la dimensionalidad fueron para obtener un dataset resultante más limpio en cuanto a sus datos, las métricas obtenidas con este, deberían ser más precisas que utilizando el dataset original.

**4) Explicar por qué es importante separar el train y test, que consideraciones hay que tener en el preprocesamiento y explicar una técnica usada en los tps.**

¿Por qué es importante separar en train y en test?

Necesitamos el conjunto de entrenamiento para que al momento de la predicción esta se haga con datos desconocidos, y el modelo a utilizar no haya tenido influencia en estos datos. A su vez, el dataset de train lo utilizamos para estudiar cómo vienen los datos (cuantos nulls tengo, cuáles variables no serán relevantes, estudio de outliers, trabajos de encoding, y demás acciones referidas a la preparación del dataset). El tener ambos conjuntos de datos, también nos permite la tarea de validar que tan bueno fue el método, utilizando las métricas correspondientes a ese modelo.

¿Qué consideraciones hay que tener en el preprocesamiento?

Dependiendo del problema a resolver y sus necesidades vamos a determinar que tipo de preprocesamiento realizaremos.

Para el TP1:

* Outliers
  + Univariados
  + Multivariados
* Valores nulls
* Valores faltantes
* Variables irrelevantes

Para el TP2:

* Construcción de la Bag of words
  + Detección de palabras más frecuentes
  + Utilizamos regex para la eliminación de etiquetas html, acentos, etc

Explicar una técnica usada en los tps.

Por ejemplo, para el análisis de valores atípicos del tp1, utilizamos las técnicas tanto de boxplot, z-score, z-score modificado, distancia de mahalanobis, dispersogramas y gráficos de densidad.

Para la imputación de datos faltantes, decidimos utilizar el método de hot deck, el cual reemplaza estos con valores obtenidos de registros que son los más similares. Una observación a tener en cuenta que no contemplamos en su momento, es que hubiésemos tenido resultados más acertados de haber utilizado la imputación del tipo cold deck, la cual selecciona valores o usa relaciones obtenidas de fuentes distintas de la base de datos actual.

En cuanto a las variables irrelevantes, eso lo pudimos determinar tanto por la varianza como por las que tienen una alta correlación entre sí. A su vez, realizamos un estudio en base al problema que debíamos resolver, y la información que brindaba cada una de las variables del input.

Pasando a la construcción de la Bag of Words, realizamos un filtrado por lo que nos parece más relevante a unificar o eliminar (etiquetas html, acentos, mayúsculas, etc). También estudiamos las variables que poseían una mayor frecuencia en el documento.

**5) Explicar técnicas usadas para ampliación del dataset en el tp2 y comentar la performance.**

Técnicas utilizadas

El dataset fue ampliado (con el anterior como base) utilizando el de property\_descriptions. A su vez, también fueron incluidas columnas por cada uno de los aspectos seleccionados previamente utilizando la técnica de Minqing Hiu y Bing Liu la cual considera las frecuencias de las palabras.

Performance

Puntualmente en nuestro trabajo, obtuvimos para el escenario sin optimización de hiperparametros un score de ~ 58%, mientras que luego de la optimización de estos, dicho score resultó en un ~ 57.9%. Más allá de los números, no esperábamos que luego de la optimización, el porcentaje sea menor, pero es un escenario que puede ocurrir porque RSCV no asegura entregar la mejor combinación de hiperparametros posible (solo asegura que sea la mejor dentro de su conjunto). Más allá de esto, para ambos casos tuvimos números bastante buenos (demuestran que tan bueno es el modelo).

OBS: Más allá de que con RSCV no se suela obtener el mejor resultado en cuanto al número obtenido (la cual si se garantiza con GSCV), se acerca lo suficiente. Cabe destacar que suele ser la técnica más utilizada dado a que termina siendo mucho más conveniente en cuanto a tiempos de ejecución.

**6) Que es un lexicón y como utilizarlo en análisis de sentimiento, que ventajas y desventajas tiene comparado con entrenar un modelo de cero.**

¿Qué es un lexicón?

Un lexicón es un diccionario que brinda la carga de valor de una palabra. Lo entrenamos cuando no queremos entrenar un algoritmo de clasificación (se usa para algoritmos supervisados).

También podemos crearnos uno propio. Para esto necesitamos un conjunto de ejemplos previamente clasificados (polaridad positiva/negativa) o bien reglas que identifiquen patrones en una frase.

¿Cómo utilizarlo en análisis de sentimiento?

El diccionario nos brinda la polaridad de una palabra, y en base a eso se puede determinar la polaridad que una frase/texto mediante distintos algoritmos que sirven para ponderar estos.

Algoritmos:

* Hatzivassiloglou y McKeown → Lo utilizamos para la ampliación de un lexicón mediante la utilización de las palabras “y” y “pero” para detectar la polaridad similar u opuesta de un par de palabras. Comete errores, con lo cual es necesario luego una supervisión manual.
* Turney → Nos ayuda para obtener la polaridad de una frase y, en definitiva, la ponderación general de un texto. Esto último podemos calcularlo haciendo el promedio de la polaridad de todas las frases.

Ventajas y desventajas contra entrenar un modelo de cero.

La principal ventaja que vamos a obtener si utilizamos un lexicón es en cuanto a lo temporal. Mientras que al entrenar un modelo desde cero vamos a tener que invertir recursos en su entrenamiento, el usar un lexicón nos ahorra esta tarea.

Por otro lado, en cuanto a las desventajas, podemos atribuirle la necesidad de revisión por parte de una persona, dado que sus clasificaciones podrían contener errores (por ejemplo cuando un texto posee sutilezas y expectativas frustradas).