Cuestionario NN-SVM, AFI 2022

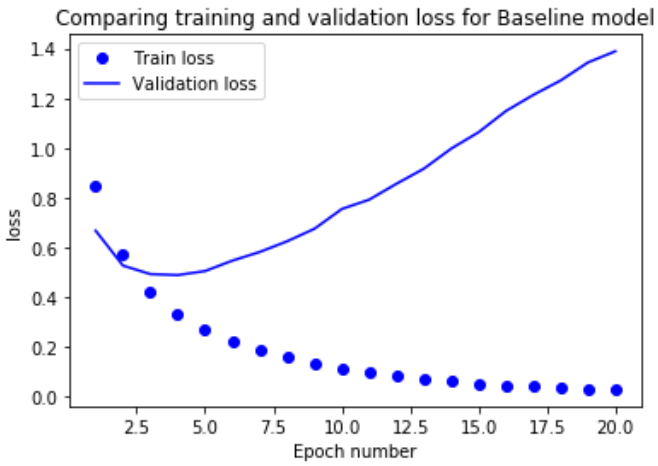
Nombre: **Javier** Apellidos: **Herráez Albarrán**

Cuestiones:

1. Estoy construyendo un modelo de regresión logística y me preocupa que esté haciendo overfit. ¿Cómo podría verlo en la función de coste?

A la hora de entrenar un modelo, éste puede tender a "aprenderse" demasiado bien los datos de entrenamiento, fallando a la hora de generalizar el problema, es decir, hace overfit. Así pues, lo que hará cualquier modelo será minimizar el valor de la función de coste lo máximo posible.

Una manera, por lo tanto, de detectar el overfitting será observar si este valor tiende extremadamente a 0. Además, lo ideal, sería utilizar un conjunto de validación para contrastar el valor de la función de coste entre el conjunto de entrenamiento y el de validación, para detener el entrenamiento en el momento en que empiezan a divergir.



2. Vamos a aplicar un modelo de 5 vecinos próximos en un problema de clasificación en 2 clases. Indicar los posibles valores que tomarán las probabilidades a posteriori de cada clase.

Planteando que tenemos un problema con dos clases identificadas con '+' y 'o', las probabilidades de que tome un valor '+' serán:

$$P(+|\,x\,) \simeq P\big(+|\,B_r(x)\,\big) = \frac{\pi_+ P(\,B_r(x)\,|\,+\,)}{P\big(B_r(x)\big)} \simeq \frac{\frac{N_+}{N}\,\frac{n_+(x)}{N_+}}{\frac{k}{N}}$$

Simplificando y sustituyendo para k = 5, tenemos que la probabilidad de tome la clase positiva será:

$$\frac{n_+(x)}{5}$$

siendo $n_+(x)$ el número de vecinos con clase '+' en el subconjunto de 5 elementos tomados para calcular la clase de x.

3. ¿Cómo podemos identificar visualmente las variables más relevantes en un problema de regresión? ¿Y en uno de clasificación?

Para un problema de regresión podemos realizar diagramas de dispersión con la variable objetivo en un eje y la variable sobre la que queremos intuir su relevancia en el otro eje. Además, podemos utilizar matrices de correlación para distinguir si nuestra variable objetivo esta correlada con las distintas features.

Para problemas de clasificación podemos utilizar gráficos de densidad o histogramas distinguiendo entre las distintas clases (por ejemplo, con colores) para cada variable que queramos estudiar.

4. ¿Qué es la maldición de la dimensionalidad? ¿Qué modelos de los estudiados en el curso se van a ver más o menos afectados por ella?

La maldición de la dimensionalidad, de forma simplificada, se podría definir como una serie de problemas que emergen cuando un conjunto de datos tiene gran dimensionalidad, es decir, cuando consta de muchas variables, lo cual causa gran dispersión de los datos. Algunos de los problemas que puede acarrear un conjunto de gran dimensionalidad son:

- Coste computacional, debido a trabajo que lleva procesar tal cantidad de variables.
- Insignificancia de correlaciones, al tener tal cantidad de variables, se podrán ver patrones en variables que estadísticamente son significativos, pero que en realidad son irrelevantes.
- Y, sobre todo, que las distancias dejan de ser significativas. En un espacio de baja dimensionalidad las distancias euclidianas pueden definir conjuntos, o diferencias entre muestras. Sin embargo, en un espacio de grandes dimensiones cada muestra parecerá ser única. Por lo tanto, y contestando a la siguiente pregunta, los modelos que se verán más afectados por ella serán aquellos que utilicen distancias euclídeas.

5. Tengo un problema de regresión con muchas features por lo que creo que un modelo SVR lineal me puede dar buenos resultados. ¿Qué hiperparámetros debo optimizar? Me da la impresión de que el modelo va a tener la pega de ser homogéneo. ¿Qué puedo hacer para mitigar esto y tener finalmente un buen modelo?

El peligro siempre presente con los datos de alta dimensión es el sobreajuste. Hay dos soluciones genéricas a este problema: la reducción de la dimensionalidad y la regularización. Obviando la reducción de dimensionalidad, lo ideal para nuestro modelo sería optimizar un hiperparámetro de regularización, que para SVM es el hiperparámetro **C**.

Para modelos que tengan problemas de homogeneidad una de las soluciones puede ser utilizar el algoritmo Sequential Minimal Optimization (SMO).

6. Nos hemos despistado en un problema de clasificación y hemos acabado construyendo un modelo de regresión logística constante (esto es, en vez de w x + w_0, el modelo solo usa w_0). ¿Cuánto valdrá w_0? ¿Qué predecirá el modelo?

(Sugerencia: hacer w = 0 en la fórmula de la log likelihood en las transparencias de la regresión logística y obtener el w_0 que la maximiza.)

Teniendo la fórmula de log-likelihood:

$$\ell(w_0, w) = \sum_p y^p (w_0 + w \cdot x^p) - \sum_p \log\left(1 + e^{w_0 + w \cdot x^p}\right)$$

Igualando $w = 0$:

$$\ell(w_0) = \sum_p y^p w_0 - \sum_p \log(1 + e^{w_0})$$

Para maximizar la función, derivamos e igualamos a 0:

$$l'(w_0) = \sum_p y^p - \sum_p \frac{e^{w_0}}{1 + e^{w_0}} = 0$$

Despejando w_0:

$$w_0 = \ln \frac{\sum_p y^p}{1 - \sum_p y^p}$$

Teniendo en cuenta la función logit:

$$logit(p) = \ln \frac{p}{1 - p}$$

Por lo tanto, tenemos que la probabilidad a predecir por el modelo teniendo en cuenta el espacio muestral será $\frac{1}{N}\sum_p y^p$, es decir, será la proporción de la muestra de la clase positiva.

7. Estamos trabajando con un modelo SVR gaussiano para una muestra de dimensión 2 pero no nos hemos dado cuenta de que las dos features son exactamente iguales. ¿Cómo afectará esto al modelo? ¿Podemos controlar su efecto? ¿Cómo?

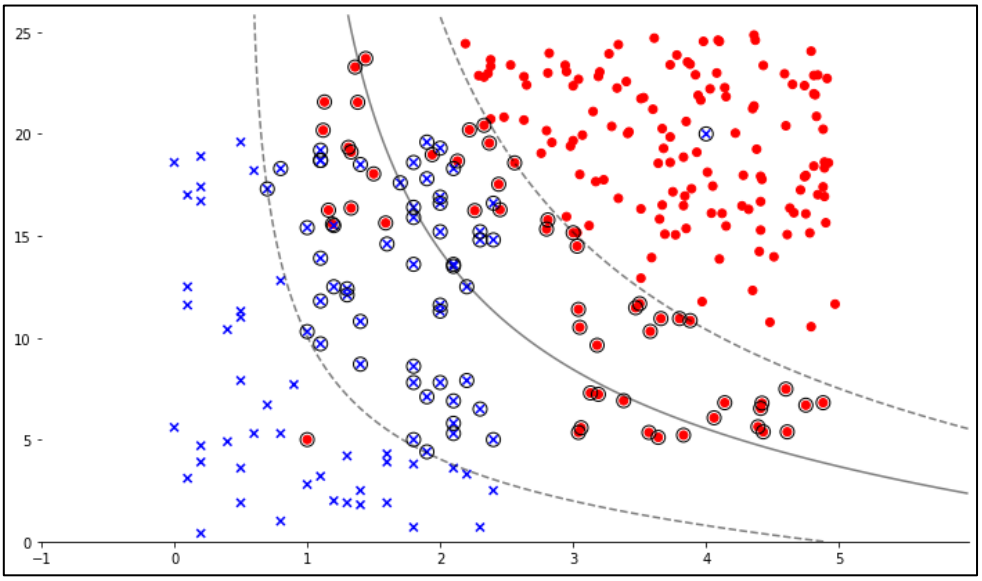
Para nuestro caso, se tendrán en cuenta las distancias entre los elementos de la muestra para entrenar un modelo. De modo que, al tener esta *feature* repetida, ésta tendrá mayor peso en las distancias que cualquier otro atributo. Por lo tanto, posiblemente, el modelo se verá más afectado por esta *feature* que por el resto.

Es decir, si las dimensiones están correlacionadas entre sí, la distancia euclidiana puede dar poca información o una información engañosa sobre lo cerca que está realmente un punto. Así pues, podemos usar otro tipo de métrica que nos ayude a lidiar con la correlación, por ejemplo, la distancia de Mahalanobis que sí que la tiene en cuenta.

8. Hemos entrenado una SVM en un problema de clasificación con una constante C=10 sobre una muestra de 1000 patrones y hemos obtenido 200 vectores soporte, de los cuáles 100 tienen coeficientes alfa con el valor C=10. ¿Cuál sería la accuracy mínima de dicho modelo?

El modelo utiliza 200 vectores soporte, sin embargo, los valores de alfa para 100 de ellos nos indican que éstos últimos estarían mal clasificados. Para el resto de los patrones, que no acaben siendo vectores soporte implica que estarán correctamente

clasificados. Esto significaría que al menos tendríamos un 10% de accuracy debido a los vectores soporte bien clasificados, además del 80% derivado de los que no son vectores soporte. Por lo tanto, acabaríamos teniendo un accuracy del 90% en el conjunto de entrenamiento.



Ejemplo del funcionamiento de un modelo SVM con kernel polinómico con 2 clase donde las vemos diferenciadas con colores y los vectores soporte dibujados con borde negro.

9. Si a continuación aplicamos el modelo a los 800 patrones que no son vectores soporte, ¿cuál sería la accuracy del modelo sobre los mismos?

Como hemos dicho anteriormente, que los patrones no hayan sido considerados vectores soporte implica que su clasificación resulta obvia, por lo que podemos asumir que el accuracy será del 100%.

10. Si en un problema dado hacer el fit de un MLP de regresión tarda 10 segundos, ¿qué tiempo de ejecución debo esperar si mantengo la arquitectura MLP, pero triplico el tamaño de la muestra? ¿Qué tiempo debería esperar como mínimo si se tratara de un modelo SVR?

El coste de entrenar un MLP manteniendo la misma arquitectura dependerá linealmente del tamaño de la muestra. Entonces, con un tamaño 3 veces mayor que la muestra original, el coste de entrenamiento será también 3 veces mayor, es decir, para nuestro caso serán 30 segundos.

Por otra parte, el coste computacional de una SVM es, al menos, cuadrático respecto al tamaño de la muestra. Por lo tanto, si triplicásemos el tamaño de la muestra de una SVM que tarda 10 segundos en entrenar, tendríamos que el nuevo modelo tardaría como mínimo 90 segundos.