Actividad: Clasificación con máquina de vectores de soporte y redes de neuronas

**Objetivos**

Mediante esta actividad se pretende que pongas en práctica la creación de modelos basados en máquinas de vector de soporte y redes de neuronas. El objetivo es que comprendas de forma práctica con un problema determinado las diferencias que existen a la hora de entrenar los diferentes modelos.

**Descripción de la actividad**

Importa los datos de Mobile Price Classification:

<https://www.kaggle.com/iabhishekofficial/mobile-price-classification#train.csv>

La página web contiene la información necesaria sobre los datos. La variable respuesta es «price\_range».

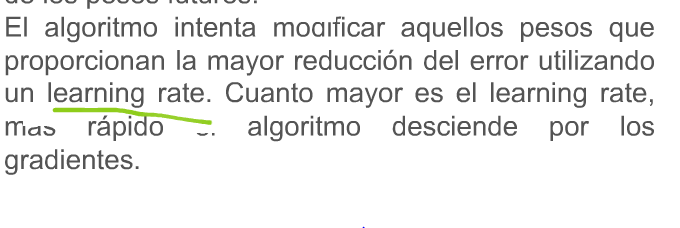
Crea modelos basados en máquinas de vector de soporte y redes de neuronas.

* Análisis descriptivo de los datos:
  + De las variables numéricas hallar el valor mínimo, el máximo, la mediana y la media.
  + De las variables categóricas, listar las diferentes categorías y hallar la frecuencia de cada una de ellas.
  + Hallar todas las correlaciones existentes entre las variables numéricas del conjunto de datos.
* Determinar el conjunto de modelización y el de validación.
* Calcular las métricas de evaluación de ajuste adecuadas:
  + Realizar una validación cruzada con mínimo 5 folds y calcular la matriz de confusión.
* Comentar los resultados obtenidos en la matriz de confusión de cada uno de los modelos. ¿Cuáles son las ventajas y las desventajas de cada uno de los modelos? ¿Cuál es el mejor modelo? Justifique su respuesta.
* Otros comentarios que parezcan adecuados.
* Se puede usar R o Python.
* Se deben comentar los resultados obtenidos y el código.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Actividad | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Metodología | 6 | 60% |
| Criterio 2 | Resultados | 2 | 20% |
| Criterio 3 | Informe | 2 | 20% |
|  |  | **10** | **100 %** |

**Extensión** **máxima:** 5 páginas en formato Word o PDF con el código adjunto aparte.



Se verifica que no existen datos faltantes para los tipos numéricos y no numéricos. Por tanto, no es necesario realizar ningún tratamiento.

Resultados matrices de confusión:

Para las funciones de clasificación la matriz de confusión, cuyos valores son métricas de las estimaciones, ambas tienen el valor de ratio de éxito mayor y el valor f1. Sin embargo, para el modelo de redes neuronales durante el proceso de validación cruzada se puede observar un comportamiento atípico frente a otros modelos de clasificación, ya que los valores de ratio de éxito ya que no lograron aprender tanto.

Por tanto, es el mejor método para clasificar en este conjunto de datos es el basado en SVM ya que en el proceso de validación cruzada como en el entrenamiento sobre los conjuntos de entrenamiento con el 80% y 20 de test respectivamente los valores de ratio de éxito se mantienen sobre el 85%.

Desventajas

There is no specific rule for determining the structure of artificial neural networks. The appropriate network structure is achieved through experience and trial and error.

No hay una regla para determinar una mayor estructura para resolver un problema. La estimación de una estructura adecuada se la consigue generalmente a través de métodos prueba y erros y se basan en la experiencia de los datos.

<https://intellipaat.com/community/21886/advantages-and-disadvantages-of-neural-networks>

<https://www.springboard.com/blog/beginners-guide-neural-network-in-python-scikit-learn-0-18/>

<https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6>

Ventajas

Son tolerantes a fallos ya que la corrupción de una de sus neuronas (en caso de tener una gran cantidad de ellas), permite que se siga realizando el proceso de estimación.

Corruption of one or more cells of ANN does not prevent it from generating output. This feature makes the networks fault-tolerant.

Artificial neural networks learn events and make decisions by commenting on similar events.

Las redes neuronales aprenden a través de eventos y realizan decisiones relacionándolos.

Here are some advantages of [Artificial Neural Networks](https://intellipaat.com/blog/tutorial/artificial-intelligence-tutorial/artificial-neural-networks/" \t "_blank) ( ANN)

* Storing information on the entire network: Information such as in traditional programming is stored on the entire network, not on a database. The disappearance of a few pieces of information in one place does not restrict the network from functioning.
* The ability to work with inadequate knowledge: After ANN training, the data may produce output even with incomplete information. The lack of performance here depends on the importance of the missing information.
* It has fault tolerance:  Corruption of one or more cells of ANN does not prevent it from generating output. This feature makes the networks fault-tolerant.
* Having a distributed memory: For ANN to be able to learn, it is necessary to determine the examples and to teach the network according to the desired output by showing these examples to the network. The network's progress is directly proportional to the selected instances, and if the event can not be shown to the network in all its aspects, the network can produce incorrect output
* Gradual corruption:  A network slows over time and undergoes relative degradation. The network problem does not immediately corrode.
* Ability to train machine: Artificial neural networks learn events and make decisions by commenting on similar events.
* Parallel processing ability:  Artificial neural networks have numerical strength that can perform more than one job at the same time.

Disadvantages of Artificial Neural Networks (ANN)

* Hardware dependence:  Artificial neural networks require processors with parallel processing power, by their structure. For this reason, the realization of the equipment is dependent.
* Unexplained functioning of the network: This is the most important problem of ANN. When ANN gives a probing solution, it does not give a clue as to why and how. This reduces trust in the network.
* Assurance of proper network structure:  There is no specific rule for determining the structure of artificial neural networks. The appropriate network structure is achieved through experience and trial and error.
* The difficulty of showing the problem to the network:  ANNs can work with numerical information. Problems have to be translated into numerical values before being introduced to ANN. The display mechanism to be determined here will directly influence the performance of the network. This depends on the user's ability.
* The duration of the network is unknown: The network is reduced to a certain value of the error on the sample means that the training has been completed. This value does not give us optimum results.

If you wish to learn more about AI, visit [Artificial Intelligence tutorial](https://intellipaat.com/blog/tutorial/artificial-intelligence-tutorial/" \t "_blank) and [Artificial Intelligence course](https://intellipaat.com/artificial-intelligence-deep-learning-course-with-tensorflow/" \t "_blank) by Intellipaat.

1. SVM works relatively well when there is clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient

Funcionan bien cuando existe una clara alta de-correlación entre las clases. Por ejemplo, con data no estructura o semi-estructurada por ejemplo textos, imagines entre otros.

Es más efectiva cuando existen más variables que datos. Adicionalmente, Corre menos riesgo de sobre estancación

Es eficiente respecto del consumo de memoria. Generalmente comparada con redes de neuronas ofrece mejores resultados.

models have generalization in practice, the risk of over-fitting is less in SVM.

ith even unstructured and semi structured data like text, Images and trees.

<https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/>

<https://medium.com/@dhiraj8899/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107>

- Realizar un análisis descriptivo de los datos tanto de variables numéricas(media, cuartiles,...) como de variables categóricas (cuántas clases existen en cada variable y la frecuencia de cada una de ellas).

- Hallar las correlaciones entre las variables e investigar que variables están altamente correlacionadas.

- Revisar si existen valores nulos y trabajar con esos datos. Eliminar columnas si existen muchos valores nulos, llenar valores nulos con media, moda, ceros, depende del tipo de datos de cada columna.

- Determinar el conjunto de datos de entrenamiento y el de validación.

- Aplicar los algoritmos de máquina de vectores de soporte y de redes de neuronas al conjunto de datos.

- Aplicar validación cruzada y calcular la matriz de confusión.

- Comentar los resultados obtenidos con cada uno de los clasificadores.