- CLASIFICACIÓN NO CONVENCIONAL: MULTILABEL - Máster Inteligencia Computacional e Internet de las Cosas

Entregable I – Multilabel Datasets:

DO NOT FORGET:

Make a small report with the answers to the exercises of this practice to later generate a PDF document that will be the "deliverable" of this practice.

Take care of the cover page, table of contents, content, style, format, bibliographical references, structure, etc. of the deliverables.

Not all the points of the practice involve including something in the deliverable. Only those points of the exercise in which it is explicitly indicated, and all those in which you are asked for the code or instructions necessary to carry out an exercise will be included in the deliverable, in which case you will indicate the code and the explanation that you consider necessary.

In the deliverable, always indicate the exercise number and its statement.

Free text answers to exercise questions, or additional explanations to exercises, should be no more than 4-5 paragraphs maximum, of about 30-40 words each.

Always write the commands you use to do each exercise in the deliverable in the corresponding section.

EJERCICIOS:

1. Realiza el tutorial colgado en: http://scikit.ml/tutorial.html, comprueba que no tienes problemas de configuración y que comprendes lo realizado en cada paso del mismo.
2. Familiarízate con los datasets presentes en Scikit-multilearn y con los que trabajaremos durante las próximas sesiones. ¿Qué información representan?. Concretamente, dispone de los siguientes 17 datasets: {'scene', 'Corel5k', 'bibtex', 'enron', 'rcv1subset5', 'tmc2007_500', 'rcv1subset3', 'rcv1subset1', 'delicious', 'rcv1subset4', 'genbase', 'birds', 'emotions', 'rcv1subset2', 'mediamill', 'medical', 'yeast'}.
3. Escribe un script en Python (car.py) que calcule para cada uno de los datasets de los dos ejercicios anteriores las siguientes medidas de caracterización de cada dataset (como mínimo los estadísticos vistos en teoría) y que las muestra en pantalla de forma ordenada:
    a. number of instances (n)
    b. number of attributes (f)
    c. number of labels (l)
    d. cardinality (car)
    e. density (den)
    f. diversity (div, represents the percentage of labelsets present in the dataset divided by the number of possible labelsets)
    g. average Imbalance Ratio per label (avgIR, measures the average degree of imbalance of all labels, the greater avgIR, the greater the imbalance of the dataset)
    h. ratio of unconditionally dependent label pairs by chi-square test (rDep, measures the proportion of pairs of labels that are dependent at 99% confidence)
4. Busca en Internet dos datasets extra. Descárgalos localmente y realiza un script en Python para cargarlos y recalcular todas las medidas del paso anterior sobre los mismos. Debes escoger datasets variados, con diferente número de etiquetas, variables e instancias.

5. Familiarízate con la documentación scikit-multilearn en: http://scikit.ml/ Prueba los métodos ML disponibles pertenecientes a las dos categorías (transformación y adaptación) que hemos visto en teoría.

6. Repasa lo que ya estudiado anteriormente sobre validación cruzada, investiga las funciones: KFold() y cross_validate(), cross_val_score y make_scorer(). ¿Cómo aplicarías validación cruzada en el contexto ML? Pega a continuación el código de un ejemplo de uso.
7. Describe brevemente las métricas que utilizaremos: *Accuracy, Hamming loss, Precision, Recall y F1_score.* a información que aporta cada una de estas métricas. Investiga su uso estudiando classification metrics y classification_report()
8. Escribe un script en Python (*cl-cv.py*) seleccionando al menos 3 métodos (que NO pertenezcan todos a la misma categoría) para evaluarlos con 5 de los datasets que recopilaste anteriormente y calcula las métricas resultantes mediante validación cruzada. El script mostrará el resultado de las métricas anteriores.

9. Responde a las siguientes preguntas:
    a. ¿Qué método(s) parece(n) comportarse mejor globalmente? Entre otros aspectos, analiza el efecto que tiene en el rendimiento de los métodos de transformación la elección que hagamos del clasificador base.
    b. ¿Cuál es el tiempo de ejecución de cada método? ¿Existen diferencias destacables?
    c. ¿Observas diferencias reseñables en cuanto a los valores de cada métrica?
    d. Analiza si observas alguna relación entre los resultados obtenidos para las métricas de evaluación y las características propias del conjunto ML: número de instancias, número de etiquetas, densidad, cardinalidad...? Etc. Escribe un breve informe sobre si los estadísticos calculados en el entregable1 dan información del rendimiento que estás observando. ¿Se pueden extraer conclusiones al respecto? Puedes buscar información en referencias bibliográficas previas.
    e. Realiza un script para visualizar gráficamente (mediante matplotlib) el comportamiento de los métodos.