Pontificia Universidad Católica de Chile

Ingeniería

IIC1005 - Exploratorio Computaci

# Informe Tarea 5

En esta tarea se buscaba utilizar la librería “scikit-learn” de python para realizar minería de datos. Para esto utilizamos distintos features que están clasificados en tres categorías, la categoría 1 donde están los features e_cosine_similarity_categories, e_total, event_regions_common y event_regions_total, la categoría 2 con m_cosine_sim_categories, m_cosine_sim_prices y m_cosine_sim_ratings y por último la categoría 3 con sb_aa_neigh, sb_neigh_common y sb_neigh_total.
Tras haber trabajado con cada categoría se obtuvo los siguientes resultados:
-Resultado de: c1. Eventos y Regiones de usuarios de SL:
LogisticRegression score: 0.527778
-Resultado de: c2. Características de los productos
LogisticRegression score: 0.731111:
-Resultado de: c3. Red Seller – Buyer
LogisticRegression score: 0.741111
-Resultado de todas las características juntas:
LogisticRegression score: 0.791111

Pudiendo concluir que la combinación de todas estas categorías lograba darnos un resultado mejor y que los features de la categoría 3 era la que lo seguía, o sea mientras más información se maneje sobre ciertos nodos en estos casos, se podrá obtener resultados más exactos y que también con los features de categoría 3 es casi tan contundente como el ocupar todos los features.
Tal como se mencionó esta combinación de categorías mostro ser mejor, ya que nos logró dar más información acerca de aquellos nodos y que la categoría 3 nos entrega la información acerca de los vecinos del nodo a predecir, lo cual nos entrega de cierta forma la cercanía siendo también información muy útil.
Al momento de obtener los resultados de los métodos para cada categoría, se tuvo lo siguiente:

| Categoría(c)/Métodos | Precision | Recall | F-1 | AUC |
|---|---|---|---|---|
| C1 | 0.59842519685 | 0.168888888889 | 0.263431542461 | 0.527777777778 |
| C2 | 0.721276595745 | 0.753333333333 | 0.736956521739 | 0.731111111111 |
| C3 | 0.833846153846 | 0.602222222222 | 0.69935483871 | 0.741111111111 |

| | | | | |
|---|---|---|---|---|
| Combinación de c. | 0.804651162791 | 0.768888888889 | 0.786363636364 | 0.791111111111 |

Una vez calculadas la media y desviaciones de las features que fueron:

La media categoría 1 es: 0.517704060153 y la desviación es: 0.0421887330232
La media categoría 2 es: 0.732686818598 y la desviación es: 0.00285462440655
La media categoría 3 es: 0.744878002 y la desviación es: 0.0111891335828
La media categoría 4 es: 0.791229352721 y la desviación es: 0.000309139562826

Lo que nos entrega valores en los cuales sus medias y desviaciones no sufren tanta variación, lo cu al nos muestra como estos resultados no son diferentes.
La normalización en el modelo de regresión genera como filtrar la gran cantidad de datos que se e ntregan en un conjunto de datos más pequeños y específicos, logrando poder entender más rápid o y fácil los features y así minimizando los procesos. De esta forma ayuda a realizar más rápido to dos los procesos de predicción según cada categoría y features.
Una vez normalizado obtuvimos resultados en donde se notó claramente que la categoría 2 era m ucho más exacta al momento de predecir un suceso de compra y venta, y esto se debe a que en la s categorías, muchas están normalizadas en distintos rangos, por lo que al normalizar todos dentro del mismo rango genera mayor “equidad” para poder predecir un suceso entre comprador y vend edor. O sea esto nos entregó un valor con más igualdad entre todos, pero que a la vez también de be haber afectado no positivamente a una categoría, en este caso la combinación de todas las cate gorías.

David Ruz Ordinola
N° alumno: 13635271