Definición de Proyecto No. 2

Laplace Smoothing y EM-GMM

Curso impartido por Samuel Chávez (sachavez@uvg.edu.gt)

CC3045 - Universidad del Valle de Guatemala

Foto: obtenida de la película Her (2013)

Instrucciones

Definición de Proyecto No. 2

A continuación se presentan dos tasks, resuelva cada uno de ellos según las instrucciones. La rúbrica a utilizar para cada uno será la siguiente:

Indicador	Ponderación
Su solución obtiene resultados congruentes	40%
Aplicación clara y correcta de los conceptos teóricos	35%
Calidad de código	25%

Task 1: SMS Spam filtering

Task 1: SMS Spam filtering

Definición de Proyecto No. 2

Construya un filtro de spam a partir del archivo **corpus.txt** (adjunto en la sección de Blackboard para este proyecto) usando el algoritmo de Laplace Smoothing.

corpus.txt fue tomado de: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

Task 1.1: Lectura del archivo

Definición de Proyecto No. 2

Construya un programa que lea e interprete el archivo **corpus.txt**. El archivo tiene las siguientes características:

- 1002 mensajes etiquetados como **ham**
- 322 mensajes etiquetados como **spam**
- El formato del archivo es tal que cada línea es un mensaje con un sufijo que indica su etiqueta. El sufijo se separa del cuerpo del mensaje por un carácter de tabulación (**\t**)

```
ham   What you doing?how are you?
ham   Ok lar... Joking wif u oni...
ham   dun say so early hor... U c already then say...
ham   MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham   Siva is in hostel aha:-.
ham   Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor..
spam  FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone
spam  Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ
spam  URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is
```

Task 1.2: Sanitización

Definición de Proyecto No. 2

Cada mensaje en **corpus.txt** tiene algunos caracteres y palabras que pueden añadir ruido al desempeño de sus filtros, sanitice cada mensaje previamente a construir su clasificador.

Task 1.3: Selección de datos

Definición de Proyecto No. 2

Particione el corpus de mensajes en las categorías siguientes:

- **Entrenamiento**: 80%
- **Cross validation**: 10%
- **Test**: 10%

Nota: asegúrese de dejar una mezcla uniforme de mensajes spam y ham en cada categoría.

Task 1.4: Entrenamiento

Definición de Proyecto No. 2

Elija un ***factor de alisado K*** inicial y construya un clasificador bayesiano usando Laplace Smoothing (K). Utilice la representación de bolsa de palabras para cada mensaje. Utilice los mensajes en la categoría de ***entrenamiento***.

Task 1.5: Optimización de K

Definición de Proyecto No. 2

Encuentre el mejor K que pueda tal que el rendimiento de su clasificador se maximice. Para este propósito utilice los mensajes en la categoría de ***cross validation***.

Task 1.6: Programa de prueba

Definición de Proyecto No. 2

Construya un programa que reciba un archivo de entrada con mensajes sin clasificar, y de como salida cada mensaje clasificado (con la misma notación de prefijo tabulado).

Task 2: EM-GMM para d = 2

Task 2: EM-GMM para d = 2

Definición de Proyecto No. 2

Construya un programa de aprendizaje no supervisado para un espacio de features en dos dimensiones. Recibe de entrada un archivo de coordenadas y da de salida una gráfica de los puntos y los gaussianos para clasificar.

Task 2.1: Lectura del archivo

Definición de Proyecto No. 2

El archivo tendrá una coordenada por línea. Por ejemplo:

```
[4, 5]
[3, 4.5]
[1.1, -2.3]
```

Las coordenadas estarán en la región **(-500, -500)** hasta **(500, 500)**

Task 2.2: Construcción de los K clústers

Definición de Proyecto No. 2

Solicite de entrada la cantidad de Gaussianos a ajustar, y construya iterativamente los mejores parámetros para cada uno de ellos usando el algoritmo de ***expectation maximization***.

Task 2.3: Resultados

Definición de Proyecto No. 2

Muestre en pantalla los puntos de entrada y un diagrama de niveles por cada gaussiano ajustado (GUI).

Task 2.4: Clasificación

Definición de Proyecto No. 2

Construya un programa que reciba de entrada un punto en el espacio de features, e indique el clúster al que pertenece según los gaussianos encontrados en el task 2.2.