Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Tarea 1: Tópicos Avanzados en Inteligencia de Máquina

Profesor : Karim Pichara Baksai
Ayudantes : Ignacio Becker, Francisco Pérez Galarce, Matías Vergara

Fecha de entrega : 3 de Mayo de 2019, 23:59 hrs

1 Objetivos

Esta tarea tiene los siguientes objetivos:

- Conocer e implementar el método *Bayesian Naive Bayes*.
- Relacionarce con *jupyter notebook* y algunos paquetes tales como *matplotlib*,*scipy*, *pandas* y *numpy* utilizados frecuentemente durante el transcurso del curso.

2 Instrucciones

Para desarrollar la tarea se recomienda a los alumnos seguir la siguiente secuencia de actividades:

- Investigar, explicar(con las ecuaciones respectivas e intuición detrás) e implementar el clasificador Naive Bayes "clásico" en jupyter notebook utilizando exclusivamente las librerías antes mencionadas.
- Investigar, explicar(con las ecuaciones respectivas e intuición detrás) e implementar Naive Bayes Bayesiano en jupyter notebook.
- Aplicar ambos métodos a la base de datos publicada junto a este enunciado y descrita en la sección 3. Se espera que para cada método se cree una clase que contenga por lo menos los métodos *fit* y *predict*. El método *fit* debe recibir solo la base de datos y realizar el entrenamiento del algoritmo y el método predict debe recibir un diccionario {y:[clase],[variable 1]:[valor 1],...,[variable n]:[valor n]} y debe retornar la probabilidad

$P(y = [clase]|[variable1] : [valor1], \ldots, [variablen] : [valorn])$. En este punto considere que puede recibir como variables input el conjunto de características completo o un subconjunto de este.

- Realizar un análisis comparativo de los resutados entregados por cada uno de los métodos considerando diferentes factores tanto de la calidad de los resultados como de la interpretación de estos mismos.

3 Base de datos

En esta tarea se trabajará con una base de datos astronómicos. A continuación se entregará una descripción del tipo de datos a utilizar, destacando su relevancia y dando intuiciones respecto a la interpretación de los atributos (descriptores) entregados.

Existen estrellas cuyo brillo puede variar en un período de días e incluso años. Estas estrellas son llamadas "estrellas variables" y se conocen desde hace cientos de años. Su estudio se ha acelerado en los últimos años debido a su importancia en muchas áreas de la astronomía. Por ejemplo, el análisis de los cambios en el brillo de las estrellas nos permite investigar su interior, así como reconstruir la historia de nuestra galaxia, sus posibles interacciones con otras galaxias, o incluso detectar el canibalismo galáctico. Además, conocer las estrellas variables nos ayuda a calcular las distancias desde la Tierra a esas estrellas, creando un mapa tridimensional en el espacio. Los mapas tridimensionales permiten entender el comportamiento a gran escala del universo, como Edwin Hubble lo hizo a inicios del siglo XX para demostrar la expansión del universo, usando justamente estrellas variables.

El estudio de estrellas variables generalmente se basa en el análisis del brillo de éstas a lo largo del tiempo, lo que es conocido en astronomía como *Light Curves*. Así, distintas clases de estrellas variables poseen distintos patrones de variabilidad de brillo, los que funcionan casi como una huella digital, permitiendo entre otras cosas su clasificación.

Idealmente, todas las observaciones se realizarían de forma simultánea en todo el cielo. Sin embargo, en la práctica, los telescopios no pueden observar el cielo con la suficiente rapidez (lo más cercano que tendremos en el mediano plazo es la llegada del LSST el 2022, para el cual en el DCC nos estamos preparando). Además, existen fenómenos climáticos e incluso las fases de la Luna que afectan dichas observaciones. El resultado son curvas de luz cuyas mediciones ocurren en forma aleatoria y con diferente número de observaciones en cada una. Para resumir y estandarizar dicha información, se calculan descriptores estadísticos o *features* a partir de las series de tiempo. Estas *features* condensan la información a una dimensión

definida, reduciendo la complejidad y permitiendo usar una amplia gama de algoritmos que permiten realizar clasificación.

En esta tarea usaremos un set de features discretizadas:

- Amplitud: Se refiere a qué tan grande es el cambio de brillo. En general cambios grandes indican procesos más intensos. Puede ir desde milésimas hasta decenas de veces el cambio de brillo.

- Periodo: Esta feature representa el periodo principal de pulsación en días. El periodo es un valor crítico para calcular distancias a estas estrellas ya que permite relacioanrlo con el brillo intrínseco de algunas clases de estrellas.

- Desviación Estándar: Mide la dispersión de la variación. Puede asociarse al tiempo que permanece la estrella alejada de su brillo promedio.

- Media: El brillo de una estrella que medimos depende en parte de su masa y su estado de evolución. Ciertos tipos de variables tienen una edad determinada que puede correlacionarse con un brillo medio determinado.

- Max Slope: Es el valor absoluto de la pendiente máxima entre dos puntos consecutivos de la serie de tiempo. Mayores valores indican fenómenos más rápidos y por lo tanto, más intensos.

- Mean variance: Es definido como la desviación estándar normalizada por la media. Valores altos indican fuerte variabilidad y es usado para separar candidatos de estrellas no variables. Estas últimas tendrían un valor cercano a cero.

- Linear trend: Esta feature es la pendiente de un ajuste lineal a la serie de tiempo. Indica posibles comportamientos a largo plazo subyacentes a la variabilidad. Por ejemplo, una estrella que está creciendo en el tiempo.

4 Entregable

El entregable de esta tarea es un archivo comprimido con:

- *Notebook* (extensión .ipynb) en el cual se encuentre la teoría detallada de *Naive Bayes* y *Bayesian Naive Bayes* (usar el *markdown* de Jupyter Notebook para este fin), código y todos los comentarios y supuestos necesarios para entenderlo. El archivo .ipynb deberán nombrarlo $[numero\_alumno]\_T1.ipynb$.

- Archivo con extensión *.py* cuyo contenido es **solamente el código** (el mismo del *.ipynb*).El archivo .py deberán nombrarlo $[numero\_alumno]\_T1.py$.

El archivo comprimido debe ser subido al cuestionario abierto en el sistema SIDING específicamente para esta tarea hasta el día y hora señalada con el nombre $[numero\_alumno]\_T1.rar$. Las tareas atrasadas enviadas por otro medio no serán evaluadas y serán calificadas con nota 1.0. **La tarea es estrictamente individual**.

5 Aspectos a evaluar

Los criterios utilizados para la evaluación serán:

- Implementación: En este ítem se evaluará el grado de funcionamiento de los métodos implementados (bugs, funcionamiento en diferentes bases de datos, cálculos, etc.) y la utilización de los paquetes ya mencionados en la sección 1.

- Comprensión del método: En este ítem se evaluarán las explicaciones entregadas en el archivo jupyter notebook. En este punto, el alumno debe presentar el método desde un punto de vista matemático pero a la vez se espera de intuiciones respecto al funcionamiento de los modelo.

- Experimentos y Análisis de resultados: Se medirá la cantidad y calidad de los experimentos (no basta solo con correrlo), el alumno debe realizar una cantidad de experimentos que permitan realizar análisis robustos. Para realizar los experimentos se recomienda dividir la base de datos en un conjunto de entramiento y un conjunto de testeo, en tanto, debe entregar como principal output las matrices de confusión de sus algoritmos en las diferentes configuraciones que determine. Es deseable que sean capaces de generar gráficas de apoyo que permitan apreciar las diferentes ventajas que tiene la versión bayesian de Naive Bayes.

6 Bibliografía Sugerida

Se sugiere la siguiente bibliografía para entender el clasificador *Naive Bayes* y su variante bayesiana:

- *"Machine Learning: A Probabilistic Perspective"(Murphy K, MIT Press)* (el capítulo está disponible en SIDING).

- https://www.youtube.com/watch?v=CW9YGii1nSA

7 Política de Integridad Académica

Los alumnos de la Escuela de Ingeniería deben mantener un comportamiento acorde al Código de Honor de la Universidad:

> *"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad"*

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por "copia" o "plagio" se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.