Métodos Numéricos
Primer Cuatrimestre 2017
Trabajo Práctico 3

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

¡No te calentés!

Contexto y motivación

Los pronósticos del clima han sido desarrollados durante milenios con diversos objetivos. Actualmente, la posibilidad de tener pronósticos precisos impacta en la capacidad de previsión en diversas actividades económicas como la industria agropecuaria. Además, las predicciones diarias permiten facilitar la vida del ciudadano de a pie (*¿llevo o no llevo paraguas?*) pero en el ámbito académico los pronósticos se desarrollan a distintos niveles.

En los pronósticos a largo plazo, es objetivo es poder predecir comportamientos en períodos de varios años. Este tipo de análisis permite detectar patrones o tendencias como aquellos correspondientes al calentamiento global. En este trabajo nos enfocaremos en pronósticos de este tipo y en análisis de zonas geográficas a partir de pocos datos disponibles.

El problema

Diversas técnicas han sido desarrolladas alrededor de los pronósticos climáticos utilizando, cada vez más, modelos computacionales. En este trabajo nos enfocaremos en modelos usando cuadrados mínimos lineales para predecir la temperatura de ciudades, regiones o el planeta.

Para realizar esta tarea, contaremos con información de las temperaturas de ciudades (junto con la latitud y la longitud), países y el planeta en distintos períodos. Para países y ciudades se cuenta con datos mes a mes dependiendo del momento en el que se comenzaron a realizar las mediciones. Esto implica que no todas los ciudades o países cuentan con información para los mismos períodos de tiempo, de modo que será importante tener en cuenta la información disponible para hacer los modelos adecuados. La temperatura del planeta está reportada con un valor por año. [1]

En el marco del trabajo práctico se pretende generar modelos predictivos en dos niveles de acuerdo a lo que se busca predecir.

Por un lado, se busca predecir la temperatura del planeta a fin de establecer y analizar patrones de dos maneras: utilizando temperaturas de países en un caso y utilizando funciones matemáticas en otro. Los dos modelos planteados deberán ser debidamente explicados justificando las elecciones tomadas para generarlos en el marco del uso de cuadrados mínimos lineales y deberán ser comparados entre sí respecto de sus capacidades predictivas.

Por otro lado, se busca predecir las temperaturas de ciudades utilizando datos de otras ciudades o países. En este caso interesará evaluar la capacidad de predicción pero además analizar resultados variando los conjuntos de datos de entrada con el objetivo de encontrar un modelo

[1] Los datos son un subconjunto de los disponibles en https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data y de https://www.ncdc.noaa.gov/cag/time-series/global/globe/land_ocean/1/3/1880-2017 con algunas modificaciones de formato.

que produzca buenos resultados pero usando la menor cantidad de variables como entrada. Dicha comparación deberá tener en cuenta, además, ls posición geográfica de la cual provienen los datos de entrada.

La realización del trabajo debe basarse en los dos enfoques descriptos (predecir la temperatura del planeta por un lado y predecir la temperatura de ciudades por otro) pero en cada caso será requisito que el grupo proponga alguna variante original para el análisis. Por ejemplo, podría buscarse explicar cierto fenómeno climático o predecir las temperaturas de ciudades (o regiones) similares en sus características pero distantes geográficamente. Las características pueden ser latitud y longitud pero también altitud sobre el nivel del mar, distancia al mar, nivel de humedad o cualquier característica que el grupo considere relevante y que pueda conseguir para los datos a utilizar. Eventualmente, podrá cambiarse el eje para, por ejemplo, predecir la latitud de las ciudades en función de las temperaturas. En cualquier caso, deberán quedar claramente explicitados cuáles son los análisis sobre los cuales trabajará el grupo.

Dado que el clima varía notablemente según la geografía, un aspecto relevante al momento de generar las predicciones será la ubicación de las ciudades o los países cuyos datos se usen. En particular, al usar datos de ciudades, es posible utilizar la información de latitud y longitud para hacer los análisis.

Otro aspecto importante es la periodicidad con la que se reportan los datos. Queda a criterio del grupo decidir si los análisis y predicciones se harán por mes, por año u otro período de tiempo (por ejemplo por lustro o década). Cualquiera sea el caso elegido, deberá justificarse la elección y explicar cómo se obtienen los valores de un nivel de periodicidad a partir de los valores de otro nivel.

Para los modelos propuestos será necesario explicar qué criterios se usaron para formularlos. Algunas ideas para ganar conocimiento sobre el impacto de los datos que se tienen es analizar el nivel de correlación (por ejemplo midiendo covarianza entre las variables) y/o basarse en bibliografía del tema (que deberá ser citada debidamente). En todos los casos se espera que el grupo sea crítico sobre la información que encuentre así como sobre las métricas propuestas fundamentando sus críticas y realizando experimentos con los datos que permitan validar o refutar sus conjeturas.

Técnicas a utilizar y métricas de evaluación

La técnica de Métodos Numéricos a utilizar para proponer los modelos es Regresiones Lineales/Cuadrados Mínimos Lineales (CML). Para determinar el modelo, se tiene una serie de N observaciones $(x_{(i)}, y_{(i)})$, con $x_{(i)} \in \mathbb{R}^k$ el vector de *features* (las variables de entrada) e $y_{(i)} \in \mathbb{R}$ la variable dependiente (la temperatura a predecir en nuestro caso). Luego, el modelo consiste en encontrar los parámetros (lineales) que definen $y_{(i)} = f(x_{(i)}) + \epsilon_i$, $i = 1, \ldots, N$, donde ϵ_i es el error de la medición i-ésima, y que minimizan el error de la aproximación en el sentido de CML.

Dado un conjunto de datos $\{(x_{(i)}, y_{(i)}\}_{i=1,\ldots,N}$ será necesario considerar distintas hipótesis sobre la función f que dan lugar a distintos modelos. Para poder decidir entre los mismos, es necesario considerar alguna métrica de evaluación. Se sugiere considerar el *Mean Squared Error* (MSE). Dado un modelo $\hat{f}$ de f y una observación $(x_{(i)}, y_{(i)})$, se define $\hat{y}_{(i)} = \hat{f}(x_{(i)})$ y $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$. Con estas definiciones, se puede calcular el MSE del modelo $\hat{f}$ como

$$MSE(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} e_{(i)}^2.$$

Otra posibilidad consiste en considerar $\max\limits_{i\in\{1\ldots N\}} e_{(i)}^2$ de manera que se busque minimizar el máximo error. Ya sea que se elija uno de estos criterios u otro propuesto por el grupo será necesario justificar brevemente la elección.

Esta metodología sirve para evaluar cuán bien ajusta el modelo en función de los datos de entrenamiento utilizados. Sin embargo, en un contexto de modelos predictivos se corre el riesgo de caer en el conocido *overfitting*. Para evitar este fenómeno, se puede considerar la técnica de *cross-validation* (CV). Es decir, particionar el conjunto de datos y variar la composición de la base de entrenamiento (*training*) y las observaciones consideradas como *test*. Una vez obtenido el modelo $\hat{f}$, se toman las observaciones en el conjunto de test, se aplica el modelo y se evalúa la métrica de evaluación obtenida. La métrica final para el modelo $\hat{f}$ consiste en tomar alguna medida sobre los resultados obtenidos para cada combinación de training/test considerado.

Es importante notar que si se consideran datos de períodos en el tiempo prolongados o muy cortos para train/test se debe tener en cuenta que el conjunto sea representativo. Es sabido que existen fenómenos climáticos que tienen duraciones acotadas en el tiempo de modo que si uno entrena con datos de un cierto período de años y evalúa en otro muy posterior es posible que estos efectos hagan que los resultados no sean demasiado confiables. Será necesario explicar datos de qué períodos se consideran y justificar brevemente la elección.

Enunciado

El Trabajo Práctico consiste en considerar los datos sobre temperaturas descriptos y formular modelos para la predicción de la temperatura del planeta (un modelo usando datos de países y otro usando expresiones matemáticas en función del año) y para la predicción de la temperatura de ciudades (a partir de la información de países u otras ciudades). Para ello, se deberá utilizar CML como técnica de análisis y modelado, tanto a nivel descriptivo de los datos como a nivel predictivo de resultados futuros. Para el desarrollo de los métodos se podrá considerar como posibles lenguajes MATLAB/Octave, Python y/o C++. Se remarca que, a diferencia de trabajos anteriores, no es necesario realizar toda la implementación desde cero y es posible utilizar rutinas provistas por dichos lenguajes. El objetivo principal de este trabajo se centra en la aplicación de las técnicas de CML a una temática práctica concreta y en la correspondiente experimentación necesaria para evaluar los desarrollos.

Junto con el enunciado se proveen una serie de archivos con información sobre temperaturas en ciudades, países y el mundo. Además, se presentan scripts para extraer determinados indicadores de los archivos como las ciudades de un país o la información sobre ciertos países. Una vez aplicados dichos scripts, los resultados podrán eventualmente resultar más fáciles de utilizar en el código que realice el modelo.

Estas herramientas son posibles puntos de partida para que los grupos extraigan la información relevante para su caso de estudio. Sin embargo, si es necesario, el grupo puede plantear sus propias herramientas de scripting para filtrar los datos de acuerdo a sus necesidades. En este sentido, es posible que los grupos compartan, a través de la lista de alumnos de la mate-

ria, herramientas de preprocesamiento y extracción de datos con otros grupos. Es importante evitar que las herramientas compartidas contengan información particular de las métricas planteadas y la experimentación a realizar.

Los resultados deben ser volcados en un informe con la estructura habitual. Sin embargo, en este caso es obligatorio escribirlo utilizando el template de la revista *Electronic Notes on Discrete Mathematics* (ENDM). Además, el informe no podrá exceder las 10 páginas de longitud y, por lo tanto, los resultados tienen que ser presentados y condensados de forma adecuada. Notar que esto no significa que la experimentación debe ser acotada, sino todo lo contrario: es importante realizar muchos experimentos y mostrar los que resulten representativos. Como en los demás trabajos, es importante proveer la información necesaria para poder replicar todos los experimentos, ya sea que se encuentren en el informe o no.

Por último, este trabajo tendrá una presentación oral frente a un grupo de docentes que será evaluada como una parte adicional de la nota. Para la misma, cada grupo diseñará una presentación incluyendo los desarrollos y resultados que considere interesantes, plasmados en el informe, y dispondrá de 15 minutos para exponerlo. La exposición puede ser de la totalidad o de un subconjunto de los integrantes, y esta decisión queda a elección del grupo. Una vez finalizada la misma, se llevará a cabo un coloquio donde los integrantes del grupo responderán a las preguntas realizadas. Cabe mencionar que los docentes podrán elegir qué alumno debe responder, con lo cual es importante que todos los integrantes estén al tanto de todas las decisiones tomadas.

Fechas de entrega

- *Formato Electrónico:* Lunes 26/6, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP3]` seguido de la lista de apellidos de los integrantes del grupo.
- *Confirmación presentación oral:* Miércoles 28/6, por correo electrónico.
- *Presentación oral:* Lunes 3/7, en horario a determinar luego de la confirmación. Será en horario de clase de la materia.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.