Tarea 1

Profesor: Felipe Tobar
Auxiliares: Mauricio Araneda, Alejandro Cuevas, Mauricio Romero
Consultas: Mauricio Araneda
Fecha entrega: 11/4/2019

Formato entrega: Informe en formato PDF, con una extensión máxima de 3 páginas (puede usar un formato de doble columna), presentando y analizando sus resultados, y detallando la metodología utilizada. Adicionalmente debe entregar el jupyter notebook (o el código que haya generado) con la resolución de la tarea.

P1. Máxima Verosimilitud (2.0 puntos, tomado de MacKay p.309)

Siete científicos con habilidades experimentales salvajemente dispares reportan distintas estimaciones de un parámetro μ

Científico	A	B	C	D	E	F	G
Estimación	-27.020	3.57	8.191	9.898	9.603	9.945	10.056

Discuta cómo encontrar el parámetro óptimo y cuán confiable es cada científico. Para este fin asuma que la estimación de cada científico puede ser considerada como una muestra de distribuciones normales de igual media (μ) pero distintas varianzas ($\sigma_1^2, \ldots, \sigma_7^2$). Observe de los datos que las estimaciones de los científicos A y B son poco confiables y que el parámetro buscado debería estar entre 9 y 10 ¿es posible abordar este problema usando máxima verosimilitud?

Valide (o rechace) su respuesta mediante simulaciones. Explicite sus supuestos e interprete sus resultados. (Hint: Grafique la verosimilitud)

P2. Regresión Lineal (3.5 puntos)

En el archivo **datos/szege_clima.csv** se encuentra un *dump* (subconjunto del elementos del dataset original) de datos climáticos de la ciudad de Szege, la tercera ciudad más grande de Hungría. El archivo consiste en dos columnas, la primera X corresponde a la proporcion de humedad ambiental, la segunda Y corresponde a la sensación térmica. Mayor detalle sobre los datos utilizados puede encontrarse en:

`https://www.kaggle.com/budincsevity/szeged-weather`

Para este conjunto de datos se pide implementar una regresión lineal regularizada en Python donde X será el regresor y Y la variable a estimar. Esto usando solo operaciones álgebra lineal. Para esto deberá:

(0) Instalar la última versión de Anaconda y lanzar Jyputer Notebook.

(a) (0.5 puntos) Cargar los datos desde el archivo **datos/szege_clima.csv** y graficarlos.

(b) (1.0 puntos) Crear y ajustar su modelo de regresión lineal regularizada, implementando regularización *ridge*, pruebe con distintos valores del factor ρ (partiendo de $\rho = 0$). Se recomienda crear la función $reg\_lineal(X, Y, \rho)$ que devuelva los parámetros θ.

(d) (0.5 puntos) Grafique el valor de los coeficientes para distintos valores de ρ.

(e) (0.5 puntos) Grafique el error cuadrático medio y varianza para distintos valores de ρ.

(f) (1.0 puntos) Grafique y discuta los resultados, puede ayudarle graficar distintas curvas regularizadas sobre el dataset. A modo de guía en la discusión puede considerar las siguientes preguntas (no se limite únicamente a éstas):

- ¿En qué afecta que la regresión esté regularizada sobre este conjunto de datos?
- ¿Qué pros y contras otorga la regularización? ¿Cuál es preferible en este contexto?
- ¿Qué modelo tiene mejor desempeño en sus predicciones? ¿Cuál tiene el menor error asociado?

No se permite el uso de paquetes predefinidos para regresión lineal. Estos pueden ser considerados para contrastar los propios resultados pero no para resolver la pregunta. E.g., `numpy.polyfit`, `scipy.stats.linregress`, `sklearn.linear_model.LinearRegression`

P3. Proyecto curso (0.5 puntos)

Discuta brevemente su idea de proyecto para el curso (50-100 palabras). El proyecto puede ser aplicado o más teórico, motivado por problemas existentes de sus proyectos/memorias/tesis, emprendimientos, o bien puede inspirarse en los siguientes sitios:

`http://games.cmm.uchile.cl/courses/MA5203/`
`https://www.kaggle.com/competitions`