Trabajo Práctico 3

Estudio de modelos lineales

Simulaciones

El archivo `tp3_train.csv` contiene 100 observaciones que se usarán para identificar los parámetros de un modelo, mientras que el archivo `tp3_test.csv` tiene 20 observaciones que se usarán para evaluar los resultados del proceso de inferencia.

Se sabe que los datos fueron generados utilizando un modelo de la forma:

$$y_i \sim \mathcal{N}(\theta_3 x_i^3 + \theta_2 x_i^2 + \theta_1 x_i + \theta_0, \sigma^2)$$

aunque no se conocen los valores de los θ_i ni de σ

En primer lugar, se estudiará el **efecto de la cantidad de datos** utilizados para el ajuste del modelo.

1. Realice un análisis exploratorio de los datos de ajuste del modelo (*train*) y proponga *priors* vagos de acuerdo a la escala de los datos. Realice simulaciones predictivas *a priori*.

2. Utilice los *priors* propuestos en el punto anterior para ajustar el modelo utilizando 10, 20, 50 o 100 observaciones de los datos de entrenamiento. Compare las distribuciones *a posteriori* de los parámetros, la media condicional y las predicciones sobre el conjunto de datos de evaluación.

En segundo lugar, se investigará qué efecto tiene la elección de **distribuciones *a priori* centradas en valores erróneos** de los parámetros de la regresión.

3. Considere las siguientes distribuciones *a priori*:

$$\begin{aligned}
\theta_0 &\sim \mathcal{N}(0.6, 0.2)\\
\theta_1 &\sim \mathcal{N}(-0.2, 0.2)\\
\theta_2 &\sim \mathcal{N}(2.1, 0.2)\\
\theta_3 &\sim \mathcal{N}(1, 0.2)
\end{aligned}$$

Compare, como hizo en el caso anterior, los resultados en las estimaciones utilizando 10, 20, 50 y 100 observaciones.

Finalmente, se considera el caso donde se propone un **modelo erróneo**.

4. Utilizando *priors* vagos, ajuste polinomios de grado 3, 4, 5 y 6 utilizando 20 observaciones. Compare gráficamente el ajuste en los datos de entrenamiento con el ajuste en los datos de evaluación. ¿Qué ocurre si se utilizan todos los datos del conjunto de entrenamiento?

5. Utilice ahora *priors* de regularización (centrados en 0 y con diferentes varianzas: 0.01, 0.1, 1 y 10) para polinomios de grado 3, 4, 5 y 6 utilizando 20 observaciones. Compare gráficamente el ajuste en los datos de entrenamiento con el ajuste en los datos de evaluación. Como hizo en el punto anterior, analice que ocurre cuando utiliza todos los datos del conjunto de entrenamiento.

Enfriamiento de agua en un termo

En un país matero como Argentina, era de esperarse que aparecieran casi tantos termos como personas. Es difícil decir si la cantidad de variantes de termos en el mercado nacional cambió, pero es indudable que, tras el furor del termo color *verde militar*, la elección de un termo adquirió un papel relevante en el ritual del mate.

De acero, de vidrio, con capa aisladora. Diferentes configuraciones dan lugar a distintas capacidades de mantener la temperatura y, obviamente, a distintos rangos de precios. Dejando de lado la cuestión monetaria, centrémonos en estudiar cómo varía la temperatura de un líquido en el interior de un termo en función del tiempo transcurrido.

La temperatura es una medida del grado de agitación de las partículas de una sustancia. Un líquido (o sólido, o gas) está *más caliente* que otro si sus partículas tienen (en promedio) mayor grado de agitación. Sabemos por evidencia empírica que si un cuerpo se pone en contacto con otro que tiene una temperatura menor, hay una transferencia de energía que hace que el primero se enfríe y el segundo se caliente, hasta que alcanzan el denominado equilibrio térmico.

De manera similar, esto es lo que ocurre con el agua que dejamos dentro del termo: en algún momento, llega al equilibrio térmico con el ambiente. La salvedad necesaria acá es que, como el ambiente es grande, no aumenta su temperatura con la energía que pierde el agua del recipiente.

Ahora bien, el ritmo con el cual el agua caliente pierde energía no es constante. Físicamente, mientras mayor sea la diferencia de temperatura entre dos cuerpos, más rápido fluirá la energía (y más rápido cambiará la temperatura). Si estudiamos la temperatura del agua en el termo en función del tiempo, notaremos que el ritmo con el que cambia decrece a medida con el que transcurre el tiempo.

La lectura del párrafo anterior debería permitir asociar el concepto de *ritmo de cambio* con la noción matemática de derivada. En efecto, la derivada de la temperatura respecto al tiempo varía con el tiempo. En otras palabras, la pendiente no es constante.

Las leyes que rigen el universo pueden muchas veces formularse en términos de lo que en matemática se conoce como ecuación diferencial. En este caso, la temperatura del agua en el termo satisface la siguiente ley:

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = r[T_{\text{amb}} - T(t)]$$

donde T_{amb} es la temperatura ambiente (un valor fijo y conocido), r es una constante y $T(t)$ es la función (en principio desconocida) que describe la temperatura del agua del termo en función del tiempo.

No se trata de una ecuación algebraica donde la solución es un valor numérico sino de una ecuación donde la solución es una función. Buscamos una función $T(t)$ que satisfaga la ecuación: su derivada debe cambiar con el valor que toma la función.

Una función que satisface esa ecuación es:

$$T(t) = T_{\text{amb}} + (T_i - T_{\text{amb}})e^{-rt}$$

siendo T_i la temperatura a la que está inicialmente el agua en el termo (un valor fijo y conocido).

1. Verificar que la función anterior satisface la ecuación diferencial
2. Grafique $T(t)$ para $T_{\text{amb}} = 20$ C° y $T_i = 90$ C°, para dos valores de r, $r_1 = 0.1$ y $r_2 = 0.3$. ¿Qué representa r?
3. Según su experiencia con termos, ¿cuál es un valor realista de r?

Estudiaremos a continuación un conjunto de mediciones de temperatura de agua en un termo en función del tiempo transcurrido.

Leonel tiene un termo Estanli™ que compró por Amason y se dispone a despejar la duda de cualquier usuario de termos Estanli™: ¿cuánto dura el agua caliente? Pone agua en la pava eléctrica, la vierte en el termo y registra la temperatura en algunos momentos posteriores. Ese día, el reporte meteorológico indica una temperatura de $T_{\text{amb}} = 23$ C°. Las temperaturas que registró Leonel son las siguientes:

t (h:mm)	T (°C)
1:20	92.0
2:30	90.5
4:00	81.4
5:15	80.8
8:30	74.2

Para simplificar la construcción de un modelo, en lugar de considerar la temperatura del agua en el termo, se considerará la diferencia entre la temperatura del agua y la temperatura ambiente $T - T_{\text{amb}}$. Además, se llamará T_{diff} a la diferencia de temperatura entre la temperatura inicial del agua y la temperatura ambiente $T_i - T_{\text{amb}}$.

$$T(t) - T_{\text{amb}} = T_{\text{diff}} e^{-rt}$$

4. Verifique que el logaritmo natural de la nueva variable $(T(t) - T_{\text{amb}})$ es una función lineal de t. ¿Qué representan el intercepto y la pendiente? Llámelos β_0 y β_1.

Se propone ajustar un modelo lineal normal utilizando los datos transformados.

5. En función del enunciado del problema y de su conocimiento de termos, elija una distribución *a priori* para β_0, β_1 y σ. ¿Cuáles son las implicancias de sus distribuciones *a priori*? Realice pruebas predictivas *a priori*.
6. Ajuste el modelo lineal utilizando `R`.
7. Encontrar el *posterior* de r, de T_{diff} y de la temperatura inicial del agua T_i.
8. Predecir la temperatura a la que estará el agua transcurridas 12:00 h del inicio de la experiencia.