Bayesian Statistics and Machine Learning Workshop 2023

Aproximación Gaussiana y Teoría de ajustes

Martín Onetto

1. Questions

1. Qué hipótesis son las que se asumen al proponer como likelihood una distribución gaussiana.
2. Discutir en qué consiste la aproximación de Laplace. Por qué es útil?
3. Demostrar que hacer un ajuste de parámetros por cuadrados mínimos coincide con calcular el MLE de una distribución gaussiana.
4. Discutir qué diferencia conceltual hay entre calcular la incerteza de los parámetros evaluando la segunda derivada en el máximo con calcular la infromación de fisher del parámetro.

1.1. Problems

1. **Global warming**

   a) Cargar los datos GlobalWarming.csv y gráficar la dependencia de las temeraturas medias con su incerteza en función del tiempo.

   b) Ajustar un modelo lineal $y = ax + b$, donde x es el número de día e y es la temperatura. Tomar como $\sigma = 2$ igual para todos los puntos. Encontrar la distribución posterior de los parámetros usando la aproximación gaussiana.

   c) Dar el intervalo de credibilidad de 95 % de ambos parámetros.

   d) Tomar muestras de la distribución posterior obtenida y graficar sobre los datos las rectas que se derivan de ellas.

   e) Discutir la proyección del modelo a datos que todavía no se observaron.

   f) Cómo cambia la incerteza en los valores de y a medida que me alejo del rango de datos en donde se ajustó el modelo? Para responder hacer un gráfico de la incerteza $P(\tilde{y}|x_{\text{not seen}}y)$ en función del tiempo. Interpretar.

   g) Repetir lo anterior proponiendo un modelo cuadrático, $y = ax^2 + bx + c$

h) (Optional) En el caso de tener una incerteza diferente en cada punto de la regresión cambian las expresiones para encontrar los máximos e incertezas de los paráemetros. Desarrollarlas. Repetir ambos ajustes tomando las incertezas que corersponden a cada día. Cómo cambian las predicciones? Cómo se manifiesta en el ajuste que algunos días tengan mayor incerteza que otros?

2. **Sun spectrum**

   a) Cargar la base de datos de sunspectra.csv que corresponde al espectro de emisión del sol.

   b) Ajustar la temperatura del sol utlizando la ley de Plank de radiación de cuerpo negro:

$$B_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/(\lambda k_\mathrm{B} T)} - 1} \tag{1}$$

   Usando como likelihood una distribución gaussiana y una de Poisson. Discutir las diferencias. Calcular la incerteza de la temperatura obtenida

3. **California Houssing**

   a) Cargar la base de datos de housing.csv que corresponde a los precios de las casas en california según diferentes variables. Numéricas y categóricas, a esta altura nos interesarán sólo las numéricas.

   b) Ajustar la variable *median house value* con un modelo lineal $y = \beta^T X$ respecto a las otras variables X. Para esto considerar que cada conjunto de variables es $X_i = (1, x_1, x_2, \ldots, x_p)_i$ donde p la dimensión y donde agregamos un 1 que nos será útil para calcular la ordenada al origen. En este esquema el óptimo valor de $\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y$ y su varianza es $Var(\beta) = (X^T X)^{-1} \sigma^2$ donde σ lo estimamos como la varianza de los datos en torno a las predicciones del modelo lineal $\hat{\sigma}$. Es decir tomando $\hat{y} = \hat{\beta}_{MLE} X$, luego $\hat{\sigma} = \frac{1}{N-p-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$