Discussion of some exercises

November 23, 2023

Predicting with Bayes

- In a Bayesian model where $X_1, \ldots, X_n, X_{n+1}$ are (conditionally on the unobservable θ) i.i.d. distributed according to $f(x|\theta)$, moreover θ is distributed according π.
- The *predictive distribution* is the distribution of X_{n+1}.
- The *posterior predictive distribution* is the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$.
- f and π are $\mathcal{N}(\theta, \sigma_0^2)$ and $\mathcal{N}(\theta_0, \tau_0^2)$. Evaluate predictive and posterior predictive distributions and discuss their limiting behavior when $n \to \infty$.
- Model

$$f(x_1, \ldots, x_n, x_{n+1}|\theta) = \prod_{i=1}^{n+1} f(x_i|\theta) \propto \exp\left(-\sum_{i=1}^{n+1} (x_i - \theta)^2 /(2\sigma_0^2)\right)$$

- Prior

$$\pi(\theta) \propto \exp\left(-\theta^2/(2\tau_0^2)\right)$$

- In the terms of random variables

$$(X_1, \ldots, X_{n+1}|\Theta = \theta) \sim \mathcal{N}(\theta, \sigma_0 \mathbf{I}_{n+1}), \quad \Theta \sim \mathcal{N}(0, \tau_0^2)$$

- What is the joint distribution $(X_1, \ldots, X_{n+1}, \Theta)$?

Example – normal distribution with missing data

- Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be i.i.d. as $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.
- Suppose that some of the Z_i and some of the Y_i are missing as follows:
    - For $1 \leq i \leq n_1$ we observe both Z_i and Y_i,for $n_1 + 1 \leq i \leq n_2$, we observe only Z_i, and for $n_2 + 1 \leq i \leq n$, we observe only Y_i. The observed data are denoted by S.
    - In this case a set of sufficient statistics for the complete data is

$$T = (\bar{Z},\ \bar{Y},\ \overline{Z^2}, \overline{Y^2}, \overline{ZY}).$$

- One wants to reconstruct the missing parts needed for T from the incomplete data given in S.

The EM algorithm for exponential families

Theorem 2.4.3

For a canonical exponential family generated by (T, h) satisfying the conditions of Theorem 2.3.1. Let $S(X)$ be any statistic, then the EM algorithm consists of the alternation

$$\dot{A}(\theta_{new}) = E_{\theta_{old}}(T(X)|S(X) = s).$$

Example (cont.): We can see that

$$\dot{A}(\theta) = E_\theta T = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \sigma_1\sigma_2\rho + \mu_1\mu_1)$$

so the left-hand side is straightforward. The right hand side can be easily derived by noting the well-known relations

$$E_\theta(Y|Z) = \mu_2 + \rho\sigma_2(Z - \mu_1)/\sigma_1$$
$$E_\theta(Y^2|Z) = (\mu_2 + \rho\sigma_2(Z - \mu_1)/\sigma_1)^2 + (1 - \rho^2)\sigma_2^2$$

leading to the new paremater values being simply regular functions of the sufficient statistic T but evaluated at the expected values in the E-step of the algorithm. See Example 2.4.6 for details.

Exercise

In the bivariate normal Example 2.4.6.

- complete the E-step by finding $E(Z_i|Y_i)$, $E(Z_i^2|Y_i)$, and $E(Z_iY_i|Y_i)$.
- verify the M-step by showing that

$$E_\theta \mathbf{T} = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \rho\sigma_1\sigma_2 + \mu_1\mu_2).$$

Hint: Use the conditional distributions of the two dimensional Gaussian vectors so $Z|Y$ is Gaussian, $E(a + bX)^2 = a^2 + 2abE(X) + b^2E(X^2)$, $E(YZ|Y) = YE(Z|Y)$.

Multivariate normal (Gaussian) distribution

Everyone believes in Gauss distribution: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact.

Quote attributed to Henri Poincaré by de Finetti. However, Cramer attributes the remark to Lippman and quoted by Poincaré; *Gabriel Lippman* – a Nobel prize winner in physics, *Henri Poincaré* – a mathematician, theoretical physicist, engineer, and a philosopher of science

- The multivariate normal or Gaussian random vector $\mathbf{X} = (X_1, \dots, X_p)$ is given by density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  that is characterized by: a vector parameter $\boldsymbol{\mu}$ and a matrix parameter $\boldsymbol{\Sigma}$.
- The notation $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ should be read as *“the random vector $\mathbf{X}$ has multivariate normal (Gaussian) distribution with the vector parameter $\boldsymbol{\mu}$ and the matrix parameter $\boldsymbol{\Sigma}$.”*

Multivariate normal (Gaussian) distribution – properties

We often drop the dimension p from the notation writing $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- The vector parameter $\boldsymbol{\mu}$ is equal to the mean of $\mathbf{X}$ and the matrix parameter $\boldsymbol{\Sigma}$ is equal to the covariance matrix of $\mathbf{X}$.
- Any coordinate X_i of $\mathbf{X}$ is also normally distributed, i.e. X_i has $\mathcal{N}(\mu_i, \sigma_i^2)$.
- If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ is a $q \times p$ (non-random) matrix, $q \leq p$, (and the matrix $\mathbf{A}$ is of the rank q), then

$$\mathbf{AX} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

Subsetting from coordinates of MND

Any vector made of a subset of different coordinates of X is also multivariate normal with the corresponding vector mean and covariance matrix.

More precisely, if $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

are partitioned into sub-vectors $\mathbf{X}_1 : q \times 1$ and $\mathbf{X}_2 : (p-q) \times 1$ then with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{X}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ and } \mathbf{X}_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Conditional distributions

If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

are partitioned into sub-vectors $\mathbf{X}_1 : q \times 1$ and $\mathbf{X}_2 : (p-q) \times 1$ then with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$, is

$$\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Regression reinterpretation of conditional distributions

Vector $\mathbf{X}_1$ given $\mathbf{X}_2$ forms a regression model

$$\mathbf{X}_1 = \mathbf{a} + \mathbf{D}\mathbf{X}_2 + \epsilon,$$

where

- The constant term $\mathbf{a} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$
- The design matrix $\mathbf{D} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$
- The error term $\epsilon \sim \mathcal{N}_q(0, \boldsymbol{\Sigma}_{11} - \boldsymbol{D}\boldsymbol{\Sigma}_{21})$

Special case $\mathbf{X}_1 = (X_i, X_j)$ – calculating partial covariances

Partial covariance matrix

Recall that the partial covariance 2×2 matrix $\mathbf{\Sigma}_{ij}$ of (X_i, X_j) is given at the covariance of their distribution conditionally all other variables:

$$(X_i, X_j) = (a_i, a_j) + \mathbf{D}\mathbf{X}_2 + \epsilon,$$

- The constant term $(a_i, a_j) = (\mu_i, \mu_j) - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$, where $\mathbf{\Sigma}_{12}$ is made of the ith and jth rows of of $\mathbf{\Sigma}$ without the ith and jth coordinates in these rows, thus it is $2 \times (p-2)$ matrix, $\mathbf{\Sigma}_{22}$ the covariance matrix with out the ith and jth columns and rows, thus it is a $(p-2) \times (p-2)$ matrix, $\boldsymbol{\mu}_2$ the mean values with the μ_i and μ_j values dropped.
- The $2 \times (p-2)$ design matrix $\mathbf{D} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$,
- The error term $\epsilon \sim \mathcal{N}_q(0, \mathbf{\Sigma}_{ij})$, $\mathbf{\Sigma}_{11} - \boldsymbol{D}\mathbf{\Sigma}_{21}$, $\mathbf{\Sigma}_{21}$ is the transpose of $\mathbf{\Sigma}_{12}$
- The(i, j)th partial correlation θ_{ij} is the correlation in the covariance matrix $\mathbf{\Sigma}_{ij}$, i.e. of the diagonal term divided by square roots of the diagonal terms.