Properties of Multivariate Gaussian Distributions

A Gaussian variable $X \sim N(\mu, \Sigma)$, where μ is the mean and Σ is the covariance matrix, has the following probability density function:

$$P(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right)$$

where $|\Sigma|$ is the determinant of Σ

The Gaussian distribution occurs frequently in real world data. This is due to the **Central Limit Theorem.**

The CLT states that the arithmetic mean of $m > 0$ samples is approximately normally distributed - independent of the original sample distribution (provided finite mean/variance).

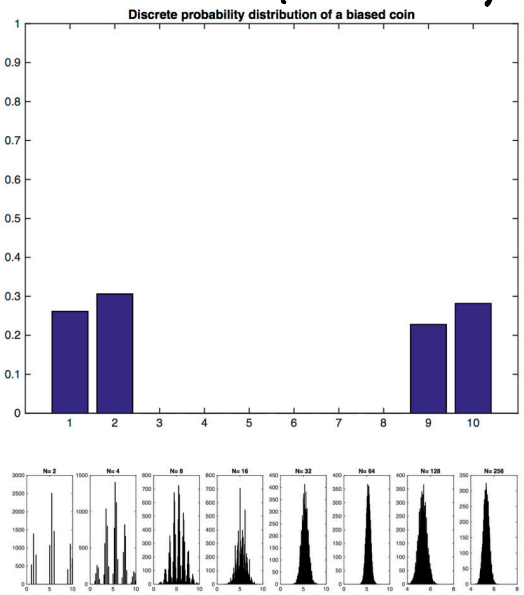

Illustration of the Central Limit Theorem: In the above graph, random variables Y are drawn from the distribution illustrated through the bar plot. Values of 1,2 and 9,10 are likely, but 3-8 have no support. This distribution does not look Gaussian at all. However, their sample means, $\bar{Y}_j = \frac{1}{N}\sum_{i=1}^{N} y_i$, of which we sample m: $\bar{Y}_1, \ldots, \bar{Y}_m$, become very "Gaussian" distributed as N increases (lower plot).

Once Gaussian, Always Gaussian

Let Gaussian variable

$$y = \begin{bmatrix} y_A \\ y_B \end{bmatrix}, \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$$

① Normalization

$$\int_y p(y; \mu, \Sigma)\, dy = 1$$

② Marginilization

The marginal distribution

$$p(y_A) = \int_{y_B} p(y_A, y_B; \mu, \Sigma)\, dy_B$$

and

$$p(y_B) = \int_{y_A} p(y_A, y_B; \mu, \Sigma)\, dy_A$$

are Gaussian!

$$y_A \sim N(\mu_A, \Sigma_{AA})$$

$$y_B \sim N(\mu_B, \Sigma_{BB})$$

③ Summation

If $y \sim N(\mu, \Sigma)$ and $y' \sim N(\mu', \Sigma')$ then

$$y + y' \sim N(\mu + \mu', \Sigma + \Sigma')$$

(4) Conditioning

The conditional distribution of y_A on y_B

$$p(y_A | y_B) = \frac{p(y_A, y_B ; \mu, \Sigma)}{\int_{y_A} p(y_A, y_B ; \mu, \Sigma) \, dy_A}$$

is also Gaussian!

Note

$$y_A | y_B = y_B \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(y_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

Gaussian Process Regression

Consider a regression problem

$$y = f(\underline{x}) + \varepsilon$$

$$y = \underline{w}^T \underline{x} + \varepsilon \qquad \text{(OLS and ridge regression)}$$

$$y = \underline{w}^T \phi(\underline{x}) + \varepsilon \qquad \text{(kernel ridge regression)}$$

In general, the posterior distribution is

$$P(Y | D, X) = \int_{\bar{w}} P(Y, \bar{w} | D, X) \, d\bar{w} = \int_{\bar{w}} P(Y | \bar{w}, D, X) P(\bar{w} | D) \, d\bar{w}$$

The above is often <u>intractable</u> in closed form.

However, for the special case of having a Gaussian likelihood and prior (ridge regression assumptions), the expression is Gaussian and we can derive its mean and covariance.

So,

$$P(y_* \mid D, \bar{x}) \sim \mathcal{N}(\mu_{y|D}, \Sigma_{y|D})$$

where

$$\mu_{y_*|D} = K_*^T (K + \sigma^2 I)^{-1} y$$

and

$$\Sigma_{y_*|D} = K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_*$$

So, instead of doing **MAP** (as in ridge regression) let's model the entire distribution and let's forget about $\bar{w}$ and the kernel trick by modeling f directly (instead of y)!

Gaussian Process

Problem: f is an infinite dimensional function! But, the multivariate Gaussian distribution is for **finite** dimensional random vectors.

Definition: A Gaussian Process is a (potentially infinite) collection of random variables such that the joint distribution of every finite subset of RVs is multivariate Gaussian:

$$f \sim GP(\mu, k),$$

where $\mu(\vec{x})$ and $k(\vec{x}, \vec{x}')$ are the mean resp. covariance function!

In order to model the predictive distribution $P(f_* | x_*, D)$ we can use a Bayesian approach by using a **GP prior**: $P(f | X) \sim N(\mu, \Sigma)$ and condition it on the training data D to model the joint distribution $f = f(X)$ [vector of training observations] and $f_* = f(\vec{x}_*)$ [prediction at test input]

Gaussian Process Regression

We assume that, before we observe the training labels, the labels are drawn from the zero-mean prior Gaussian distribution:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ y_t \end{bmatrix} \sim N(0, \Sigma)$$

All training and test labels are drawn an $(n+m)$-dimension Gaussian distribution, where

$n \rightarrow$ number of training points

$m \rightarrow$ number of testing points

Note: real training labels $y_1, \ldots, y_n$ we observe are samples of $Y_1, \ldots, Y_n$

Whether this distribution gives us meaningful distribution or not depends on how we choose the covariance matrix Σ.

We consider the following properties of Σ:

① $\Sigma_{ij} = \mathbb{E}\left[(Y_i - \mu_i)(Y_j - \mu_j)\right]$

② Σ is **ALWAYS** positive semi-definite

③ $\Sigma_{ii} = \text{Variance}(Y_i)$, thus $\Sigma_{ii} \geq 0$

④ If Y_i and Y_j are VERY independent then $\Sigma_{ij} = \Sigma_{ji} = 0$ (i.e x_i is VERY different from x_j)

⑤ If x_i is similar to x_j, then $\Sigma_{ij} = \Sigma_{ji} > 0$

We can observe that this is very similar to the kernel matrix in SVMs.

Therefore, we can simply let $\Sigma_{ij} = K(x_i, x_j)$.

For example, if we use RBF kernel, then

$$\Sigma_{ij} = \tau e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}}$$

if we use polynomial kernel, then

$$\Sigma_{ij} = \tau(1 + x_i^T x_j)^d$$

etc.

Thus, we can decompose Σ as $\begin{pmatrix} K, & K_* \\ K_*^T, & K_{**} \end{pmatrix}$, where

- $K \rightarrow$ training kernel matrix
- $K_* \rightarrow$ training-testing kernel matrix
- $K_*^T \rightarrow$ testing-training kernel matrix
- $K_{**} \rightarrow$ testing kernel matrix

The conditional distribution of (noise-free) values of the latent function f can be written as:

$$f_* | (Y_1 = y_1, \ldots, Y_n = y_n, \bar{x}_1, \ldots, \bar{x}_n, \bar{x}_t) \sim N(K_*^T K^{-1} y, K_{**} - K_*^T K^{-1} K_*),$$

where the kernel matrices K_*, K_{**}, K are functions of $\bar{x}_1, \ldots, \bar{x}_n, \bar{x}_*$.

Additive Gaussian Noise

In most applications, the observed labels can be noisy.

If we assume this noise is independent and zero-mean Gaussian, then we observe

$$\hat{Y}_i = f_i + \varepsilon_i$$

where f_i is the true (unobserved) target and the noise is denoted by

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this case the new covariance matrix becomes

$$\hat{\Sigma} = \Sigma + \sigma^2 I$$

To see this, observe that for non-diagonal entries

$$\hat{\Sigma}_{ij} = \mathbb{E}[(f_i+\varepsilon_i)(f_j+\varepsilon_j)] = \mathbb{E}[f_i f_j] + \mathbb{E}[f_i]\mathbb{E}[\varepsilon_j] + \mathbb{E}[f_j]\mathbb{E}[\varepsilon_i] + \mathbb{E}[\varepsilon_i]\mathbb{E}[\varepsilon_j]$$

$$= \mathbb{E}[f_i f_j] = \Sigma_{ij}$$

and for diagonal entries

$$\hat{\Sigma}_{ii} = \mathbb{E}[(f_i+\varepsilon_i)^2] = \mathbb{E}[f_i^2] + 2\mathbb{E}[f_i]\mathbb{E}[\varepsilon_i] + \mathbb{E}[\varepsilon_i^2] = \mathbb{E}[f_i^2] + \mathbb{E}[\varepsilon_i^2]$$

$$= \Sigma_{ii} + \sigma^2$$

Plugging this updated covariance matrix into the Gaussian Process posterior distribution leads to

$$Y_* \mid (Y_1 = y_1, \ldots, Y_n = y_n, \vec{x}_1, \ldots, \vec{x}_n) \sim \cdots$$

$$\cdots \sim \mathcal{N}\left(K_*^T (K + \sigma^2 I)^{-1} y, (K_{**} + \sigma^2 I) - K_*^T (K + \sigma^2 I)^{-1} K_*\right)$$