Multivariate Gaussian

Probabilistic Approach to Linear Regression

- X: random variable
  $p(x)$: probability density function of X
  $\iff \quad p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1 \quad$ and

$$P[a \leq X \leq b] = \int_a^b p(x)dx$$

- For example, the *normal* random variable with mean μ and variance σ^2 has the density function

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{(x-\mu)^2}{2\sigma^2}\right).$$

  In this case, we write $X \sim \mathcal{N}(\mu, \sigma^2)$ and

$$p(x) = \mathcal{N}(x|\mu, \sigma^2).$$

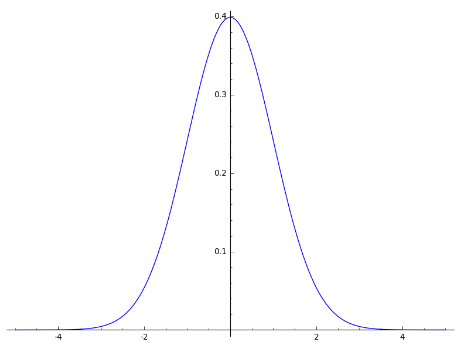

Figure 1: Graph of $\mathcal{N}(x|0, 1)$

- X, Y: two random variables
  $p(x, y)$: (joint) probability density function
  $\iff \quad p(x, y) \geq 0, \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)dxdy = 1 \quad$ and

$$P[(X, Y) \in A] = \iint_A p(x, y)dxdy$$

- Marginal density functions:

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y)\, dy \quad \text{and} \quad p_Y(y) = \int_{-\infty}^{\infty} p(x, y)\, dx$$

- The *covariance* of X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).$$

- The *conditional density* of X given that $Y = y$ is defined to be

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)} = \frac{p(x, y)}{\int p(u, y)du}.$$

- More generally, we consider

$$T, X_1, \ldots, X_m : \text{ random variables.}$$

  Let $\boldsymbol{x} = (x_1, \ldots, x_m)$. Then we have
  $p(t, \boldsymbol{x})$: probability density,
  $p(t|\boldsymbol{x})$: conditional density

- Given random variables $X_1, \ldots, X_m$, the *covariance matrix* is defined to be

$$\Sigma = [\mathrm{Cov}(X_i, X_j)].$$

Recall the settings of linear regression.

- Input: x Output: t
  Observations: $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$

- In many applications we expect some noise in determining the output, and the following assumption is reasonable.

- Assume that given x, the corresponding value of t has a normal distribution with a mean equal to the value $y(x, \boldsymbol{w})$ of the polynomial curve

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_D x^D,$$

  where $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]^\top$.

- That is to say,

$$t = y(x, \boldsymbol{w}) + \epsilon,$$

  where ϵ is a Gaussian noise with variance σ^2. Then we can write

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}),$$

  where $\beta = 1/\sigma^2$ is the inverse variance, called *precision*.

- Given $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\mathbf{t} = (t_1, \ldots, t_N)$, we have

$$p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}).$$

- Let $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\mathbf{t} = (t_1, \ldots, t_N)$ be given.
  Task: Determine $\boldsymbol{w}$ and β by maximum likelihood.
  This is a probabilistic approach to the regression problem.

- We have

$$p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}).$$

To maximize this function, we take logarithm:

$$\ln p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

{#eq:log}

Exercise: Verify equality +@eq:log.

- Thus maximizing likelihood with respect to $\boldsymbol{w}$ is equivalent to minimizing the error function $E(\boldsymbol{w})$:

$$E(\boldsymbol{w}) = \tfrac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2.$$

Thus this probabilistic approach leads to the same computation as the usual linear regression to determine $\boldsymbol{w}$. Nevertheless, we can also determine the parameter β to get maximum likelihood as follow.

- After finding out $\boldsymbol{w}_{\mathrm{ML}}$, take the derivative with respect to β to obtain

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}_{\mathrm{ML}}) - t_n\}^2.$$

- Finally, the *predictive distribution* is given by

$$\boxed{p(t|x, \boldsymbol{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}(t|y(x, \boldsymbol{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1})}.$$

!!Example or code??

Bayesian Linear Regression

- Bayesian linear regression avoids the over-fitting problem of maximum likelihood.

- We need multi-dimensional normal distributions.

Recall one-dimensional normal distribution:

$$p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{(x-\mu)^2}{2\sigma^2}\right).$$

- D-dimensional Gaussian distribution:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

  where the D-dimensional vector $\boldsymbol{\mu}$ is the mean, the $D \times D$ matrix Σ is the covariance, and $|\Sigma|$ is the determinant of Σ.

- Assume

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda^{-1}),$$
$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|A\boldsymbol{x} + \boldsymbol{b}, L^{-1}).$$

- Then we have

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|A\,\boldsymbol{\mu} + \boldsymbol{b}, L^{-1} + A\,\Lambda^{-1}A^\top),$$
$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\Sigma\,\{A^\top L(\boldsymbol{y}-\boldsymbol{b}) + \Lambda\,\boldsymbol{\mu}\}, \Sigma),$$

  where $\Sigma = (\Lambda + A^\top LA)^{-1}$.

!! Do we need to verify this??

Recall the settings of linear regression.

- Input: x; Output: t
  Observations: $(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)$
- Assume that given x, the corresponding value of t has a normal distribution with a mean equal to the value $y(x, \boldsymbol{w})$ of the polynomial curve

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_D x^D,$$

  where $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]^\top$.

- Consider a prior distribution for $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}I).$$

  Note that we are taking the initial vector for $\boldsymbol{w}$ to be the zero vector $\mathbf{0}$.

- Recall that we have

$$p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}).$$

- Bayes' Theorem says

$$(\text{posterior}) \propto (\text{likelihood}) \times (\text{prior}).$$

  In our situation, it becomes

$$p(\boldsymbol{w}|\boldsymbol{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\alpha).$$

  Task: Given the data, determine $\boldsymbol{w}$ so that the posterior is maximized. This process is called *maximum posterior* (MAP).

- Take the negative logarithm of the posterior

$$\begin{aligned}
&-\ln p(\boldsymbol{w}|\boldsymbol{x}, \mathbf{t}, \alpha, \beta) \\
&\qquad = -\ln\left[p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\alpha)\right] + \text{constant} \\
&\qquad = \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2} \boldsymbol{w}^\top \boldsymbol{w} + \text{constants}
\end{aligned}$$

- The maximum of the posterior is given by the minimum of

$$\tilde{E}(\boldsymbol{w}) = \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2} \boldsymbol{w}^\top \boldsymbol{w}.$$

  Thus maximizing the posterior distribution is equivalent to minimizing the *regularized* sum-of-square error function.

- Let $\phi_i(x) = x^i$ and

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_N) \\ \phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_2(x_N) \\ \vdots & \vdots & & \vdots \\ \phi_{M-1}(x_1) & \phi_{M-1}(x_2) & \cdots & \phi_{M-1}(x_N) \end{bmatrix}.$$

  Then

$$\tilde{E}(\boldsymbol{w}) = \frac{\beta}{2} \|X^\top \boldsymbol{w} - \mathbf{t}\|^2 + \frac{\alpha}{2} \boldsymbol{w}^\top \boldsymbol{w},$$

  and

$$\nabla \tilde{E}(\boldsymbol{w}) = \beta X (X^\top \boldsymbol{w} - \mathbf{t}) + \alpha \boldsymbol{w} = 0.$$

  Thus

$$\boldsymbol{w} = \beta S X \mathbf{t} \quad \text{with} \quad S^{-1} = \beta X X^\top + \alpha I.$$

We can choose values of the parameters α and β.

Recall the maximum likelihood gave us

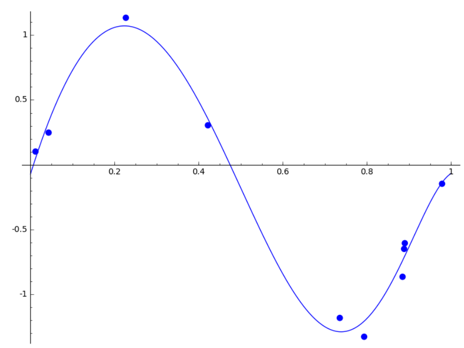

Figure 2: $N = 9$, $\alpha = 0.01$, $\beta = 1000$

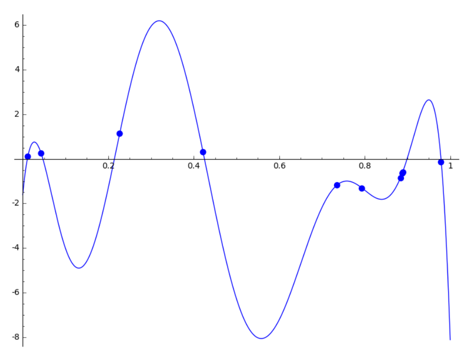

Figure 3: Over-fitting

Predictive distribution

- The posterior can be computed explicitly, since the prior and the likelihood are all Gaussian. Indeed, we obtain

$$p(\boldsymbol{w}|\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|m_N, S_N),$$

where

$$m_N = \beta S_N X \mathbf{t} \quad \text{and} \quad S_N^{-1} = \alpha I + \beta X X^\top.$$

- Furthermore, we can compute the predictive distribution $p(t|x, \boldsymbol{x}, \mathbf{t})$. Assume that α and β are fixed. Then the predictive distribution is given by

$$p(t|x, \boldsymbol{x}, \mathbf{t}) = \int p(t|x, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{x}, \mathbf{t}) d\boldsymbol{w},$$

where

$$p(t|x, \boldsymbol{w}) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}).$$

- One can compute the integral to obtain

$$p(t|x, \boldsymbol{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x)),$$

where

$$\begin{aligned}
m(x) &= \beta \boldsymbol{\phi}(x)^\top S \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n, \\
s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^\top S \boldsymbol{\phi}(x), \\
S^{-1} &= \alpha I + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^\top, \\
\boldsymbol{\phi}(x) &= [1, \phi_1(x), \phi_2(x), \ldots, \phi_M(x)]^\top, \quad \phi_i(x) = x^i.
\end{aligned}$$