

Linear Models for Regression & Classification and Sparse Kernel Machines

Lecturer: Prof. Ko Nishino

Scribe: L.Kratz & P.Bariya

1 Bayes' theorem for Gaussian variables

Let us consider a Gaussian marginal distribution

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \quad (1)$$

and a Gaussian conditional distribution

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}), \quad (2)$$

where, $\boldsymbol{\mu}$, \mathbf{A} , \mathbf{b} are parameters that govern the means, and $\boldsymbol{\Lambda}$ and \mathbf{L} are the precision matrices respectively. The conditional distribution $p(\mathbf{y}|\mathbf{x})$ has a mean that is a linear function of \mathbf{x} and a covariance which is independent of \mathbf{x} .

Now, let us define \mathbf{z} as:

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}^T & \mathbf{y}^T \end{pmatrix}^T$$

The joint probability distribution over \mathbf{x} and \mathbf{y} is given by:

$$\begin{aligned} p(\mathbf{z}) &= p(\mathbf{x}, \mathbf{y}) \\ &= p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \end{aligned}$$

Then,

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) + \text{const.} \end{aligned} \quad (3)$$

1.1 Covariance

To find the precision of this Gaussian distribution, we will consider the second-order terms of the above equation 3:

$$\begin{aligned} & -\frac{1}{2}\mathbf{x}^T(\mathbf{A}^T\mathbf{L}\mathbf{A} + \boldsymbol{\Lambda})\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{y} \\ &= -\frac{1}{2}\begin{pmatrix} \mathbf{x}^T \\ \mathbf{y}^T \end{pmatrix} \begin{bmatrix} \mathbf{A}^T\mathbf{L}\mathbf{A} + \boldsymbol{\Lambda} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\ &= -\frac{1}{2}\mathbf{z}^T\mathbf{R}\mathbf{z}. \end{aligned}$$

Here, the precision (inverse covariance) of the distribution is given by

$$\mathbf{R} = \begin{bmatrix} \mathbf{A}^T \mathbf{L} \mathbf{A} + \mathbf{\Lambda} & -\mathbf{A}^T \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}. \quad (4)$$

Then, the covariance is given by

$$\text{cov}[x, y] = \mathbf{R}^{-1} = \begin{bmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1} \mathbf{A}^T \\ \mathbf{A} \mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T \end{bmatrix}. \quad (5)$$

1.2 Mean

To find the mean of this Gaussian distribution, we will consider the first-order terms of the above equation 3:

$$\begin{aligned} & \mathbf{x}^T \mathbf{\Lambda} \boldsymbol{\mu} - \mathbf{x}^T \mathbf{A}^T \mathbf{L} \mathbf{b} + \mathbf{y}^T \mathbf{L} \mathbf{b} \\ &= \begin{pmatrix} \mathbf{x}^T \\ \mathbf{y}^T \end{pmatrix} \begin{bmatrix} \mathbf{\Lambda} \boldsymbol{\mu} - \mathbf{A}^T \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{bmatrix}. \end{aligned}$$

Comparing this with the linear term inside the exponent obtained after completing the square over the quadratic form of a multivariate Gaussian, i.e. $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, the mean of the distribution is given by

$$\begin{aligned} \mathbb{E}[\mathbf{z}] &= \mathbf{R}^{-1} \begin{bmatrix} \mathbf{\Lambda} \boldsymbol{\mu} - \mathbf{A}^T \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1} \mathbf{A}^T \\ \mathbf{A} \mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} \boldsymbol{\mu} - \mathbf{A}^T \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A} \boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \end{aligned} \quad (6)$$

1.3

The mean and covariance of the marginal distribution $p(\mathbf{y})$ is given by

$$\mathbb{E}[\mathbf{y}] = \mathbf{A} \boldsymbol{\mu} + \mathbf{b},$$

and

$$\text{cov}[\mathbf{y}] = \mathbf{L}^{-1} + \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T.$$

The marginal distribution $p(\mathbf{y})$ is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T). \quad (7)$$

The mean and covariance of the conditional distribution $p(\mathbf{x} | \mathbf{y})$ is given by

$$\mathbb{E}[\mathbf{x} | \mathbf{y}] = (\mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1} \{ \mathbf{A}^T \mathbf{L} (\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda} \boldsymbol{\mu} \},$$

and

$$\text{cov}[\mathbf{x} | \mathbf{y}] = (\mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1}.$$

The conditional distribution of x given y , $p(\mathbf{x} | \mathbf{y})$, is given by

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\Sigma} \{ \mathbf{A}^T \mathbf{L} (\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda} \boldsymbol{\mu} \}, \boldsymbol{\Sigma}) \quad (8)$$

where,

$$\boldsymbol{\Sigma} = (\mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1}.$$

2 Maximum Margin Classifiers

Consider a two-class classification problem using linear models of the form

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b, \quad (9)$$

where $\phi(\mathbf{x})$ is the feature-space transformation.

The training data set contains N input vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$ with corresponding target values t_1, \dots, t_N where $t_n \in \{-1, 1\}$.

For all training data:

$$t_n y(\mathbf{x}_n) > 0,$$

i.e. $y(\mathbf{x}_n) > 0$ for points having $t_n = +1$ and $y(\mathbf{x}_n) < 0$ for points having $t_n = -1$.

The objective here is to maximize the margin, which is the smallest distance between the decision boundary and any of the samples. The distance of any point \mathbf{x}_n to the decision surface is given by

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}.$$

The maximum margin solution is given by

$$\operatorname{argmax}_{\mathbf{w}, b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n [t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b)] \right\}$$

For the point closest to the decision surface

$$t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b) = 1.$$

And the canonical representation of the decision hyperplane is given by

$$t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1, \quad n = 1, \dots, N \quad (10)$$

where equality holds for *active* data points.

For optimization, we need to maximize $\|\mathbf{w}\|^{-1}$, which is equivalent to

$$\operatorname{argmin}_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

subject to the constraints given above by equation 10.

Lagrange multipliers $a_n \geq 0$ are introduced, giving the Lagrangian function

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N a_n \{t_n (\mathbf{w}^T \phi(\mathbf{x}_n) + b) - 1\} \quad (11)$$

where $\mathbf{a} = (a_1, \dots, a_N)^T$.

Then, setting $\frac{\partial L}{\partial \mathbf{w}} = 0$ and $\frac{\partial L}{\partial b} = 0$, we get

$$\mathbf{w} = \sum_{n=1}^N a_n t_n \phi(\mathbf{x}_n) \quad (12)$$

$$0 = \sum_{n=1}^N a_n t_n. \quad (13)$$

Eliminating \mathbf{w} and b from equation 11 using these conditions, we get the dual representation of the maximum margin problem

$$\begin{aligned}
\tilde{L}(\mathbf{a}) &= \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m) - \sum_{n=1}^N \sum_{m=1}^N a_n \left\{ t_n (a_m t_m \phi(\mathbf{x}_m)^T \phi(\mathbf{x}_n) + b) - 1 \right\} \\
&= -\frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m) - \sum_{n=1}^N \sum_{m=1}^N a_n t_n b + \sum_{n=1}^N a_n \\
&\quad \left[\text{since, } \sum_{n=1}^N \sum_{m=1}^N a_n t_n = 0 \right] \\
&= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m) \\
&= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)
\end{aligned}$$

subject to the constraints

$$\begin{aligned}
a_n &\geq 0, (n = 1, \dots, N) \\
\sum_{n=1}^N a_n t_n &= 0.
\end{aligned}$$

And the kernel function is defined by

$$k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m).$$