Machine Learning 1 - Homework Week 1

Minh Ngo 10897402[1]

September 11, 2015

[1] University of Amsterdam
minh.ngole@student.uva.nl

1.1

1

Following random variables can be defined:

1. R = {r, nr}, where **R** - random variable that defines a weather, **r** - raining, **nr** - not raining.
2. P = {Amsterdam, Rotterdam} - random variable that defines a place where you are staying.

From the task, we can determine following probabilities:

1. Raining in Amsterdam: $p(R = r|P = Amsterdam) = 0.5$
2. Raining in Rotterdam: $p(R = r|P = Rotterdam) = 0.75$
3. Being in Amsterdam: $p(P = Amsterdam) = 0.8$
4. Being in Rotterdam: $p(P = Rotterdam) = 0.2$

2

$p(P = Rotterdam|R = nr) = 1-p(P = Rotterdam|R = r) = 1-0.75 = 0.25$

3

According to the sum rule of probability:

$$\begin{aligned}
p(R=r) &= p(R=r|P=Amsterdam) \times p(P=Amsterdam)+ \\
&\quad p(R=r|P=Rotterdam) \times p(P=Rotterdam) \\
&= 0.8 * 0.5 + 0.2 * 0.75 = 0.55
\end{aligned}$$

4

According to the Bayes' theorem:

$$\begin{aligned}
p(P=Amsterdam|R=r) &= \frac{p(R=r|P=Amsterdam)p(P=Amsterdam)}{p(R=r)} \\
&= \frac{0.5 \times 0.8}{0.55} = 0.727
\end{aligned}$$

1.2

1

Population of the city **N** = 500000. Estimated number of people that have cancer **c** = 500.

$$p(cancer) = \frac{c}{N} = \frac{500}{500000} = 0.001$$
$$p(not\ cancer) = 1 - p(cancer) = 0.999$$

2

For this task we can build a confusion matrix of size 2x2 that describe the behavior of the blood test:

		Predict	
		Cancer	Not cancer
Actual	Cancer	99	1
	Not cancer	5	95

$$p(has\ cancer) = \frac{c(true\ positives)}{c(true\ positives) + c(false\ negatives)} = \frac{99}{99+5} = 0.95$$

3

We assume that cancer & not cancer predictions are made independently by the blood test.

1.3

1

$$p(\theta, D) = p(D|\theta)p(\theta) \qquad p(\theta, D) = p(D, \theta)$$
$$\Rightarrow p(D|\theta)p(\theta) = p(\theta|D)p(D)$$

$$\Leftrightarrow p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\sum_\theta p(\theta, D)} = \frac{p(D|\theta)p(\theta)}{\sum_\theta p(D|\theta)p(\theta)}$$

We call $p(\theta)$ as a prior, $p(D|\theta)$ as a likelihood, $p(D)$ as an evidence, $p(\theta|D)$ as a posterior.

2

For our example:

$$p(D|\theta) = p(x_1, .., x_n|\mu, \sigma^2)$$

$$p(D|\theta) = p(x|\mu, \sigma^2) = N(x|\mu, \sigma^2) = \prod_{i=1}^{n} p(x_i|\mu, \sigma^2) = \prod_{i=1}^{n} N(x_i|\mu, \sigma^2)$$

$$p(\theta) = N(\mu|\mu_0, \sigma^2)$$

$$\Leftrightarrow p(\theta, D) = \frac{(\prod_{i=1}^{n} N(x_i|\mu, \sigma^2))N(\mu|\mu_0, \sigma^2)}{\int(\prod_{i=1}^{n} N(x_i|\mu, \sigma^2))N(\mu|\mu_0, \sigma^2)d\mu}$$

2.1

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ 2 & 3 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 9 \\ 5 \end{pmatrix}$$

1

$$\mathbf{Ab} = \begin{pmatrix} 3 \times 9 + 5 \times 5 \\ 2 \times 9 + 3 \times 5 \end{pmatrix} = \begin{pmatrix} 52 \\ 33 \end{pmatrix}$$

2

$$\mathbf{b^T A} = \begin{pmatrix} 9 & 5 \end{pmatrix} \times \begin{pmatrix} 3 & 5 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 9 \times 3 + 5 \times 2 & 9 \times 5 + 5 \times 3 \end{pmatrix} = \begin{pmatrix} 37 & 60 \end{pmatrix}$$

3

$$\mathbf{Ac} = \mathbf{b}$$

We can use Gaussian elimination to find $\mathbf{c}$:

$$\left(\begin{array}{cc|c} 3 & 5 & 9 \\ 2 & 3 & 5 \end{array}\right) \rightarrow \left(\begin{array}{cc|c} 6 & 10 & 18 \\ 6 & 9 & 15 \end{array}\right) \rightarrow \left(\begin{array}{cc|c} 6 & 10 & 18 \\ 0 & -1 & -3 \end{array}\right)$$

$$c_2 = 3 \qquad c_1 = \frac{18 - 10 \times 3}{6} = -2 \Leftrightarrow \mathbf{c} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$$

4

$$\mathbf{A^{-1}} = \frac{1}{det(\mathbf{A})} adj(\mathbf{A})$$

$$adj(\mathbf{A}) = \begin{pmatrix} 3 & -5 \\ -2 & 3 \end{pmatrix} \qquad det(\mathbf{A}) = 3 \times 3 - 2 \times 5 = -1$$

$$\Rightarrow \mathbf{A^{-1}} = -1 \times \begin{pmatrix} 3 & -5 \\ -2 & 3 \end{pmatrix} = \begin{pmatrix} -3 & 5 \\ 2 & -3 \end{pmatrix}$$

5

$$\mathbf{A^{-1}b} = \begin{pmatrix} -3 & 5 \\ 2 & -3 \end{pmatrix} \ times \begin{pmatrix} 9 \\ 5 \end{pmatrix} = \begin{pmatrix} -3 \times 9 + 5 \times 5 \\ 18 - 15 \end{pmatrix} = \begin{pmatrix} -2 \\ 3 \end{pmatrix} = \mathbf{c} \qquad q.e.d.$$

2.2

By definition, $\bigtriangledown f = \frac{\partial f}{\partial x_1}\mathbf{e_1} + .. + \frac{\partial f}{\partial x_n}\mathbf{e_n}$, where $\{\mathbf{e_i}\}_{i=1}^{n}$ are standard unit vectors.

1. $f(x) = x^2 + 2x + 3 \Rightarrow \bigtriangledown f = (2x + 2)\mathbf{i}$
2. $g(x) = (2x^3 + 1)^2 => \bigtriangledown g = (2(2x^3 + 1)6x^2)\mathbf{i} = 12x^2(2x^3 + 1)\mathbf{i}$

Partial derivatives:

1. $f(x, y, z) = (x + 2y)^2 sin(xy)$

$$\frac{\partial f}{\partial x} = 2(x + 2y)sin(xy) + (x + 2y)^2 cos(xy)y$$

$$\frac{\partial f}{\partial y} = 4(x + 2y)sin(xy) + (x + 2y)^2 cos(xy)x$$

$$\frac{\partial f}{\partial z} = 0$$

2. $f(x, y, z) = 2log(x + y^2 - z)$

$$\frac{\partial f}{\partial x} = \frac{2}{x + y^2 - z}$$

$$\frac{\partial f}{\partial y} = \frac{4y}{x + y^2 - z}$$

$$\frac{\partial f}{\partial z} = -\frac{2}{x + y^2 - z}$$

3. $f(x, y, z) = exp(x \cos(y + z))$

$$\frac{\partial f}{\partial x} = exp(x \cos(y + z)) \cos(y + z)$$

$$\frac{\partial f}{\partial y} = exp(x \cos(y + z))(-x \sin(y + z))$$

$$\frac{\partial f}{\partial z} = -exp(x \cos(y + z))(-x \sin(y + z))$$

2.3

1.

$$\begin{aligned}
\mathbf{T} &= (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
&= (\mathbf{x}^T - \boldsymbol{\mu}^T) \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + (\boldsymbol{\mu}^T - \boldsymbol{\mu}_0^T) \mathbf{S}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
&= \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \\
&\quad \boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^T \mathbf{S}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}_0^T \mathbf{S}^{-1} \boldsymbol{\mu}_0 \\
&= \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \\
&\quad \boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T \mathbf{S}^{-1} \boldsymbol{\mu}_0
\end{aligned}$$

2.

$$\begin{aligned}
\frac{\partial \mathbf{T}}{\partial \boldsymbol{\mu}} &= \frac{\partial \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}}{\partial \boldsymbol{\mu}} - \frac{\partial 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{\partial \boldsymbol{\mu}} + \frac{\partial \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{\partial \boldsymbol{\mu}} + \\
&\quad \frac{\partial \boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu}}{\partial \boldsymbol{\mu}} - \frac{\partial 2\boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu}_0}{\partial \boldsymbol{\mu}} + \frac{\partial \boldsymbol{\mu}_0^T \mathbf{S}^{-1} \boldsymbol{\mu}_0}{\partial \boldsymbol{\mu}} \\
&= 0 - 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^T) \boldsymbol{\mu} + \\
&\quad (\mathbf{S}^{-1} + (\mathbf{S}^{-1})^T) \boldsymbol{\mu} - 2\mathbf{S}^{-1} \boldsymbol{\mu}_0^T + 0 \\
&= -2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} + 2\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + 2\mathbf{S}^{-1} \boldsymbol{\mu} - 2\mathbf{S}^{-1} \boldsymbol{\mu}_0^T
\end{aligned}$$

3.

$$\frac{\partial \mathbf{T}}{\partial \boldsymbol{\mu}} = 0$$

$$\Leftrightarrow -2\mathbf{x}^T\boldsymbol{\Sigma}^{-1} + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\mathbf{S}^{-1}\boldsymbol{\mu} - 2\mathbf{S}^{-1}\boldsymbol{\mu}_0^T = 0$$

$$\Leftrightarrow \mu = (\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1})^{-1}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\boldsymbol{\mu}_0^T)$$

3

1

In the case of a single variable x, the Gaussian distribution is represented as following:

$$N(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\bigg(-\frac{1}{2\sigma^2}(x-\mu)^2 \bigg)$$

Let $\mathbf{t} = \{t_i\}_{i=1}^N \qquad \boldsymbol{\Phi} = \{\boldsymbol{\phi}_i\}_{i=1}^K \qquad \mathbf{X} = \{x_i\}_{i=1}^N$.

$$\begin{aligned}
p(\mathbf{X}|\mathbf{w}) &= \prod_{i=1}^N N(t_i|\mathbf{w}^T\boldsymbol{\phi}_i, \frac{1}{\beta}) \\
&= \prod_{i=1}^N \frac{1}{\sqrt{\frac{2\pi}{\beta}}} exp\bigg(-\frac{1}{2/\beta}(t_i - \mathbf{w}^T\boldsymbol{\phi}_i)^2 \bigg) \\
&= \prod_{i=1}^N \sqrt{\frac{\beta}{2\pi}} exp\bigg(-\frac{\beta}{2}(t_i - \mathbf{w}^T\boldsymbol{\phi}_i)^2 \bigg) \\
&= (\frac{\beta}{2\pi})^{N/2} exp\bigg(-\frac{\beta}{2}\sum_{i=1}^N (t_i - \mathbf{w}^T\boldsymbol{\phi}_i)^2 \bigg) \\
&= (\frac{\beta}{2\pi})^{N/2} exp\bigg(-\frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) \bigg)
\end{aligned}$$

2

The expression for multivariate Gaussian distribution

$$N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} exp\bigg(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \bigg)$$

where $\boldsymbol{\mu}$ is a D-minensional mean vector, $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix, and $|D|$ is its determinant.

$$
\begin{aligned}
\Rightarrow p(\mathbf{w}) &= N(\mathbf{w}|\mathbf{0},\mathbf{I}/\alpha)\\
&= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{I}/\alpha|^{1/2}}exp\Big(-\frac{1}{2}(\mathbf{w}-\mathbf{0})^T(\mathbf{I}/\alpha)^{-1}(\mathbf{w}-\mathbf{0})\Big)\\
&= \Big(\frac{\alpha}{2\pi}\Big)^{D/2}exp\Big(-\frac{\alpha\mathbf{w}^T\mathbf{w}}{2}\Big)
\end{aligned}
$$

$$
\Rightarrow log(p(\mathbf{w})) = \frac{D}{2}log\frac{\alpha}{2\pi}-\frac{\alpha\mathbf{w}^T\mathbf{w}}{2}
$$

3

$$
p(\mathbf{X}) = \sum_w p(\mathbf{X}|\mathbf{w})p(\mathbf{w}) = \int p(\mathbf{t}|\boldsymbol{\phi},\mathbf{w},\beta)p(\mathbf{w})d\mathbf{w}
$$

And posterior over $\mathbf{w}$ is:

$$
p(\mathbf{w}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{X})}
$$

where likelihood, prior and evidence are defined by previous equations.

4

$$
\begin{aligned}
log(p(\mathbf{w}|\mathbf{X})) &= log\frac{p(\mathbf{X}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{X})} = log(p(\mathbf{X}|\mathbf{w}))+log(p(\mathbf{w}))-log(p(\mathbf{X}))\\
&= \frac{N}{2}log\frac{\beta}{2\pi}-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w})+\frac{D}{2}log\frac{\alpha}{2\pi}-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}-log(p(\mathbf{X}))\\
&= -\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w})-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}+C
\end{aligned}
$$

where C is some constant not dependent on $\mathbf{w}$. Finding MAP is simpler because we can neglect an evidence value, that is difficult to compute.

5

To compute the derivative the following equations are needed:

$$
\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{I}\mathbf{X}
$$

$$
\frac{\partial(\mathbf{B}\mathbf{x}+\mathbf{b})^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d})}{\partial\mathbf{x}} = \mathbf{B}^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d})+\mathbf{D}^T\mathbf{C}^T(\mathbf{B}\mathbf{x}+\mathbf{b})
$$

$$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B}+\mathbf{B}^T)\mathbf{x}$$

$$\begin{aligned}
\frac{\partial log(p(\mathbf{w}|\mathbf{X}))}{\partial \mathbf{w}} &= -\frac{\beta}{2}\frac{\partial(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^T(\mathbf{t}-\mathbf{\Phi}\mathbf{w})}{\partial \mathbf{w}} - \frac{\alpha}{2}\frac{\partial \mathbf{w}^T\mathbf{w}}{\partial \mathbf{w}} \\
&= -\frac{\beta}{2}\frac{\partial(\mathbf{\Phi}\mathbf{w}-\mathbf{t})^T(\mathbf{\Phi}\mathbf{w}-\mathbf{t})}{\partial \mathbf{w}} - \frac{\alpha}{2}\frac{\partial \mathbf{w}^T\mathbf{I}\mathbf{w}}{\partial \mathbf{w}} \\
&= -\frac{\beta}{2}(\mathbf{\Phi}^T\mathbf{I}(\mathbf{\Phi}\mathbf{w}-\mathbf{t}) + \mathbf{\Phi}^T\mathbf{I}^T(\mathbf{\Phi}\mathbf{w}-\mathbf{t})) - \alpha\mathbf{I}\mathbf{w} \\
&= -\beta\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{w}-\mathbf{t}) - \alpha\mathbf{I}\mathbf{w} \\
&= \beta\mathbf{\Phi}^T\mathbf{t} - (\beta\mathbf{\Phi}^T\mathbf{\Phi}+\alpha\mathbf{I})\mathbf{w} = 0
\end{aligned}$$

$$\Rightarrow \mathbf{\Phi}^T\mathbf{t} = (\mathbf{\Phi}^T\mathbf{\Phi} + \frac{\alpha}{\beta}\mathbf{I})\mathbf{w}$$

$$\Leftrightarrow \mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi} + \frac{\alpha}{\beta}\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{t} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{t} \quad , where \lambda = \frac{\alpha}{\beta} \quad q.e.d.$$

The derivative $\frac{\partial log(p(\mathbf{w}|\mathbf{X}))}{\partial \mathbf{w}}$ can also be computed with the likelihood represented as a product over N:

$$\begin{aligned}
\frac{\partial log(p(\mathbf{w}|\mathbf{X}))}{\partial \mathbf{w}} &= -\frac{\beta}{2}\sum_{i=1}^{N}\frac{\partial(t_i - \mathbf{w}^T\boldsymbol{\phi}_i)^2}{\partial \mathbf{w}} - \alpha\mathbf{I}\mathbf{w} \\
&= -\beta\sum_{i=1}^{N}(\mathbf{w}^T\boldsymbol{\phi}_i - t_i)\boldsymbol{\phi}_i - \alpha\mathbf{I}\mathbf{w} \\
&= -\beta\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{w}-\mathbf{t})
\end{aligned}$$

6

$\boldsymbol{\phi}_0 = 1$ allows us to have a bias $w_0\boldsymbol{\phi}_0$ parameter to provide a fixed offset in the data. From the expression for multivariate Gaussian distribution presented in the subsection 2 we can rewrite $p(\mathbf{w})$ in such way that the first basis function has its own prior/penalty:

$$p(\mathbf{w}) = \left(\frac{\alpha}{2\pi}\right)^{(D-1)/2} exp\left(-\frac{\alpha(\mathbf{w}^T\mathbf{w}-w_0^2)}{2}\right)\sqrt{\frac{\gamma}{2\pi}}exp\left(-\frac{\gamma w_0^2}{2}\right)$$