# Course on probabilistic graphical models
# Master MVA 2015-2016
# Review exercises
# Solution part II

December 15, 2015

## Bayesian regression

Consider the Gaussian probabilistic conditional model seen in class for linear regression in which given a pair of variables $(X, Y)$ with $X$ taking values in $\mathbb{R}^d$ and $Y$ in $\mathbb{R}$, we model the conditional distribution of $Y$ given $X = x$ by a Gaussian distribution $\mathcal{N}(w^\top x, \sigma^2)$ parametrized by $w$ and $\sigma^2$. Assume that $\sigma^2$ is fixed and $w$ unknown and that the problem of learning the linear regression is approached from a Bayesian point of view, by placing a Gaussian prior distribution on $w$ of the form $\mathcal{N}(0, \tau^2 I_d)$.

*In the text, a single pair variables $(x, Y)$ is considered. I will write the solution in the more general case where we assume that a sample of size $n$ of the form $(x_1, y_1), \dots, (x_n, y_n)$ has been observed and one wants to consider the posterior distribution $p(w|x_1, y_1, \dots, x_n, y_n)$. Since no distribution has been specified for $x_1, \dots, x_n$ and since we are only interested in modeling the conditional distribution of $Y$ given $X = x$, we will treat the observations $x_1, \dots, x_n$ as fixed and therefore as non random quantities. This entails that $p(w|x_1, y_1, \dots, x_n, y_n)$ is in fact the same as $p(w|y_1, \dots, y_n)$ (you can think of $x_1, \dots, x_n$ as acting like fixed parameters). We will write $\mathbf{y} = (y_1, \dots, y_n)^\top$. Normally we should write it with capital letters because it is a random variable but to avoid confusions, let's keep it uncapitalized, while keeping in mind that it is a random variable. Likewise, we will write $\mathbf{X} \in \mathbb{R}^{n \times d}$ the design matrix. This one is fixed. The use of bold letters is just to differentiate the variables that have all the data stacked, from a single observation. Finally we have $w \in \mathbb{R}^d$ which is also a random variable. I will use $\mathbf{w}$ for $w$, because it is strange to have just this one not bold.*

1. Compute the parameters of the joint distribution of $(w, Y)$.

   *So, we will compute the joint distribution of $(\mathbf{w}, \mathbf{y})$. This would reduce to the question asked, if there was only a single observation $\mathbf{y} = y_1 = y$. First note that all the distributions that we are going to manipulate are Gaussian, in particular the joint distribution on $(\mathbf{w}, \mathbf{y})$ is Gaussian. To characterize this distribution, since it is Gaussian, we need to compute its expectation and its covariance matrix. The first thing to notice is that we can write $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$ and independent of $w$.*

   *We start be computing expectations. We know that $\mathbb{E}[\mathbf{w}] = 0$ by construction. Now $\mathbb{E}[\mathbf{y}] = \mathbf{X}\mathbb{E}[\mathbf{w}] + \mathbb{E}[\boldsymbol{\varepsilon}] = 0$.*

   *We have thus proved that*

$$\mathbb{E}\begin{bmatrix}\mathbf{w}\\ \mathbf{y}\end{bmatrix} = 0$$

*Let's consider the covariance matrix. Since the expectation is 0, the covariance matrix is also the matrix of second moments:*

$$\mathbb{E}\begin{bmatrix}\mathbf{w}\mathbf{w}^\top & \mathbf{w}\mathbf{y}^\top \\ \mathbf{y}\mathbf{w}^\top & \mathbf{y}\mathbf{y}^\top\end{bmatrix}$$

*We know that* $\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \tau^2 I_d$. *We have*

$$\mathbb{E}[\mathbf{y}\mathbf{w}^\top] = \mathbb{E}[\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}\mathbf{w}^\top] = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top] + \mathbb{E}[\boldsymbol{\varepsilon}]\mathbb{E}[\mathbf{w}]^\top = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = \tau^2\mathbf{X}.$$

*and*

$$\mathbb{E}[\mathbf{y}\mathbf{y}^\top] = \mathbb{E}[(\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon})(\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon})^\top] = \mathbf{X}\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\mathbf{X}^\top + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \tau^2\mathbf{X}\mathbf{X}^\top + \sigma^2 I_n.$$

*So that finally*

$$\Sigma = \mathbb{E}\begin{bmatrix}\mathbf{w}\mathbf{w}^\top & \mathbf{w}\mathbf{y}^\top \\ \mathbf{y}\mathbf{w}^\top & \mathbf{y}\mathbf{y}^\top\end{bmatrix} = \tau^2\begin{bmatrix} I_d & \mathbf{X}^\top \\ \mathbf{X} & (\mathbf{X}\mathbf{X}^\top + \lambda I_n)\end{bmatrix} \qquad \textit{with} \quad \lambda := \frac{\sigma^2}{\tau^2}.$$

2. Compute de posterior distribution on $w$. *Now that we have computed the joint distribution. We can easily compute the conditional distribution of* $\mathbf{w}$ *given* $\mathbf{y}$. *Indeed denoting* $\Sigma_{\mathbf{w},\mathbf{w}}, \Sigma_{\mathbf{w},\mathbf{y}}, \Sigma_{\mathbf{y},\mathbf{w}}$ *and* $\Sigma_{\mathbf{y},\mathbf{y}}$, *the four blocks of the covariance matrix, reading them in row first order, we have shown in the course that (using the fact that marginal expectations are equal to 0) we have*

$$\mathbb{E}[\mathbf{w} \mid \mathbf{y}] = \Sigma_{\mathbf{w},\mathbf{y}}\Sigma_{\mathbf{y},\mathbf{y}}^{-1}\mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda I_n)^{-1}\mathbf{y} = (\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y},$$

*where the last identity is not completely obvious, can be shown using the matrix inversion lemma for example and was not expected as part of the desired answer. It has however the merit of showing us that we retrieve the same form as for ridge regression.*

*Then for the conditional covariance, we have that*

$$Cov(\mathbf{w}|\mathbf{y}) = \Sigma_{\mathbf{w},\mathbf{w}} - \Sigma_{\mathbf{w},\mathbf{y}}\Sigma_{\mathbf{y},\mathbf{y}}^{-1}\Sigma_{\mathbf{y},\mathbf{w}} = \tau^2\big(I_d - \mathbf{X}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}\big) = \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}.$$

3. Compute the predictive distribution over a new output variable $y'$ given a new input $x'$.

*Finally, for a new* $y'$ *we have* $y' = {x'}^\top\mathbf{w} + \varepsilon'$. *The joint distribution of* $(\mathbf{w}, \mathbf{y}, y')$ *is Gaussian (here implicitly given* $\mathbf{X}$ *and* $x'$*), which entails that the predictive distribution of* $y'$, *which is by definition the marginal distribution of* $y'$ *given the data is* $p(y' \mid \mathbf{y})$. *This last distribution must also be Gaussian, and so, one more time it is characterized by its mean and covariance.*

*But using* $y' = {x'}^\top\mathbf{w} + \varepsilon'$, *we have*

$$\mathbb{E}[y'|\mathbf{y}] = {x'}^\top\mathbb{E}[\mathbf{w}|\mathbf{y}] + \mathbb{E}[\varepsilon'|\mathbf{y}] = {x'}^\top\mathbb{E}[\mathbf{w}|\mathbf{y}] + \mathbb{E}[\varepsilon'] = {x'}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^\top\mathbf{y}.$$

*For the conditional variance, first we have*

$$Var[y'|\mathbf{y}] = Var[\mathbf{w}^\top x'|\mathbf{y}] + Var[\varepsilon'|\mathbf{y}]$$

*because* $\varepsilon' \perp\!\!\!\perp \mathbf{w} \mid \mathbf{y}$ *for the simple reason that* $\varepsilon' \perp\!\!\!\perp (\mathbf{w}, \mathbf{y})$. *and so*

$$Var[y'|\mathbf{y}] = {x'}^\top Cov(\mathbf{w}|\mathbf{y})x' + Var(\varepsilon') = \sigma^2\big({x'}^\top(\mathbf{X}^\top\mathbf{X} + \lambda I_d)^{-1}x' + 1\big)$$