École polytechnique
MAP 534: introduction to machine learning
A. Durmus alain.durmus@polytechnique.edu

3 Bayesian inference and Bayesian linear model

3.1 Gaussian random variables

3.1.1 Reminder

Definition 3.1. *1. For any* $\mathrm{m} \in \mathbb{R}$ *and* $\sigma > 0$*, we denote by* $\mathrm{N}(\mathrm{m}, \sigma^2)$ *the Gaussian distribution on* $\mathbb{R}$ *which has density with respect to the Lebesgue measure given by*

$$x \mapsto (2\pi\sigma^2)^{-1/2} \exp(-(x-\mathrm{m})^2/(2\sigma^2)) \; . \tag{7}$$

We can extend this definition for $\sigma = 0$ *by setting for any* $\mathrm{m} \in \mathbb{R}$, $\mathrm{N}(\mathrm{m}, 0) = \delta_{\mathrm{m}}$*, where* δ_{m} *is the Dirac distribution at* m*.*

2. A real-valued random variable Y *is said to follow a Gaussian distribution if there exists* $\mathrm{m} \in \mathbb{R}$ *and* $\sigma \geq 0$*, such that* Y *has distribution* $\mathrm{N}(\mathrm{m}, \sigma^2)$*.*

3. A random variable X *on* $\mathbb{R}^d$ *is said to be a Gaussian random variable if for any* $t \in \mathbb{R}^d$*, the real-valued random variable* $\langle t, X \rangle$ *follows a Gaussian distribution.*

Proposition 3.2. *The random variable* X *on* $\mathbb{R}^d$ *is Gaussian if and only if there exists* $\mathrm{m} \in \mathbb{R}^d$ *and a semi-definite positive matrix* Σ *such that for any* $t \in \mathbb{R}^d$, $\mathbb{E}[\mathrm{e}^{\mathrm{i}\langle t, X\rangle}] = \exp(\mathrm{i}t\mathrm{m} - (1/2)\langle \Sigma t, t\rangle)$*. In that case,* m *and* Σ *are the mean and covariance matrix of the vector* X*, i.e.* $\mathrm{m} = \mathbb{E}[X]$ *and* $\Sigma = \mathbb{E}\left[X X^{\mathrm{T}}\right]$*. We say then that* X *follows a* d*-dimensional Gaussian distribution with mean* m *and covariance matrix* Σ*, denoted* $\mathrm{N}(\mathrm{m}, \Sigma)$*.*

Proposition 3.3. *Let* X *be* d*-Gaussian random variable with distribution* $\mathrm{N}(\mathrm{m}, \Sigma)$*. Let* $a \in \mathbb{R}^n$ *and* $\mathrm{M} \in \mathbb{R}^{n \times d}$*. Then* $Y = a + \mathrm{M}X$ *is a* n*-dimensional Gaussian random variable with distribution* $\mathrm{N}(a + \mathrm{M}\mathrm{m}, \mathrm{M}\Sigma\mathrm{M}^{\mathrm{T}})$*.*

Proposition 3.4. *Let* X *be* d*-dimensional Gaussian random variable with distribution* $\mathrm{N}(\mathrm{m}, \Sigma)$*. Then if for any* $i \in \{1, \ldots, d\}$*,* X_i *is the* i*-th component of* X*, the family of one dimensional r.v.s* $(X_i)_{i \in \{1,\ldots,d\}}$ *is independent if and only if* $\Sigma_{i,j} = 0$ *for* $i, j \in \{1, \ldots, d\}$*,* $i \neq j$*.*

Proposition 3.5. *Let* X *be a* $\sum_{i=1}^d n_i$*-dimensional Gaussian random variable with distribution* $\mathrm{N}(\mathrm{m}, \Sigma)$*. Then for any* $i \in \{1, \ldots, d\}$*, define* $Z_i = (X_{n_{i-1}+1}, \ldots, X_{n_i})$*, where* $n_0 = 0$ *and* X_i *is the* i*-th component of* X *for* $j \in \{1, \ldots, \sum_{i=1}^d n_i$*. The family of r.v.s* $(Z_i)_{i \in \{1,\ldots,d\}}$ *are independent if and only if* $\Sigma_{i,j} = 0$ *for* $i, j \notin \bigcup_{i=1}^d \{n_{i-1} + 1, \ldots n_i\}$*.*

Theorem 3.6 (Cochran's theorem)**.** *Let* X *be a* d*-dimensional Gaussian random variable with distribution* $\mathrm{N}(0, \mathrm{I}_d)$*. Consider a decomposition of* $\mathbb{R}^d$ *as an orthogonal direct sum of* p *orthogonal subspaces* $(\mathsf{E}_i)_{i \in \{1,\ldots,p\}}$ *such that for any* $i \in \{1, \ldots, p\}$*, the dimension of* E_i *is equal to* n_i*. Then let* $(\mathrm{proj}\,(\cdot \,|\, \mathsf{E}_i))_{i \in \{1,\ldots,p\}}$ *be the associated orthogonal projections and* $Y_i = \mathrm{proj}\,(X \,|\, \mathsf{E}_i)$ *for any* $i \in \{1, \ldots, p\}$*. Then the family of r.v.s* $(Y_i)_{i \in \{1,\ldots,p\}}$ *are independent and for any* $i \in \{1, \ldots, p\}$*, the distribution of* $\|Y_i\|^2$ *is* $\chi^{\mathbf{2}}(n_i)$*.*

Corollary 3.7. *Let* $(X_i)_{i \in \{1,\ldots,n\}}$ *be i.i.d. real-valued rvs with distribution* $\mathrm{N}(\mathrm{m}, \sigma^2)$*. Consider* $\bar{X}_n = n^{-1} \sum_{k=1}^n X_k$ *and* $S_n^2 = (n-1)^{-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$*. Then,* $\bar{X}_n$ *and* $(n-1)S_n^2/\sigma^2$ *are independent with distribution* $\mathrm{N}(\mathrm{m}, \sigma^2/n)$ *and* $\chi^{\mathbf{2}}(n-1)$ *respectively. In particular, the distribution of* $\sqrt{n}(\bar{X}_n - \mathrm{m})/S_n$ *is* $\mathbf{T}(n-1)$*.*

Theorem 3.8. *Let X be a d-dimensional random variable with distribution $\mathrm{N}(\mathrm{m}, \Sigma)$ and Y be a p-dimensional random variable with distribution given X, $\mathrm{N}(\mathbf{A}X + b, \mathbf{C})$, for $\mathbf{A} \in \mathbb{R}^{p\times d}$ and $b \in \mathbb{R}^p$. Then, the distribution of Y is $\mathrm{N}(\mathbf{A}\mathrm{m} + b, \mathbf{C} + \mathbf{A}\Sigma\mathbf{A}^{\mathrm{T}})$ and the distribution of X given Y is $\mathrm{N}(\mathrm{m}_{X|Y}, \Sigma_{X|Y})$ with*

$$\mathrm{m}_{X|Y} = \Sigma_{X|Y}\{\mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}(Y - b) + \Sigma^{-1}\mathrm{m}\} \tag{8}$$

$$\Sigma_{X|Y} = (\Sigma^{-1} + \mathbf{A}^{\mathrm{T}}\mathbf{C}^{-1}bfA)^{-1} \, . \tag{9}$$

3.1.2 Exercises

Exercise 3.1. Let $(G_i)_{i\in\{1,\dots,n\}}$ be i.i.d. one-dimensional Gaussian random variables with distribution $\mathrm{N}(0, 1)$.

1. Show that the n-dimensional vector $X = (G_1, \dots, G_n)$ is Gaussian and specify its mean and covariance matrix.
2. Deduce how to get a random variable with distribution $\mathrm{N}(\mathrm{m}, \Sigma)$ from $(G_i)_{i\in\{1,\dots,n\}}$.

Exercice 3.2. Let $X = (X_1, X_2, X_3)$ be a Gaussian random vector with mean 0 and covariance matrix

$$\Gamma = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

1. What can be said about X_3 and (X_1, X_2)?
2. What is the distribution of (X_1, X_2)?
3. Show that for any $a \in \mathbb{R}$, the vector $(X_2, X_2 + aX_1)$ is a Gaussian random vector.
4. Choosing appropriately a such that X_2 and $X_2 + aX_1$ is independent, give an expression for $\mathbb{E}[X_1|X_2]$.

3.2 Bayesian inference

Exercise 3.3. Here we aim to prove the result regarding the posterior distribution of a linear regression models when we choose a Gaussian prior.

(i) Consider a density p on $\mathbb{R}^d$ satisfying:

$$p(w) \propto \exp\left(-\frac{1}{2}\langle Aw, w\rangle + \langle b, w\rangle\right) \, , \tag{10}$$

for $A \in \mathbb{R}^{d\times d}$ symmetric definite positive and $b \in \mathbb{R}^d$. Show that p is the density of $\mathrm{N}(m_p, \Sigma_p)$ with

$$\Sigma_p = A^{-1} \, , \quad m_p = \Sigma_p b \, . \tag{11}$$

For $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we now consider a linear regression model:

$$p_w(\mathcal{D}) \propto \prod_{i=1}^n \exp\left(-\frac{\beta}{2}(y_i - \sum_{j=1}^d w_j\phi_j(x_i))^2\right) \, , \tag{12}$$

for some fixed precision parameter $\beta > 0$ and w is the parameter of interest.

(ii) Write p_w of the form

$$p_w(\mathcal{D}) \propto \exp\left(-\frac{\beta}{2}\|y - \Phi_x w\|^2\right) \, , \tag{13}$$

where $y \in \mathbb{R}^n$ and $\Phi_x \in \mathbb{R}^{n\times d}$ have to be specified.

we choose as prior $p(w)$ for w, $\mathrm{N}(m_0, S_0)$

(iii) Find the posterior distribution $p(w|\mathcal{D})$ for the considered model.

We now choose $m_0 = 0$ and $S_0 = \alpha^{-1}\mathrm{I}_d$.

(iv) Give the MAP and the posterior mean.

(v) Identify the predictive distribution $p(\tilde{y}_{\text{new}}|\tilde{x}_{\text{new}}, \mathcal{D}) = \int p_w(\tilde{y}_{\text{new}}|\tilde{x}_{\text{new}})p(w|\mathcal{D})\mathrm{d}w$. You may find Theorem 3.8 useful.

(vi) Give an explicit expression for the marginal likelihood $p(y|x) = \int p_w(y|w)p(w)\mathrm{d}w$.

Exercise 3.4. Find the likelihoods function and the posterior distribution for the following models and priors:

(i) $Y_1, \ldots, Y_n|(\mu, \lambda) \overset{\text{iid}}{\sim} \mathrm{N}_1(\mu, \lambda^{-1})$, $\quad (\mu, \lambda) \sim p(\mu|\lambda)p(\lambda) = \mathrm{N}(\mu_0, (\beta\lambda)^{-1})\,\Gamma(a, b)$, for some fixed $\mu_0 \in \mathbb{R}$ and $\beta, a, b > 0$.

(ii) $Y_1, \ldots, Y_n|q \overset{\text{iid}}{\sim} \mathrm{Neg}(10, q)$ and $q \sim \mathrm{Beta}(1/2, 1/2)$, where $\mathrm{Neg}(r, q)$ is the negative binomial distribution on $\mathbb{N}$ with density

$$p(y|r, q) = \begin{pmatrix} y + r - 1 \\ y \end{pmatrix} (1 - q)^r q^y \ . \tag{14}$$

(iii) For each of these models give the marginal distribution defined for a likelihood $p_w(y_1, \ldots, y_n)$ and a prior $p(w)$ by $p(y_1, \ldots, y_n) = \int p_w(y_1, \ldots, y_n)p(w)\mathrm{d}w$, where w is the parameter of interest.

Exercise 3.5 (Numerical homework)**.** We consider here still the file:

`year-sunspots-republicans.csv`,

but we aim to perform Bayesian inference now from the Bayesian linear model:

$$Y = \sum_{j=1}^{d} w_j \phi_j(X) + \sigma\varepsilon \ , \quad \varepsilon \sim \mathrm{N}(0, 1) \ , \tag{15}$$

with $\sigma = 1$. We choose $\mathrm{N}(0, \alpha^{-1}\mathrm{I})$, with $\alpha = 0.01$ as a prior.

Recall that the numbers in the *Year* column are large (between 1960 and 2006), especially when raised to various powers. To avoid numerical instability due to ill-conditioned matrices in most numerical computing systems, we will scale the data first: specifically, we will scale all "year" inputs by subtracting 1960 and then dividing by 40.

1. Implement these procedures to obtain new features (you can use the function you implement for the last homework Exercise 2.2). In the sequel, we only use these new features.

2. - Plot the data and the Bayesian predictor for each of the following sets of basis functions, and include the generated plot as an image in your submission. You will therefore make 4 total plots:
     (a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$
         ie, use basis $y = a_1x^1 + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$ for some constants $\{a_1, ..., a_5\}$.
     (b) $\phi_j(x) = \exp -(40x - \mu_j)^2/25$ for $\mu_j = 0, 5, 10, \ldots, 50$
     (c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 5$
     (d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 25$

     * Note: Please make sure to add a bias term for all your basis functions above in your implementation

- For each set of basis functions, compute the posterior standard deviation $x' \mapsto \sigma_N(x')$ and illustrate the predictive uncertainty by plotting a shaded region, that spans $\sigma_N(x')$ either side of $\hat{y}^{\mathcal{D}}(x')$. Again, you therefore make 4 total plots.

3. Repeat the same exact process as above but for **Number of Sunspots (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Here, to avoid numerical instability with numbers in the *Sunspot_Count* column, we will also scale the data first by dividing all "sunspot count" inputs by 20. In addition, only use data from before 1985, and only use basis functions (a), (c), and (d) – ignore basis (b). You will therefore make 3 total plots. For each plot make sure to also include the train error.

4. Compute the marginal likelihood (or normalizing constant) for the the three bases (a, c, d).

5. Deduce from your last answer, which of the three bases (a, c, d) provided the "best" fit from a Bayesian perspective?

6. Discuss on the choice of α and σ.