Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik

Probabilistic and Bayesian Modelling in Machine Learning and Artificial Intelligence

Manfred Opper, Théo Galy-Fajou

Summer Term 2018

Problem Sheet 3

Solutions

Problem 1 – Bayes inference for the variance of a Gaussian

Use a Bayesian approach to estimate the inverse variance λ of a univariate Gaussian distribution

$$p(x|\lambda) = \sqrt{\frac{\lambda}{2\pi}}\ \exp\left[-\frac{\lambda x^2}{2}\right].$$

Here we have assumed for simplicity that the data has zero mean $\mu = 0$. To apply Bayesian inference we specify a *Gamma* prior distribution for λ,

$$p(\lambda) = \frac{\lambda^{\alpha-1}\exp\left[-\lambda/\beta\right]}{\Gamma(\alpha)\beta^{\alpha}}$$

where the positive numbers α and β, the *hyperparameters* of the model are assumed to be known and $\Gamma(\alpha)$ is Euler's *gamma* function (`gamma` in Octave and R). We then observe a dataset $D = (x_1, x_2, \ldots, x_N)$ comprising N independent random samples from $p(x|\lambda)$.

(a) Show that the posterior probability $p(\lambda|D)$ of the inverse variance is also a *gamma* distribution with parameters

$$\alpha_p = \alpha + \frac{N}{2}, \qquad \frac{1}{\beta_p} = \frac{1}{\beta} + \frac{1}{2}\sum_{i=1}^{N} x_i^2.$$

- Likelihood of the data set

$$p(D|\lambda) = \prod_{i=1}^{N}\sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}x_i^2\right)$$

- Joint distribution for D and λ

$$\begin{aligned}
p(D,\lambda) &= p(D|\lambda)p(\lambda) \\
&= \frac{\lambda^{\alpha-1}\exp(-\lambda\beta^{-1})}{\Gamma(\alpha)\beta^{\alpha}} \prod_{i=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}x_i^2\right) \\
&= \frac{\lambda^{(\alpha+N/2)-1}}{(2\pi)^{N/2}\Gamma(\alpha)\beta^{\alpha}} \exp\left[-\lambda\left(\frac{1}{\beta}+\frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\right] \\
&= \frac{\lambda^{\alpha_p-1}\exp(-\lambda\beta_p^{-1})}{(2\pi)^{N/2}\Gamma(\alpha)\beta^{\alpha}} \\
&= \frac{\Gamma(\alpha_p)\beta_p^{\alpha_p}}{(2\pi)^{N/2}\Gamma(\alpha)\beta^{\alpha}} \frac{\lambda^{\alpha_p-1}\exp(-\lambda\beta_p^{-1})}{\Gamma(\alpha_p)\beta_p^{\alpha_p}}
\end{aligned}$$

- Posterior for λ

$$p(\lambda|D) = \frac{\lambda^{\alpha_p-1}\exp(-\lambda\beta_p^{-1})}{\Gamma(\alpha_p)\beta_p^{\alpha_p}}$$

(b) Compute the mean of the posterior distribution of λ. Compare the result with the result from the *maximum-likelihood* estimation, $\lambda_{\mathrm{ML}} = 1/\sigma_{\mathrm{ML}}^2$ and explain what happens if $N \to \infty$.

- Mean of the posterior distribution:

$$\begin{aligned}
\langle\lambda_p\rangle &= \int_0^{\infty} \lambda\, p(\lambda|D)\, d\lambda \\
&= \int_0^{\infty} \lambda \frac{\lambda^{\alpha_p-1}\exp(-\lambda\beta_p^{-1})}{\Gamma(\alpha_p)\beta_p^{\alpha_p}}\, d\lambda \\
&= \frac{1}{\Gamma(\alpha_p)\beta_p^{\alpha_p}} \int_0^{\infty} \lambda^{\alpha_p}\, e^{-\lambda/\beta_p}\, d\lambda \\
&= \frac{\beta_p}{\Gamma(\alpha_p)} \int_0^{\infty} z^{\alpha_p}\, e^{-z}\, dz \\
&= \frac{\Gamma(\alpha_p+1)}{\Gamma(\alpha_p)}\beta_p \\
&= \alpha_p\beta_p \\
&= \left(\alpha+\frac{N}{2}\right)\left(\frac{1}{\beta}+\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}
\end{aligned}$$

- Negative logarithm of the likelihood

$$\mathcal{L} = -\log p(D|\lambda) = \frac{\lambda}{2}\sum_{i=1}^{N} x_i^2 - \frac{N}{2}\log\lambda + \frac{N}{2}\log(2\pi)$$

- Maximum likelihood estimate

$$\frac{d\mathcal{L}}{d\lambda} = 0 \quad \Longleftrightarrow \quad \frac{1}{2}\sum_{i=1}^{N} x_i^2 - \frac{N}{2\lambda} = 0 \quad \Longleftrightarrow \quad \lambda_{\mathrm{ML}} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-1}$$

- For $N \to \infty$ the posterior mean $\langle \lambda_p \rangle$ approaches the maximum likelihood estimate λ_{ML} asymptotically:

$$\langle \lambda_p \rangle = \frac{\alpha + N/2}{\beta^{-1} + N/2\,\lambda_{\mathrm{ML}}^{-1}} \quad \Longrightarrow \quad \lim_{N\to\infty} \langle \lambda_p \rangle = \lambda_{\mathrm{ML}}$$

(c) Show that the variance of the posterior distribution $\mathrm{Var}(\lambda_{\mathrm{post}}) = \langle \lambda^2 \rangle - \langle \lambda \rangle^2$ shrinks to zero as $N \to \infty$. Here we have used the notation $\langle \ldots \rangle$ for posterior expectations.

- Variance of the posterior distribution

$$\begin{aligned}
\langle \lambda_p^2 \rangle - \langle \lambda_p \rangle^2 &= \int_0^\infty \lambda^2\, p(\lambda|D)\, d\lambda - \alpha_p^2\, \beta_p^2 \\
&= \int_0^\infty \lambda^2\, \frac{\lambda^{\alpha_p - 1} \exp(-\lambda \beta_p^{-1})}{\Gamma(\alpha_p)\beta_p^{\alpha_p}}\, d\lambda - \alpha_p^2\, \beta_p^2 \\
&= \frac{1}{\Gamma(\alpha_p)\beta_p^{\alpha_p}} \int_0^\infty \lambda^{\alpha_p+1}\, e^{-\lambda/\beta_p}\, d\lambda - \alpha_p^2\, \beta_p^2 \\
&= \frac{\beta_p^2}{\Gamma(\alpha_p)} \int_0^\infty z^{\alpha_p+1}\, e^{-z}\, dz - \alpha_p^2\, \beta_p^2 \\
&= \frac{\Gamma(\alpha_p + 2)}{\Gamma(\alpha_p)}\, \beta_p^2 - \alpha_p^2\, \beta_p^2 \\
&= \alpha_p\, (\alpha_p + 1)\, \beta_p^2 - \alpha_p^2\, \beta_p^2 \\
&= \alpha_p\, \beta_p^2 \\
&= \left(\alpha + \frac{N}{2}\right)\left(\frac{1}{\beta} + \frac{1}{N}\sum_{i=1}^{N} x_i^2\right)^{-2}
\end{aligned}$$

- Asymptotic behaviour for $N \to \infty$

$$\lim_{N\to\infty} \mathrm{Var}(\lambda_p) \;=\; \lim_{N\to\infty} \frac{1}{N}\frac{\alpha/N + 1/2}{(\beta^{-1}/N + \lambda_{\mathrm{ML}}^{-1}/2)^2} \;=\; 2\lambda_{\mathrm{ML}}^2 \lim_{N\to\infty} \frac{1}{N} \;=\; 0$$

(d) Show that the predictive distribution is

$$p(x|D) = \frac{1}{\sqrt{2\pi}}\frac{\Gamma(\alpha_p + 1/2)}{\Gamma(\alpha_p)}\sqrt{\beta_p}\left(1 + \frac{x^2 \beta_p}{2}\right)^{-\alpha_p - 1/2}$$

where α_p and β_p were defined above. Note, this is **not a Gaussian**!

$$\begin{aligned}
p(x|D) &= \int_0^\infty p(x|\lambda)p(\lambda|D)d\lambda \\
&= \int_0^\infty \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}x^2\right) \frac{\lambda^{\alpha_p-1}\exp(-\lambda\beta_p^{-1})}{\Gamma(\alpha_p)\beta_p^{\alpha_p}} d\lambda \\
&= \frac{1}{\sqrt{2\pi}\,\Gamma(\alpha_p)\beta_p^{\alpha_p}} \int_0^\infty \lambda^{\alpha_p-1/2} \exp\left[-\lambda\left(\frac{1}{\beta_p}+\frac{1}{2}x^2\right)\right] d\lambda \\
&= \frac{1}{\sqrt{2\pi}} \frac{\Gamma(\alpha_p+1/2)}{\Gamma(\alpha_p)} \frac{1}{\beta_p^{\alpha_p}} \left(\frac{1}{\beta_p}+\frac{1}{2}x^2\right)^{-\alpha_p-1/2} \\
&= \frac{1}{\sqrt{2\pi}} \frac{\Gamma(\alpha_p+1/2)}{\Gamma(\alpha_p)} \sqrt{\beta_p} \left(1+\frac{x^2\,\beta_p}{2}\right)^{-\alpha_p-1/2}
\end{aligned}$$

For all conjugate priors used in Bayesian analysis of the Gaussian distribution (including Normal, Gamma-Normal, Wishart etc...), see this review from Kevin Murphy : Conjugate Bayesian analysis of the Gaussian distribution

Problem 2 – Hyperparameter estimation for a generalised linear model

Consider a model for a set of data $D = (y_1, \ldots, y_n)$ defined by

$$p(D|\mathbf{w}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left[-\sum_{i=1}^{N} \frac{\beta}{2}\left(y_i - \sum_{j=1}^{K} w_j \Phi_j(x_i)\right)^2\right]$$

with a fixed set $\{\Phi_1(x), \ldots, \Phi_k(x)\}$ of K basis functions. The prior distribution on the weights is given by

$$p(\mathbf{w}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{K/2} \exp\left[-\frac{\alpha}{2}\sum_{j=1}^{K} w_j^2\right].$$

This *generalised linear model* assumes that the observations are generated from a weighted linear combination of the basis functions with additive Gaussian noise.

(a) The posterior distribution $p(\mathbf{w}|D)$ of the vector of weights is a Gaussian. Compute the posterior mean vector $\mathrm{E}[\mathbf{w}]$ and the posterior covariance in terms of the matrix $\mathbf{X}$ where $X_{lk} = \Phi_k(x_l)$.

- Joint distribution in matrix notation

$$\begin{aligned}
&p(D, \mathbf{w}|\alpha, \beta) \\
= \;& \left(\frac{\alpha}{2\pi}\right)^{\frac{K}{2}} \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} - \frac{\beta}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\right)
\end{aligned}$$

- Mean value of the posterior (see Fisher information)

$$
\begin{aligned}
& \left.\frac{\partial \log p(D, \mathbf{w}|\alpha, \beta)}{\partial \mathbf{w}}\right|_{\mathbf{w}=\langle \mathbf{w} \rangle} = 0 \\
\iff \quad & -\alpha \langle \mathbf{w} \rangle + \beta \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\langle \mathbf{w} \rangle) = 0 \\
\iff \quad & (\alpha \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X}) \langle \mathbf{w} \rangle = \beta \mathbf{X}^\top \mathbf{y} \\
\iff \quad & \langle \mathbf{w} \rangle = \left(\frac{\alpha}{\beta} \mathbf{I} + \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}
\end{aligned}
$$

- Covariance of the posterior (see Fisher information)

$$
\begin{aligned}
& \left.\frac{\partial^2 \log p(D, \mathbf{w}|\alpha, \beta)}{\partial \mathbf{w}^2}\right|_{\mathbf{w}=\langle \mathbf{w} \rangle} = -\mathrm{Cov}(\mathbf{w})^{-1} \\
\iff \quad & \mathrm{Cov}(\mathbf{w})^{-1} = \alpha \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X} \\
\iff \quad & \mathrm{Cov}(\mathbf{w}) = \frac{1}{\beta} \left(\frac{\alpha}{\beta} \mathbf{I} + \mathbf{X}^\top \mathbf{X} \right)^{-1}
\end{aligned}
$$

(b) Derive an EM algorithm for optimising the hyperparameter β by maximising the log-evidence

$$
p(D|\alpha, \beta) = \int p(D|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) \, d\mathbf{w}
$$

Hint: Treat the weights $\mathbf{w}$ as a set of latent variables similar to the procedure for α given in the lecture. Express your result in terms of the posterior mean and variance.

- Expectation step

$$
\begin{aligned}
\mathcal{L} &= \langle \log p(D, \mathbf{w}|\alpha, \beta) \rangle \\
&= \left\langle \frac{K}{2} \log \frac{\alpha}{2\pi} + \frac{N}{2} \log \frac{\beta}{2\pi} - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} - \frac{\beta}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \right\rangle \\
&= \frac{K}{2} \log \frac{\alpha}{2\pi} + \frac{N}{2} \log \frac{\beta}{2\pi} - \frac{\alpha}{2} \langle \mathbf{w}^\top \mathbf{w} \rangle - \frac{\beta}{2} \langle (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \rangle
\end{aligned}
$$

- Expected length of the weight vector

$$
\begin{aligned}
\langle \mathbf{w}^\top \mathbf{w} \rangle &= \mathrm{Tr}[\langle \mathbf{w}\mathbf{w}^\top \rangle] \\
&= \mathrm{Tr}[\mathrm{Cov}(\mathbf{w}) + \langle \mathbf{w} \rangle \langle \mathbf{w}^\top \rangle] \\
&= \mathrm{Tr}[\mathrm{Cov}(\mathbf{w})] + \langle \mathbf{w}^\top \rangle \langle \mathbf{w} \rangle
\end{aligned}
$$

- Expected distance between model and observations

$$
\begin{aligned}
& \langle (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \rangle \\
= \; & \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X} \langle \mathbf{w} \rangle - \langle \mathbf{w}^\top \rangle \mathbf{X}^\top \mathbf{y} + \mathrm{Tr}[\mathbf{X} \langle \mathbf{w}\mathbf{w}^\top \rangle \mathbf{X}^\top] \\
= \; & \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X} \langle \mathbf{w} \rangle - \langle \mathbf{w}^\top \rangle \mathbf{X}^\top \mathbf{y} + \mathrm{Tr}[\mathbf{X} \mathrm{Cov}(\mathbf{w}) \mathbf{X}^\top] \\
+ \; & \mathrm{Tr}[\mathbf{X} \langle \mathbf{w} \rangle \langle \mathbf{w}^\top \rangle \mathbf{X}^\top] \\
= \; & \mathrm{Tr}[\mathbf{X} \mathrm{Cov}(\mathbf{w}) \mathbf{X}^\top] + (\mathbf{y} - \mathbf{X}\langle \mathbf{w} \rangle)^\top (\mathbf{y} - \mathbf{X}\langle \mathbf{w} \rangle)
\end{aligned}
$$

- Maximization step

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha} = 0 \quad & \Longleftrightarrow \quad \frac{K}{2\alpha} - \frac{1}{2} \langle \mathbf{w}^\top \mathbf{w} \rangle = 0 \\
& \Longleftrightarrow \quad \alpha = \frac{K}{\langle \mathbf{w}^\top \mathbf{w} \rangle} \\
\frac{\partial \mathcal{L}}{\partial \beta} = 0 \quad & \Longleftrightarrow \quad \frac{N}{2\beta} - \frac{1}{2} \langle (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \rangle = 0 \\
& \Longleftrightarrow \quad \beta = \frac{N}{\langle (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \rangle}
\end{aligned}
$$