02457 Non-Linear Signal Processing,
Note for week 12: Dynamic linear models

This note is based on Christopher Bishop: *Machine Learning and Pattern Recognition*, sections 3.3, 13.1, 13.3.

Bayesian linear models

To warm up we first (re-)turn to Bayesian inference in the linear model.

Let $\mathcal{D} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), ..., (\mathbf{x}_N, t_N)\}$ be a data set of N samples with $\mathbf{x} \in \mathbb{R}^d$. We will focus on function approximation and assume the generative model is linear

$$t = \mathbf{w}^\top \mathbf{x} + \epsilon \tag{1}$$

with $\epsilon \sim \mathcal{N}(0, \beta^{-1})$, i.e., normally distributed white noise with precision β. With these assumptions the likelihood function becomes

$$p(\mathcal{D}|\mathbf{w}, \beta) = \left(\sqrt{\frac{\beta}{2\pi}}\right)^N \exp\left(-\frac{\beta}{2}\sum_{n=1}^N (t_n - \mathbf{w}^\top \mathbf{x}_n)^2\right) \tag{2}$$

As in exercise 4 we will assign a standard Gaussian prior to the weights $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$, where $\mathbf{I}$ is a d-dimensional unit matrix, leading to the posterior distribution

$$\begin{aligned} p(\mathbf{w}|\alpha, \beta, \mathcal{D}) &= \frac{p(\mathcal{D}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha)}{p(\mathcal{D}|\alpha, \beta)} \\ &\propto \left(\sqrt{\frac{\beta}{2\pi}}\right)^N \exp\left(-\frac{\beta}{2}\sum_{n=1}^N (t_n - \mathbf{w}^\top \mathbf{x}_n)^2\right) \left(\sqrt{\frac{\alpha}{2\pi}}\right)^d \exp\left(-\frac{\alpha}{2}||\mathbf{w}||^2\right). \end{aligned}$$

The posterior is a product of two normal probability density functions. Combining the exponents we obtain a quadratic form in $\mathbf{w}$, hence again a normal distribution. For such a product there is a combination rule that we will need again below: The product between $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, is proportional to $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ with mean vector and covariance matrix given by,

$$\begin{aligned} \boldsymbol{\mu}_p &= \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \\ \boldsymbol{\Sigma}_p &= \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)^{-1}. \end{aligned} \tag{3}$$

In this case the prior is given by $\boldsymbol{\mu}_2 \equiv \boldsymbol{\mu}_{\text{prior}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_2 \equiv \boldsymbol{\Sigma}_{\text{prior}} = \alpha^{-1}\mathbf{I}$. For the likelihood a bit of algebra leads to,

$$\begin{aligned} \boldsymbol{\mu}_1 &\equiv \left(\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top\right)^{-1} \sum_{n=1}^N \mathbf{x}_n t_n \\ \boldsymbol{\Sigma}_1 &\equiv \left(\beta \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top\right)^{-1}. \end{aligned} \tag{4}$$

Hence the posterior mean vector and covariance matrix are found as

$$\begin{aligned}\boldsymbol{\mu}_p &\equiv \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top\right)^{-1}\beta\sum_{n=1}^{N}\mathbf{x}_n t_n \\ \boldsymbol{\Sigma}_p &\equiv \left(\alpha\mathbf{I} + \beta\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top\right)^{-1}.\end{aligned} \tag{5}$$

The predictive density is computed as

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathcal{D}) = \int p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w} \tag{6}$$

This is again a normal distribution. We note

$$t_{N+1} = \mathbf{w}_N^\top\mathbf{x}_{N+1} + \epsilon_{N+1}, \quad \text{and} \quad \mathbf{w}_N \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p),$$

which leads to the predictive mean and variance,

$$\begin{aligned}\mu_{t_{N+1}} &= \boldsymbol{\mu}_p^\top\mathbf{x}_{N+1}, \\ \sigma^2_{t_{N+1}} &= \beta^{-1} + \mathbf{x}_{N+1}^\top\boldsymbol{\Sigma}_p\mathbf{x}_{N+1}.\end{aligned} \tag{7}$$

Dynamic Bayesian models

The basic assumption in the previous derivation is that the parameter vector is stationary. In a dynamic setting we relax this assumption and assume that $\mathbf{w}_n$ is changing as data arrives. A possible prior could be the simple Markovian random walk

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \boldsymbol{\nu}_n \tag{8}$$

with $\boldsymbol{\nu}_n \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$. The pdf reads

$$p(\mathbf{w}_n|\mathbf{w}_{n-1}, \alpha) = \left(\sqrt{\frac{\alpha}{2\pi}}\right)^d \exp\left(-\frac{\alpha}{2}||\mathbf{w}_n - \mathbf{w}_{n-1}||^2\right). \tag{9}$$

A high value of precision parameter α means small changes in $\mathbf{w}_n$ as time progresses.

To simplify the notation, let us define $\mathbf{z}_n = (t_n, \mathbf{x}_n)$ and let us denote the set of all data observed until n by $\mathbf{z}_{1:n}$. We are interested in the 'dynamic posterior' $p(\mathbf{w}_n|\mathbf{z}_{1:n})$ and this quantity, it turns out, can be computed in a recursive manner.

For the proportional quantity, the joint density $p(\mathbf{w}_n, \mathbf{z}_{1:n})$, the forward recursion can be derived with these manipulations

$$p(\mathbf{w}_n, \mathbf{z}_{1:n}) = \int p(\mathbf{w}_n, \mathbf{w}_{n-1}, \mathbf{z}_{1:n})d\mathbf{w}_{n-1} \tag{10}$$

$$= \int p(\mathbf{w}_n, \mathbf{w}_{n-1}, \mathbf{z}_n, \mathbf{z}_{1:(n-1)})d\mathbf{w}_{n-1} \tag{11}$$

$$= \int p(\mathbf{z}_n|\mathbf{w}_n, \mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})p(\mathbf{w}_n, \mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})d\mathbf{w}_{n-1} \tag{12}$$

$$= p(\mathbf{z}_n|\mathbf{w}_n)\int p(\mathbf{w}_n|\mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})p(\mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})d\mathbf{w}_{n-1} \tag{13}$$

$$= p(\mathbf{z}_n|\mathbf{w}_n)\int p(\mathbf{w}_n|\mathbf{w}_{n-1})p(\mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})d\mathbf{w}_{n-1}. \tag{14}$$

Here $p(\mathbf{z}_n|\mathbf{w}_n) = p(\mathbf{z}_n|\mathbf{w}_n, \beta)$ is the observation likelihood, while $p(\mathbf{w}_n|\mathbf{w}_{n-1}) = p(\mathbf{w}_n|\mathbf{w}_{n-1}, \alpha)$ is Markov prior. As $p(\mathbf{w}_{n-1}, \mathbf{z}_{1:(n-1)})$ is the sought joint distribution evaluated at the previous time step $n-1$, we see that by performing a single d-dimensional integral (wrt. $\mathbf{w}_{n-1}$) and subsequent multiplication by $p(\mathbf{z}_n|\mathbf{w}_n)$ we arrive at the 'updated' joint distribution. The posterior distribution of $\mathbf{w}_n$, in turn, can be obtained by normalization,

$$p(\mathbf{w}_n|\mathbf{z}_{1:n}) = \frac{p(\mathbf{w}_n, \mathbf{z}_{1:n})}{\int p(\mathbf{w}_n, \mathbf{z}_{1:n}) d\mathbf{w}_n}. \tag{15}$$

Here we recognize the normalization constant is the model likelihood $p(\mathbf{z}_{1:n}|\text{Model}) = \int p(\mathbf{w}_n, \mathbf{z}_{1:n}) d\mathbf{w}_n$, hence it is a by product of the 'forward recursion'.

For the linear model we analyzed above, we get specifically,

$$\begin{aligned}
p(\mathbf{z}_n|\mathbf{w}_n, \beta) &= p(t_n|\mathbf{w}_n, \mathbf{x}_n, \beta)p(\mathbf{x}_n) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2\right) p(\mathbf{x}_n) \\
p(\mathbf{w}_n|\mathbf{w}_{n-1}, \alpha) &= \left(\sqrt{\frac{\alpha}{2\pi}}\right)^d \exp\left(-\frac{\alpha}{2}||\mathbf{w}_n - \mathbf{w}_{n-1}||^2\right).
\end{aligned}$$

The recursion starts as $p(\mathbf{w}_1, \mathbf{z}_1) \propto p(\mathbf{z}_1|\mathbf{w}_1)p(t_1|\mathbf{x}_1, \mathbf{w}_1)p(\mathbf{x}_1)$, hence the quantity of interest starts out being proportional to a normal density function in terms of $\mathbf{w}_1$. We also see that in order to compute the update for $p(\mathbf{w}_2, \mathbf{z}_{1:2})$ we perform an integral over the product of two normal distributions

$$p(\mathbf{w}_2, \mathbf{z}_{1:2}) = p(\mathbf{z}_2|\mathbf{w}_2, \beta) \int p(\mathbf{w}_2|\mathbf{w}_1, \alpha)p(\mathbf{w}_1, \mathbf{z}_1) d\mathbf{w}_1. \tag{16}$$

The result of this integral is a again a normal distribution and as this is followed by multiplication by the local likelihood, also an un-normalized normal density, we obtain a $p(\mathbf{w}_2, \mathbf{z}_{1:2})$ which itself is proportional to a normal distribution. It then follows by induction that all the following terms $p(\mathbf{w}_n, \mathbf{z}_{1:n})$ are (un-normalized) normal density functions, in case of the linear model.

Tracking the means and covariances of these un-normalized density functions, which involves the product rule (Eq. 3) two times, we get a message passing scheme for mean and covariance of the un-normalized posterior $p(\mathbf{w}_n, \mathbf{z}_{1:n})$

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w},n} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{w},n-1} + \alpha^{-1}\mathbf{I}\right)^{-1} + \beta \mathbf{x}_n \mathbf{x}_n^\top\right)^{-1} \left(\left(\boldsymbol{\Sigma}_{\mathbf{w},n-1} + \alpha^{-1}\mathbf{I}\right)^{-1} \boldsymbol{\mu}_{\mathbf{w},n-1} + \beta t_n \mathbf{x}_n\right) \\
\boldsymbol{\Sigma}_{\mathbf{w},n} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{w},n-1} + \alpha^{-1}\mathbf{I}\right)^{-1} + \beta \mathbf{x}_n \mathbf{x}_n^\top\right)^{-1} \qquad (17)
\end{aligned}$$

At any given time we can use the predictive means in Eq. 7 to find the estimator and the associated uncertainty.

Lars Kai Hansen, November 2016.