Stochastic Machine Learning
Chapter 04 - Filtering

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

Filtering

- Filtering is a dynamic Bayesian way of estimation
- The key tool is Bayes' formula - which we recall for sets with positive probability

$$P(A|B) = \frac{P(A \cap B)}{P(B))}.$$

- Think of say the outcome of the first dice ω_1, when we throw two dices and observe the sum S :

$$P(\omega_1 = 1|S = 2) = 1 \tag{1}$$

$$P(\omega_1 = 1|S = 3) = \frac{1/12}{2/12} = \frac{1}{2} \tag{2}$$

$$\vdots \tag{3}$$

- With now information we have the **a priori** probability

$$P(\omega_1 = i) = \frac{1}{6}, \qquad i = 1, \ldots, 6.$$

- But depending on our observation we update this distribution to the **a posteriori** distribution given above. In a setting with discrete random variables we can compute everything !

Continuous random variables

- If we have continuous distributions, we can work with densities. In this case, Bayes' formula reads
$$f(x|y) = \frac{f(x,y)}{f(y)}.$$
- In many cases, this is not so easy to evaluate, but in the Gaussian case we obtain that the a posteriori distribution is again **Gaussian**.
- The key to this is the following simple observation: Consider X, Y standard normal and independent. Then,
$$E[X + \rho Y|Y] = \rho Y.$$
- Now, two Gaussian random variables ξ_1 and ξ_2 with correlation ρ may always be represented as
$$(\xi_1, \rho\xi_1 + \sqrt{1-\rho^2}\eta)$$
with standard normal η, independent of ξ_1. **Why?**
- So, as above
$$E[\xi_2|\ \xi_1] = \rho\xi_1.$$

- Computing the a posteriori distribution is a little bit more work.
- We could start from the Bayes' formula with densities - but we can also exploit our knowledge on conditional expectations.

$$\begin{aligned} P(\xi_2 \leq x|\xi_1) &= P\big(\rho\xi_1 + \sqrt{1-\rho^2}\eta \leq x|\xi_1\big) \\ &= P\Big(\eta \leq \frac{x-\rho\xi_1}{\sqrt{1-\rho^2}}|\xi_1\Big) \\ &= \Phi\Big(\frac{x-\rho\xi_1}{\sqrt{1-\rho^2}}\Big). \end{aligned}$$

- Hence - as expected -

$$\xi_2|\xi_1 \sim \mathcal{N}\big(\rho\xi_1, 1-\rho^2\big)$$

Lemma

Assume $\xi \sim \mathcal{N}_2(\mu, \Sigma)$ *with*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{4}$$

Then

$$\xi_1 |\ \xi_2 \sim \mathcal{N}\Big(\mu_1 + \rho \cdot \sigma_1 \frac{\xi_2 - \mu_2}{\sigma_2}, \sigma_1^2 \cdot (1 - \rho^2)\Big).$$

Proof.

Of course we use our above representation, now with appropriate rescaling.

$$(\xi_1, \xi_2) = \Big(\mu_1 + \sigma_1 X_1, \mu_2 + \sigma_2 X_2\Big) \tag{5}$$

where $(X - 1, X_2)$ is standard normal with correlation ρ. Hence,

$$\mu_1 + \sigma_1 X_1 | X_2 \sim \mathcal{N}\Big(\mu_1 + \rho \cdot \sigma_1 X_2, \sigma_1^2 \cdot (1 - \rho^2)\Big)$$

Now we replace X_2 by $\xi_2 - \mu_2/\sigma_2$ and the proof is finished. □

The Kalman filter

- We obtain the one-dimensional Kalman Filter as follows:
- The unobserved signal $X = (X_n)_{n \geq 1}$ is a Gaussian process given by
$$X_n = aX_{n-1} + b\epsilon_n,$$
where (ϵ_n) are i.i.d. $\mathcal{N}(0, 1)$.
- The observation is given by
$$Y_n = AX_n + B\eta_n,$$
where also (η_n) are also i.i.d., independent of (ϵ_n).
- If we just want to estimate the state of X_n, we denote
$$\hat{X}_n = E[X_n | Y_1, \ldots, Y_n]$$
- The look to a past X_j, $j < n$ is called smoothing and the look into a future $j > n$ prediction.

Theorem (Kalman filter)

The conditional distribution of the signal given the observation Y until time n is normal with mean

$$\mu_n = \hat{X}_{n|n-1} + K_n \cdot \left(Y_n - A\hat{X}_{n|n-1}\right). \tag{6}$$

and variance

$$\sigma_n^2 = \sigma_{n|n-1}^2 \cdot \left(1 - AK_n\right), \tag{7}$$

where

$$\hat{X}_{n|n-1} = E[X_n|X_{n-1}] = aX_{n-1} \tag{8}$$

$$\sigma_{n|n-1}^2 = a^2\sigma_{n-1}^2 + b^2 \tag{9}$$

$$K_n = \frac{A\sigma_{n|n-1}^2}{A^2\sigma_{n|n-1}^2 + B^2}. \tag{10}$$

Proof

- The proof proceeds in a number of steps. First we note that the **transition** of X from $n-1$ to n is given by the following transition probabilities:

$$X_n|\ X_{n-1} \sim \mathcal{N}(aX_{n-1}, b^2) =: \mathcal{N}(\hat{X}_{n|n-1}, \sigma^2_{n|n-1})$$

- The next step is to update the conditional distribution by the new incoming information (called the **updating step**): so we are looking for the distribution of $X_n|X_{n-1}, Y_n$.
- for this we compute the correlation:

$$\begin{aligned}\mathrm{Cov}(X_n, Y_n|X_{n-1}) &= \mathrm{Cov}(X_n, AX_n + B\eta_n|\ X_{n-1}) \\ &= A\,\mathrm{Var}(X_n|X_{n-1}) = A\sigma^2_{n|n-1}, \end{aligned} \tag{11}$$

so the correlation computes to

$$\rho = A\frac{\sqrt{\mathrm{Var}(X_n|X_{n-1})}}{\sqrt{\mathrm{Var}(Y_n|X_{n-1})}} =: \frac{A\sigma_{n|n-1}}{\Sigma_{n|n-1}}.$$

Observe that $\Sigma^2_{n|n-1} = A^2\sigma^2_{n|n-1} + B^2$ such that

$$A^2\sigma^2_{n|n-1} < \Sigma^2_{n|n-1}$$

and hence $\rho \in [-1, 1]$.

- The next step is to apply Lemma 1. This gives

$$X_n | Y_n, X_{n-1} \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

with

$$\mu_n = \hat{X}_{n|n-1} + \frac{\rho\sigma_{n|n-1}}{\Sigma_{n|n-1}} \cdot \big(Y_n - A\hat{X}_{n|n-1}\big). \tag{12}$$

The quantity

$$K_n := \frac{\rho\sigma_{n|n-1}}{\Sigma_{n|n-1}} = \frac{A\sigma^2_{n|n-1}}{\Sigma^2_{n|n-1}} = \frac{A\sigma^2_{n|n-1}}{A^2\sigma^2_{n|n-1} + B^2}$$

is the so-called **Kalman gain**. It replaces the correlation in the more general setting here by rescaling appropriately.

- Moreover,

$$\begin{aligned}\sigma_n^2 &= \sigma^2_{n|n-1} \cdot (1 - \rho^2) = \sigma^2_{n|n-1} \cdot \Big(1 - \frac{A^2\sigma^2_{n|n-1}}{\Sigma^2_{n|n-1}}\Big) \\ &= \sigma^2_{n|n-1} \cdot \big(1 - AK_n\big). \end{aligned} \tag{13}$$

- The formulas generalize suitably to multi-dimensions. Of course, we only need joint normality, so essentially this can be generalized to quite general settings.
- There is also a Kalman filter in continuous time, see for example Liptser and Shiryaev (2001), Chapter 8. The arguments turn out to be a little bit more subtle, but essentially it works in the same direction as shown here.
- Filtering in discrete time can be solved completely, as we will soon see. Filtering in continuous time requires semimartingale techniques and we refer to Grigelionis and Mikulevicius (2011) for a general treatment. (Also lecture notes from David Criens, Philipp Harms and Josef Teichmann and myself are available on request).

 Grigelionis, B. and R. Mikulevicius (2011). „Nonlinear filtering equations for stochastic processes with jumps“. In: **The Oxford handbook of nonlinear filtering**. Oxford: Oxford Univ. Press, pp. 95–128.

 Liptser, R. and A.N. Shiryaev (2001). **Statistics of Random Processes: II - Applications**. 2nd. Berlin: Springer Verlag.