16-264: Probabilistic reasoning

Combining information

Suppose we have two experts that each predict the temperature at noon tomorrow. How should we combine their estimates?

Suppose we believe one more than the other, how should we combine their estimates?

Modeling knowledge with probability distributions

One way to combine information is to model the information sources with probability distributions.

A well known probability distribution that is common and easy to work with is the Normal (Gaussian) distribution. This distribution looks like a symmetric hill. It is characterized by where the hill is (the mean) and how wide the hill is (the variance). The Normal distribution is what you get if you add a large number of independent identically distributed random variables (coin flips or dice rolls, for example).

The mean is the average of the numbers that are generated by a probability distribution. A way to write this is using expectations (E):

$$\mathbf{E}(\mathbf{x}) = \bar{\mathbf{x}} = \sum_{1}^{N} \mathbf{x}_i / N \tag{1}$$

In estimating the mean from data the formula $\sum_1^N \mathbf{x}_i/(N-1)$ to avoid bias. See a Statistics textbook.

The variance is the expectation of the deviation from the mean squared:

$$\mathbf{E}((\mathbf{x} - \bar{\mathbf{x}})^2) = \sigma^2 \tag{2}$$

If we are more sure, the variance is smaller. If we are less sure, the variance is larger.

Normal/Gaussian random variables

We will use some facts about Gaussian random variables. George Kantor's notes on Kalman Filtering `http://www.cs.cmu.edu/~cga/controls-intro/kantor/16_299_Kalman_Filter.pdf` have a nice review of Gaussian random variables.

Fact 1: A Gaussian random vector is fully characterized by its mean (first moment) and variance (2nd moment). A compact notation is $\mathbf{x} \sim \mathrm{N}(\mathrm{mean}, \mathrm{variance})$.

Fact 2: For any random vector $\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \Sigma)$, $\mathbf{Ax} \sim \mathrm{N}(\mathbf{Am}, \mathbf{A}\Sigma\mathbf{A}^{\mathrm{T}})$. For the scalar case, $a\mathbf{x} \sim \mathrm{N}(a\mathbf{m}, a^2\sigma^2)$.

Fact 3: If any two independent random vectors ($\mathbf{x}_1 \sim \mathrm{N}(\mathbf{m}_1, \Sigma_1)$ and $\mathbf{x}_2 \sim \mathrm{N}(\mathbf{m}_2, \Sigma_2)$), are added, the result is $\mathrm{N}(\mathbf{m}_1 + \mathbf{m}_2, \Sigma_1 + \Sigma_2)$

Fact 4: If you are given two predictions or belief states about a random variable $\mathbf{x}$, and the accuracy of these predications is $\hat{\mathbf{x}}_1 \sim \mathrm{N}(\mathbf{m}_1, \Sigma_1)$ (the belief of expert 1) and $\hat{\mathbf{x}}_2 \sim \mathrm{N}(\mathbf{m}_2, \Sigma_2)$ (the belief of expert 2), your best linear unbiased estimate (BLUE) of $\mathbf{x}$ is $\mathbf{W}_1\mathbf{m}_1 + \mathbf{W}_2\mathbf{m}_2$, with $\mathbf{W}_1 = \Sigma_2(\Sigma_1 + \Sigma_2)^{-1}$ and $\mathbf{W}_2 = \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}$. The variance of this estimate is:

$$\Sigma_2(\Sigma_1 + \Sigma_2)^{-1}\Sigma_1\Sigma_2(\Sigma_1 + \Sigma_2)^{-1} + \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}\Sigma_2(\Sigma_1 + \Sigma_2)^{-1}\Sigma_1 \quad (3)$$

What a mess! However, all the above matrices are symmetric, so we can reorder them and get

$$\Sigma_1\Sigma_2(\Sigma_1 + \Sigma_2)^{-1} \quad (4)$$

A useful way to express the same thing (since $\mathbf{W}_1 = (1 - \mathbf{W}_2)$) that we will use in the derivation of the Kalman Filter is:

$$\mathbf{m} = \mathbf{m}_1 + \mathbf{W}_2(\mathbf{m}_2 - \mathbf{m}_1) \quad (5)$$

and

$$\Sigma = \Sigma_1 - \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}\Sigma_1 \quad (6)$$

Scalar derivation of the optimal combination rule

If you are given two predictions or belief states about a random variable x, and the accuracy of these predications is $\hat{x}_1 \sim \mathrm{N}(m_1, \sigma_1^2)$ (the belief of expert 1) and $\hat{x}_2 \sim \mathrm{N}(m_2, \sigma_2^2)$ (the belief of expert 2), your best linear unbiased estimate (BLUE) of x is $w_1 m_1 + w_2 m_2$, with $w_1 = \sigma_2^2(\sigma_1^2 + \sigma_2^2)^{-1}$ and $w_2 = \sigma_1^2(\sigma_1^2 + \sigma_2^2)^{-1}$.

Taking into account transformations

To combine measurements it is easiest if both quantities are measurements of the same thing. This might require transforming a measurement: $\mathbf{x}' = \mathbf{T}(\mathbf{x})$. The new mean is $\mathbf{T}(\bar{\mathbf{x}})$ and the new variance is

$$\left(\frac{\partial \mathbf{T}}{\partial \mathbf{x}}\right) \sum\nolimits_{\mathbf{x}} \left(\frac{\partial \mathbf{T}}{\partial \mathbf{x}}\right)^{\mathrm{T}} \tag{7}$$

The Kalman filter (next section) uses a transformation from one time to another (the prediction step), as well as converting measurements from different sensors into corrections of the same belief state (the measurement updates).

The Kalman Filter

The Kalman Filter estimates the state of a dynamic process, using a model of sensor noise and process noise (deviations from the dynamics due to perturbations and/or modeling error). Minimizing the variance of the state estimation error (a form of optimization) drives the design.

Kalman Filter Derivation: The Prediction Step

The Kalman Filter alternates between predicting the probability distribution of the belief state on the next step (the prediction step), and incorporating an observation (the measurement update step). After a prediction step we have a belief state $\hat{\mathbf{x}} \sim \mathrm{N}(\mathbf{m}^-, \Sigma^-)$ and after a update step we have a belief state $\hat{\mathbf{x}} \sim \mathrm{N}(\mathbf{m}^+, \Sigma^+)$. The superscripts - and + keep track of whether we have incorporated the current measurement or not.

For a nonlinear discrete time system $\mathbf{F}()$, the belief state mean is propagated forward in time just using the nonlinear dynamics:

$$\mathbf{m}^-_{\mathrm{next}} = \mathbf{F}(\mathbf{m}^+, \mathbf{u}) \tag{8}$$

The variance Σ is propagated by linearizing $\mathbf{F}()$ about $\mathbf{m}$:

$$\Sigma^-_{\mathrm{next}} = \mathbf{A}\Sigma^+\mathbf{A}^{\mathrm{T}} + \Sigma_p \tag{9}$$

Σ_{p} is the variance of the process noise. The Gaussian process noise is a perturbation, or a way to model modeling error. Note that $\mathbf{u}$ plays no role in uncertainty propagation, since the commands are known perfectly and the local dynamics are linear.

Kalman Filter Derivation: The Measurement Update Step

Let's model sensor noise as additive Gaussian noise with $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \Sigma_o)$:

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{w} \tag{10}$$

$\widehat{\mathbf{C}\mathbf{x}}$ is a prediction of a measurement. One way to predict it is to use $\mathbf{C}\hat{\mathbf{x}} = \mathbf{C}\mathbf{m}^-$, which has variance $\mathbf{C}\Sigma^-\mathbf{C}^\mathrm{T}$.

Another way to predict it is to use the actual measurement $\mathbf{y}$, which has variance Σ_o.

Now we use Gaussian Fact 4 to combine these predictions. The optimal estimate is:

$$\widehat{\mathbf{C}\mathbf{x}} = \mathbf{W}_p\mathbf{C}\mathbf{m}^- + \mathbf{W}_m\mathbf{y} \tag{11}$$

where the weight on the prediction is

$$\mathbf{W}_p = \Sigma_o(\Sigma_o + \mathbf{C}\Sigma^-\mathbf{C}^\mathrm{T})^{-1} \tag{12}$$

and the weight on the measurement is

$$\mathbf{W}_m = \mathbf{C}\Sigma^-\mathbf{C}^\mathrm{T}(\Sigma_o + \mathbf{C}\Sigma^-\mathbf{C}^\mathrm{T})^{-1} \tag{13}$$

We will use the definition $\mathbf{S} = \Sigma_o + \mathbf{C}\Sigma^-\mathbf{C}^\mathrm{T}$, and the fact that symmetric matrices commute in matrix multiplication to simplify what follows

So the optimal estimate for $\widehat{\mathbf{C}\mathbf{x}}$ is:

$$\begin{aligned} \mathrm{mean}(\widehat{\mathbf{C}\mathbf{x}}) &= (1 - \mathbf{W}_m)\mathbf{C}\mathbf{m}^- + \mathbf{W}_m\mathbf{y} \\ &= \mathbf{C}\mathbf{m}^- - \mathbf{W}_m(\mathbf{C}\mathbf{m}^- - \mathbf{y}) \\ &= \mathbf{C}(\mathbf{m}^- - \Sigma^-\mathbf{C}^\mathrm{T}\mathbf{S}^{-1}(\mathbf{C}\mathbf{m}^- - \mathbf{y})) \end{aligned} \tag{14}$$

Since $\mathbf{C}$ is a constant and the $\mathrm{mean}()$ operation is linear,

$$\begin{aligned} \mathrm{mean}(\hat{\mathbf{x}}) = \mathbf{m}^+ &= \mathbf{m}^- - \Sigma^-\mathbf{C}^\mathrm{T}\mathbf{S}^{-1}(\mathbf{C}\mathbf{m}^- - \mathbf{y}) \\ &= \mathbf{m}^- - \mathbf{K}^*(\mathbf{C}\mathbf{m}^- - \mathbf{y}) \end{aligned} \tag{15}$$

so the optimal Kalman filter gain is $\mathbf{K}^* = \Sigma^-\mathbf{C}^\mathrm{T}\mathbf{S}^{-1}$.

We also need to propagate the variance

$$\mathrm{Var}(\widehat{\mathbf{Cx}}) = \mathbf{C}\Sigma^{-}\mathbf{C}^{\mathrm{T}} - \mathbf{C}\Sigma^{-}\mathbf{C}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{C}\Sigma^{-}\mathbf{C}^{\mathrm{T}} \tag{16}$$

Peeling off the left $\mathbf{C}$ and right $\mathbf{C}^{\mathrm{T}}$:

$$\mathrm{Var}(\hat{\mathbf{x}}) = \Sigma^{-} - \Sigma^{-}\mathbf{C}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{C}\Sigma^{-} \tag{17}$$

and substituting in Σ^{+} and $\mathbf{K}^{*}$ gives the update equation for Σ:

$$\Sigma^{+} = \Sigma^{-} - \mathbf{K}^{*}\mathbf{C}\Sigma^{-} \tag{18}$$

We can see that the reduction in variance of the belief state due to the Kalman Filter is $\mathbf{K}^{*}\mathbf{C}\Sigma^{-}$. Interestingly, it is proportional to the variance of the belief state before the update Σ^{-}. It makes sense that when there is no uncertainty before the update, the update can't reduce it further.