2020 March 6

Scribed by Roh

hyunwoo@uchicago.edu

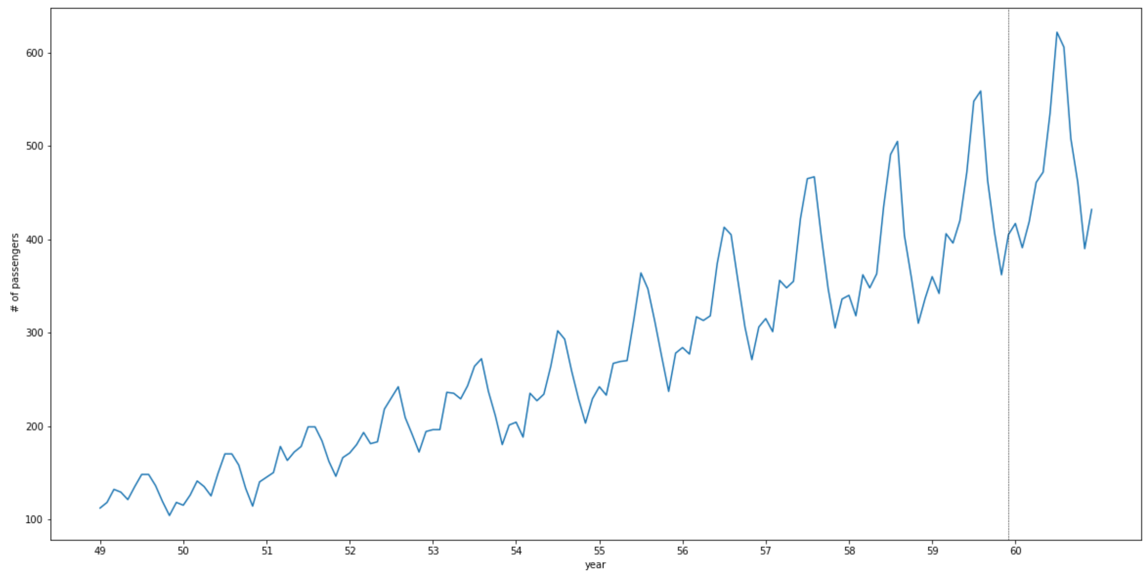
1 Better Time-series Forecasting Using Expert Knowledge (Original Source)

We have been heard that the machine learning algorithm is now seeking domain knowledge free approach so that we basically don't need any expert whose knowledge was necessary and played a core role in building up the prediction models. In other ML applications such as image classification, translation, and rule based games, the algorithm requires no prior knowledge at all. However, it still remains a very challenging task to accomplish it in time series prediction. We still believe that we need a domain knowledge and it would be better if it is from the expert. For example, long-standing senior employees in many cases have a very good overview of customer behaviour, market situation and development, economic conditions and many other important factors.

It makes sense to include this expert knowledge in the predictions of machine learning algorithm. **Then how?** The simple way of incorporating expert's opinion can be found in Black Litterman model in finance application.

There exists a forecasting model developed from Facebook called "Prophet". The model Prophet provides not only provide us with the point forecasts, but also with associated MCMC samples y_i^* from the posterior predictive distribution of each forecast step. We can guess here that we modify the posterior distribution based on the experts opinion in order to include the expert knowledge. How?

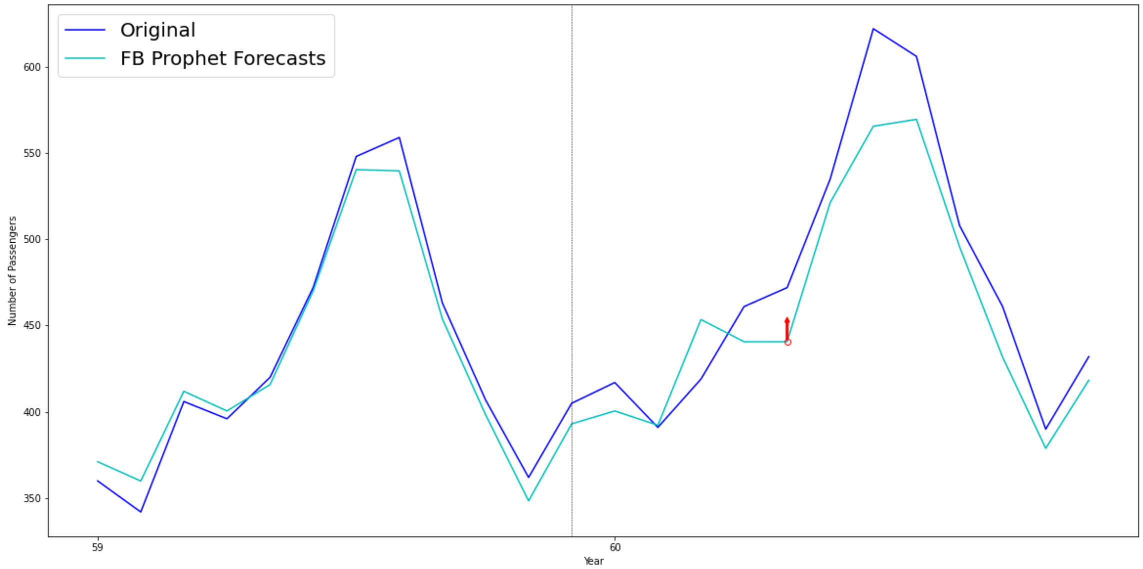
Example. Monthly totals of international airline passengers between 1949 to 1960 in thousands from which we would like to predict the year 1960



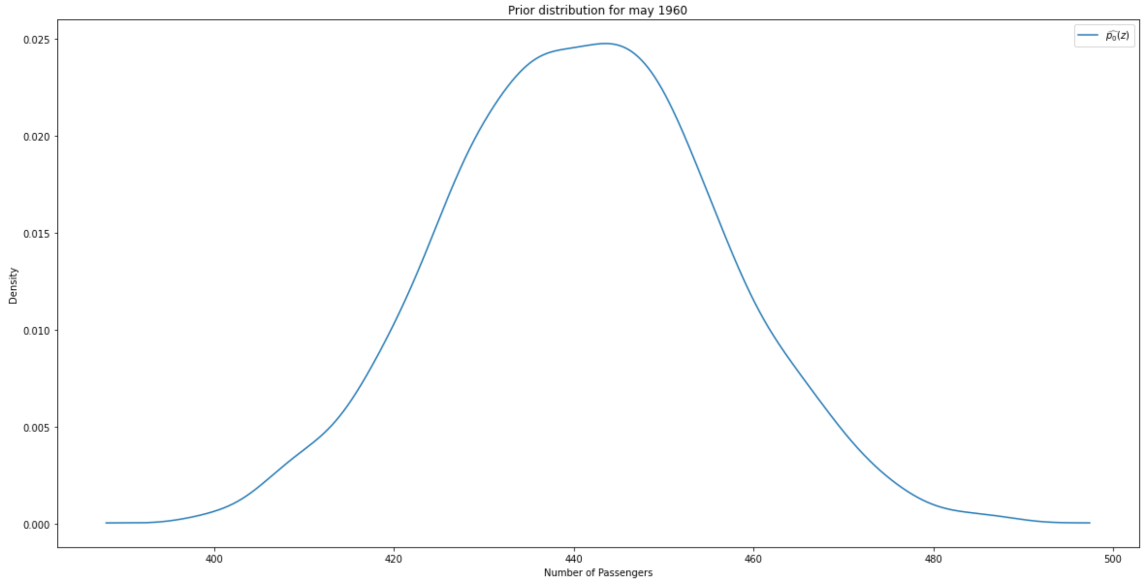
We use the first 132 months as a training set and use the last one year as a test set to evaluate the predicted values. Here, we are only interested in the value at time 1960 May.

$$Y_{1960.May} = 472, \qquad Y_{1959.May} = 420, \qquad \hat{Y}_{1960.May} = 440$$

The expected number is clearly off by large margin.



If we you see the red circle with the up-looking arrow, we can see that the predicted value is visibly off. Let's take a look at the kernel density estimate of the posterior predicted distribution $p_0(y)$ of the forecast for May 1960. We can draw a 2000 samples using the built in function.



We can see that the points are around the predicted value ≈ 440. Calculate the integral

$$\begin{aligned}\int_{-\infty}^{\infty} y p_0(y) dy &= E[Y] \\ &\approx \frac{1}{n}\sum y_i^*\end{aligned}$$

yields the point forecast for May which is $\hat{Y}_{1960.May} = 440$.

1.1 Expert's Opinion

In order to improve the forecast, it would be useful if we are able to enrich the posterior predictive distribution by expert views about future events. This does not mean that we can incorporate any type of expert's view. Here is an example of the expert's view that can bes used to modify the posterior distribution.

> **Look last May (1959 May)! We had 420.000 Passengers and we will definitely "not" have fewer in this May 1960. (['-inf', 420] has probability 1%). Furthermore, given the numbers of the last three years, I am sure that a growth rate compared to this May between 7.5% and 15% is extremely probable ([451, 483] has probability 80%). However, an increase of 15% or more compared to this May, in my opinion, is unrealistic ([483, 'inf'] has probability 1%).**

How can we include this information to the distribution and how to modify it?

1.2 Mathematical background

The starting point is continuous version of Kullback-Leibler (KL) divergence:

$$\begin{aligned}KL[p, p_0] &= KL[p||p_0] \\ &= \int_{-\infty}^{\infty} p(y)\log\frac{p(y)}{p_0(y)}dy \qquad (1) \\ &= E_p\left[\log\frac{p(y)}{p_0(y)}\right] \qquad (2)\end{aligned}$$

Given the prior $p_0(y)$, we seek the distribution $p(y)$ that minimizes the functional **KL**, given certain constraints. Here, we consider the distribution p after considering the expert's opinion closer to the true data-generating distribution for random variable Y. Be careful that we are not just approximating the prior distribution of p_0. We are targeting p_0 with given constraints. And the constraints reflect the expert's opinion. In other words, we are looking for the distribution $p(y)$ that has some pre-defined properties and comes as close as possible to our prior knowledge. The distribution $p(y)$ then is the called **Maximum Entropy distribution**. What could these constraints look like?

Before looking into the constraint, lets take a look at the KL divergence more. The log likelihood ratio can be interpreted as the amount of evidence the data provide for one model versus another ($p_0(y)$), so the KL divergence tells us how much evidence we can expect our data to provide in favor of the true model ($p(y)$). We can also interpret it as the measure of the expected number of extra bits required to code samples from $p(y)$ when using a code based on $p_0(y)$.

The integral in (1) might be rather complicated, and if we try to derive the KL divergence using brute force integration/summation it can get a little hairy. The representation in (2) makes our life a lot easier, because for many common distributions it reduces the bulk of the derivation to some simple algebra.

Now let's take a look at the possible forms of constraints on the minimization problem of KL divergence.

- The probability of 400 or fewer passengers for next July in my view is 5%.

$$\int_{-\infty}^{400} p(y)dy = 0.05$$

We can rewrite it using indicator function as

$$\begin{aligned} & E\left[I\{-\infty < y < 400\}\right] \\ =& P\left(-\infty < y < 400\right) \\ =& 0.05 \end{aligned}$$

- Corona virus is now spreading. I think with 80% probability we have between 50 and 100 passengers

$$\int_{50}^{100} p(y)dy = 0.8$$

- We have a strong growing economy, so I think with 50% probability we have between 440 and 480 passengers

$$\int_{440}^{480} p(y)dy = 0.5$$

- I expect 460 passengers

$$\int_{-\infty}^{\infty} yp(y)dy = 460$$

Above 4 are all constraints that we would like to include in approximating the distribution p_0 (= in minimizing the KL divergence). We can generally write our constraints $k = 1, 2, ..., m$ in the following form:

$$\int_{-\infty}^{\infty} F_k(y)p(y)dy = f_k$$

What does $F_k(y)$ mean? This is best understood by inspecting the third and fourth constraint-example. For the third example, it is

$$F(y) = \begin{cases} 1 & \text{, if } y \in [440, 480] \\ 0 & \text{else} \end{cases} \quad , \qquad f_k = 0.5$$

We can also think of as an indicator function where

$$\begin{aligned} & E\left[I_y\{440, 480\}\right] \\ =& P\left(440 < y < 480\right) \\ =& 0.5 \end{aligned}$$

For the fourth example, we simply have

$$F(y) = y$$

Now, we can put things into one fancy looking optimization statement.

$$\begin{aligned} p^*(y) &= arg \min KL[p, p_0] \qquad \text{subject to constraints} \\ &= arg \min \left[\int_{-\infty}^{\infty} p(y) \log \frac{p(y)}{p_0(y)} dy\right] \qquad \text{subject to} \\ & \int_{-\infty}^{\infty} F_1(y)p(y)dy = f_1 \\ & \vdots \\ & \int_{-\infty}^{\infty} F_m(y)p(y)dy = f_m \end{aligned}$$

In order to minimize the **KL** under constraints, the Lagrange multipliers $\lambda = \lambda_1, \lambda_2, ..., \lambda_m$ have to be introduced. Then we arrive at the functional:

$$\begin{aligned}\mathcal{L}[p, \lambda] = &\int_{-\infty}^{\infty} p(y) \log \frac{p(y)}{p_0(y)} dy \\ &- \lambda_1 \left(\int_{-\infty}^{\infty} F_1(y) p(y) dy - f_1 \right) \\ &\vdots \\ &- \lambda_m \left(\int_{-\infty}^{\infty} F_m(y) p(y) dy - f_m \right)\end{aligned}$$

The first step is to calculate the derivatives of $\mathcal{L}$ with respect to p and λ and to set them to zero. Beginning with the functional derivative with respect to p, we get

$$\frac{\partial \mathcal{L}}{\partial p} = \underbrace{\log \frac{p(y)}{p_0(y)} + 1}_{\text{how?}} - \lambda_1 F_1(y) - \cdots - \lambda_m F_m(y) = 0$$

How can we get this? To be continued on DM_Mar_9.

Reference:

1. https://blog.codecentric.de/en/2019/02/forecasts-machine-learning-facebook-prophet-maximum-entropy/

2. Technical Notes on Kullback-Leibler Divergence by Alexander Etz

3. https://www.shakirm.com/papers/VITutorial.pdf

4. https://physicspages.com/pdf/Lancaster%20QFT/Lancaster%20Problems%2001.03-04.pdf