Fisher Information Matrix

Suppose we have a model parameterized by parameter vector θ that models a distribution $p(x|\theta)$. In frequentist statistics, the way we learn θ is to maximize the likelihood $p(x|\theta)$ wrt. parameter θ. To assess the goodness of our estimate of θ we define a score function:

$$s(\theta) = \nabla_\theta \log p(x|\theta)\,,$$

that is, score function is the gradient of log likelihood function. The result about score function below is important building block on our discussion.

Claim: The expected value of score wrt. our model is zero.

Proof. Below, the gradient is wrt. θ.

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(x|\theta)}[s(\theta)] &= \mathop{\mathbb{E}}_{p(x|\theta)}[\nabla \log p(x|\theta)] \\
&= \int \nabla \log p(x|\theta)\, p(x|\theta)\, \mathrm{d}x \\
&= \int \frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta)\, \mathrm{d}x \\
&= \int \nabla p(x|\theta)\, \mathrm{d}x \\
&= \nabla \int p(x|\theta)\, \mathrm{d}x \\
&= \nabla 1 \\
&= 0
\end{aligned}$$

□

But how certain are we to our estimate? We can define an uncertainty measure around the expected estimate. That is, we look at the covariance of score of our model. Taking the result from above:

$$\mathop{\mathbb{E}}_{p(x|\theta)}\left[(s(\theta) - 0)\,(s(\theta) - 0)^{\mathrm{T}}\right].$$

We can then see it as an information. The covariance of score function above is the definition of Fisher Information. As we assume $\boldsymbol{\theta}$ is a vector, the Fisher Information is in a matrix form, called Fisher Information Matrix:

$$\mathrm{F} = \mathop{\mathbb{E}}_{p(x|\theta)}\left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^{\mathrm{T}}\right].$$

However, usually our likelihood function is complicated and computing the expectation is intractable. We can approximate the expectation in F using empirical distribution $q(x)$, which is given by our training data $X = \{x_1, x_2, \cdots, x_N\}$. In this form, F is called Empirical Fisher:

$$\mathrm{F} = \frac{1}{N} \sum_{i=1}^{N} \nabla \log p(x_i|\theta)\, \nabla \log p(x_i|\theta)^{\mathrm{T}}\,.$$

Fisher and Hessian

One property of F that is not obvious is that it has the interpretation of being the negative expected Hessian of our model's log likelihood.

Claim: The negative expected Hessian of log likelihood is equal to the Fisher Information Matrix F.

Proof. The Hessian of the log likelihood is given by the Jacobian of its gradient:

$$\begin{aligned}
\mathrm{H}_{\log p(x|\theta)} &= \mathrm{J}\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right) \\
&= \frac{\mathrm{H}_{p(x|\theta)}\, p(x|\theta) - \nabla p(x|\theta)\, \nabla p(x|\theta)^{\mathrm{T}}}{p(x|\theta)\, p(x|\theta)} \\
&= \frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)} - \frac{\nabla p(x|\theta)}{p(x|\theta)} \frac{\nabla p(x|\theta)^{\mathrm{T}}}{p(x|\theta)} \\
&= \frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}},
\end{aligned}$$

where the second line is a result of applying quotient rule of derivative. Taking expectation wrt. our model, we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(x|\theta)}\left[\mathrm{H}_{\log p(x|\theta)}\right] &= \mathop{\mathbb{E}}_{p(x|\theta)}\left[\frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \mathop{\mathbb{E}}_{p(x|\theta)}\left[\frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)}\right] - \mathop{\mathbb{E}}_{p(x|\theta)}\left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \int \frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)} p(x|\theta)\, \mathrm{d}x - \mathop{\mathbb{E}}_{p(x|\theta)}\left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^{\mathrm{T}}\right] \\
&= \mathrm{H}_{\int p(x|\theta)\, \mathrm{d}x} - \mathrm{F} \\
&= \mathrm{H}_1 - \mathrm{F} \\
&= -\mathrm{F}\,.
\end{aligned}$$

Thus we have $\mathrm{F} = -\mathbb{E}_{p(x|\theta)}\left[\mathrm{H}_{\log p(x|\theta)}\right]$. □

Indeed knowing this result, we can see the role of F as a measure of curvature of the log likelihood function.

Conclusion

Fisher Information Matrix is defined as the covariance of score function. It is a curvature matrix and has interpretation as the negative expected Hessian of log likelihood function. Thus the immediate application of F is as drop-in replacement of H in second order optimization methods.

One of the most exciting results of F is that it has connection to KL-divergence. This gives rise to natural gradient method, which we shall discuss further in the next article.

References

1. Martens, James. "New insights and perspectives on the natural gradient method." arXiv preprint arXiv:1412.1193 (2014).
2. Ly, Alexander, et al. "A tutorial on Fisher information." Journal of Mathematical Psychology 80 (2017): 40-55.

← PREVIOUS POST

NEXT POST →