

Draft

13. maj 2014

1 The derivative of prediction or Sensitivity

We wish to find the effect that a datapoint's class has on the predicted class for that datapoint.

$$\frac{\delta \hat{Y}_n}{\delta Y_n} \quad (1.1)$$

Our prediction is

$$\hat{Y}_n = p(y|\bar{x}, \bar{w}) \quad (1.2)$$

where \bar{w} is subject to

$$\frac{\delta L}{\delta \bar{w}} = 0 \quad (1.3)$$

Which means that we have found a locally optimal solution.

We now assume that when we move y by a small amount δy then 1.3 still holds. (can we do this with a discrete y ?)

Essentially assuming some smoothness around the optimum.

Using this and the fact that 1.3 depends both directly and indirectly on y we see that

$$\begin{aligned} \frac{\delta}{\delta y} \frac{\delta L}{\delta \bar{w}} &= 0 \\ \Downarrow \\ \frac{\delta^2 L}{\delta y \delta \bar{w}} + \frac{\delta^2 L}{\delta \bar{w} \delta \bar{w}^T} \frac{\delta \bar{w}}{\delta y} &= 0 \end{aligned}$$

and from this we can isolate

$$\frac{\delta \bar{w}}{\delta y} = - \left[\frac{\delta^2 L}{\delta \bar{w} \delta \bar{w}^T} \right]^{-1} \frac{\delta^2 L}{\delta y \delta \bar{w}} \quad (1.4)$$

Rewriting (1.1) we get

$$\frac{\delta \hat{Y}_n}{\delta Y_n} = \frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta Y_n} = \frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T} \frac{\delta \bar{w}}{\delta y} \quad (1.5)$$

And inserting (1.4)

$$\frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T} \frac{\delta \bar{w}}{\delta y} = - \frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T} \left[\frac{\delta^2 L}{\delta \bar{w} \delta \bar{w}^T} \right]^{-1} \frac{\delta^2 L}{\delta y \delta \bar{w}} \quad (1.6)$$

And this is our leverage score for this where \bar{w} is gained from a training based on a small dataset.

2 Randomised algorithm

Uncertainty based on asymptotic likelihood and \bar{w} -distribution

Let \mathcal{L}_∞ be the log-likelihood function for a distribution, now let \mathcal{L}_N denote the log-likelihood function based on N observations from this distribution. Furthermore, let N be a large number, for which $L_N \approx L_\infty$.

$$\mathcal{L}_N = \frac{1}{N} \sum_{n=1}^N \ell_n \quad \bar{w} \text{ s.t. } \frac{\delta \mathcal{L}}{\delta \bar{w}} = \bar{0} \quad (2.1)$$

Where ℓ_n is the log-likelihood of the n^{th} observation. And \bar{w} is the true weights for the distribution, then we combine the expressions from (2.1), such that for the true weights the following must be fulfilled:

$$\frac{1}{N} \sum_{n=1}^N \frac{\delta \ell_n}{\delta \bar{w}} = 0 \quad (2.2)$$

(Skal vi lige skrive lidt om at $\Delta w = w - w_0$ og er en lille forskydelse i vægtene? Eller er det en lille forskydelse?) For each of the N observations, we can approximate the log-likelihood of the n^{th} observation with this Taylor expansion:

$$\ell_n(\bar{w}) = \ell_n(\bar{w}_0) + \frac{\delta \ell_n}{\delta \bar{w}} \Big|_{\bar{w}_0} \Delta \bar{w} + \frac{1}{2} Tr \left[\frac{\delta^2 \ell_n}{\delta \bar{w} \delta \bar{w}^T} \Big|_{\bar{w}_0} \Delta \bar{w} \Delta \bar{w}^T \right] \quad (2.3)$$

Or for the entire log-likelihood function: **Where did the trace go ?**

$$\mathcal{L}_N(\bar{w}) = \mathcal{L}_N(\bar{w}_0) + \left(\frac{\delta \mathcal{L}_N}{\delta \bar{w}} \Big|_{\bar{w}_0} \right)^T \cdot \Delta \bar{w} + \frac{1}{2} \Delta \bar{w}^T \left(\frac{\delta^2 \mathcal{L}_N}{\delta \bar{w} \delta \bar{w}^T} \Big|_{\bar{w}_0} \right) \Delta \bar{w} + R \quad (2.4)$$

Where R is the error of the approximation and assumed to be 0. Furthermore, we define $\bar{H}_N = \frac{\delta^2 \mathcal{L}_N}{\delta \bar{w} \delta \bar{w}^T} \Big|_{\bar{w}_0}$, and $\bar{g} : N = \frac{\delta \mathcal{L}_N}{\delta \bar{w}} \Big|_{\bar{w}_0}$. And evaluate condition (2.1) on \bar{w} :

$$\frac{\delta \mathcal{L}_N}{\delta \bar{w}} = \bar{g}_N + \bar{H}_N \Delta \bar{w} = \bar{0} \quad (2.5)$$

2.1 Covariance of \bar{w} - distribution

We replace $\Delta\bar{w}$ with $\hat{\Delta}\bar{w}$ as N is a finite number, thus only approximating $\Delta\bar{w}$. Isolating $\hat{\Delta}\bar{w}$, and using Ljung [REFERENCE?], we get: **Is this to soon to involve Ljung?**

$$\hat{\Delta}\bar{w} = -\bar{H}_N^{-1} \cdot \bar{g}_N \stackrel{\text{Ljung}}{=} -\bar{H}_0^{-1} \cdot \bar{g}_{\bar{w}}(\bar{w}_0) \quad (2.6)$$

(Forklaring af at H_0 er uafhængig af datasæt, mens g nu er afhængig af w evalueret i w_0) Besides getting an estimate for $\hat{\Delta}\bar{w}$, we can find the mean of the distribution:

$$\langle \hat{\Delta}\bar{w} \rangle = -\bar{H}_0^{-1} \bar{g}_0 = 0$$

As $\delta\bar{w} = \bar{w} - \bar{w}_0$??mistet tråden?

2.1 Covariance of \bar{w} - distribution

Why do we do this???

$$\langle \delta\bar{w} \delta\bar{w}^T \rangle_N = \langle \bar{H}^{-1} \bar{g} \bar{g}^T \bar{H}^{-1} \rangle \stackrel{\text{Ljung}}{=} \bar{H}_0^{-1} \langle \bar{g} \bar{g}^T \rangle \bar{H}_0^{-1} + R' \quad (2.7)$$

With error $R' = O(\frac{1}{N}) \approx 0$, for large N . We look at the covariance of the gradient function

$$\begin{aligned} \langle \bar{g} \bar{g}^T \rangle_N &= \frac{1}{N^2} \sum_{n,n'=1}^N \left\langle \frac{\delta \ell_n}{\delta \bar{w}} \bigg|_{\bar{w}_0} \frac{\delta \ell_{n'}}{\delta \bar{w}} \bigg|_{\bar{w}_0} \right\rangle \\ &= \frac{1}{N^2} \left(\underbrace{\sum_{n \neq n'} \left\langle \frac{\delta \ell_n}{\delta \bar{w}} \bigg|_{\bar{w}_0} \right\rangle \cdot \left\langle \frac{\delta \ell_{n'}}{\delta \bar{w}} \bigg|_{\bar{w}_0} \right\rangle}_0 + \sum_{n=1}^N \left\langle \frac{\delta \ell_n}{\delta \bar{w}} \bigg|_{\bar{w}_0} \frac{\delta \ell_n}{\delta \bar{w}^T} \bigg|_{\bar{w}_0} \right\rangle \right) \end{aligned} \quad (2.8)$$

(2.9)

Due to the assumption of independence, only the N diagonal elements are non-zero. So;

$$\langle \bar{g} \bar{g}^T \rangle_N = \frac{1}{N} \left\langle \frac{\delta \mathcal{L}}{\delta \bar{w}} \bigg|_{\bar{w}_0} \frac{\delta \mathcal{L}}{\delta \bar{w}^T} \bigg|_{\bar{w}_0} \right\rangle \quad (2.10)$$

2.2 Proof that $\left\langle \frac{\delta \mathcal{L}}{\delta \bar{w}} \bigg|_{\bar{w}_0} \frac{\delta \mathcal{L}}{\delta \bar{w}^T} \bigg|_{\bar{w}_0} \right\rangle = \bar{H}_0$

Tekst test

$$\langle \bar{g} \bar{g}^T \rangle_N = \frac{1}{N^2} \sum_{n=1}^N \int_{\Omega} \frac{\delta \ell_n(\bar{x})}{\delta \bar{w}} \bigg|_{\bar{w}_0} \frac{\delta \ell_n(\bar{x})}{\delta \bar{w}^T} \bigg|_{\bar{w}_0} p(\bar{x}) d\bar{x} \quad (2.11)$$

2.3 Uncertainty of prediction

From (2.10) and (2.11), and setting $\ell_n(\bar{x}) = p(\bar{x})$:

$$\bar{H}\Big|_{\bar{w}_0} = \frac{1}{N} \sum_{n=1}^N \int_{\Omega} \frac{\delta}{\delta \bar{w} \delta \bar{w}^T} - \log p(\bar{x}|\bar{w}) p(\bar{x}) \delta \bar{x} \quad (2.12)$$

$$= \frac{1}{N} \sum_{n=1}^N \int_{\Omega} -\frac{\delta}{\delta \bar{w}} \frac{1}{p(\bar{x})} \frac{\delta}{\delta \bar{w}^T} p(\bar{x}|\bar{w}) p(\bar{x}) \delta \bar{x} \quad (2.13)$$

$$(2.14)$$

Now if $p(\bar{x}|\bar{w}_0) = p(\bar{x})$, then

2.3 Uncertainty of prediction

For a number of weight-vectors \bar{w} , we take the mean of predictions based on these weight-vectors;

$$\langle p(y|\bar{x}, \bar{w}) \rangle \approx p(y|\bar{x}, \hat{\bar{w}}) = p(y|\bar{x}, \mathbf{E}(\bar{w})) \quad (2.15)$$

We now look at a small change in the prediction Δp , caused by a change of $\Delta \bar{w}$ in true weight vector \bar{w}_0 .

$$\Delta p = p(y|\bar{x}, \bar{w}_0 + \Delta \bar{w}) - p(y|\bar{x}, \bar{w}_0) \approx \left. \frac{\delta p}{\delta \bar{w}} \right|_{\bar{w}_0} \cdot \Delta \bar{w} \quad (2.16)$$

The variance of Δp , can then be computed as

$$\langle (\Delta p)^2 \rangle = Tr \left[\left. \frac{\delta p}{\delta \bar{w}} \left(\frac{\delta p}{\delta \bar{w}} \right)^T \langle \Delta \bar{w} \Delta \bar{w}^T \rangle \right] = \frac{1}{N} \left(\frac{\delta p}{\delta \bar{w}} \right)^T \bar{H}^{-1} \frac{\delta p}{\delta \bar{w}} \quad (2.17)$$

2.3.1 For a linear model with known σ^2

The prediction in a linear model is:

$$p(y|\bar{x}, \bar{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(\bar{x}|\bar{w}))^2}{2\sigma^2}} \quad (2.18)$$

Where y is the target and $f(\bar{x}|\bar{w})$ is the prediction. Differentiating (??) with respect to \bar{w} : (Hvorfor er det vi gør det??)

$$\frac{\delta p}{\delta \bar{w}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(\bar{x}|\bar{w}))^2}{2\sigma^2}} - (y - f(\bar{x}|\bar{w})) \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (2.19)$$

We let $y = f(\bar{x}|\bar{w}) + \epsilon$. (Targets kan beskrives som en approximativ funktion + en fejl ..)

$$\frac{\delta p}{\delta \bar{w}} = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}}_{\text{const. w.r.t. } \bar{x}} \epsilon^2 \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}}^T \bar{H}_0^{-1} \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (2.20)$$