Draft

7. 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.

$$\frac{\delta \hat{Y}_n}{\delta Y_n} \tag{1.1}$$

Our prediction is

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

where $\bar{w}$ is subject to

$$\frac{\delta L}{\delta \bar{w}} = 0 \tag{1.3}$$

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

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

and from this we can isolate

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

Rewriting (1.1) we get

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

And inserting (1.4)

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

And this is our leverage score for this

2 Randomised algorithm

Uncertainty based on asymptotic likelihood and $\bar{w}$-distribution
Let $\mathcal{L}_\infty$ 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 $L_N \approx L_\infty$.

$$\mathcal{L}_\mathcal{N} = \frac{1}{N} \sum_{n=1}^{N} \ell_n \qquad \bar{w}\, s.t. \frac{\delta \mathcal{L}}{\delta \bar{w}} = \bar{0} \tag{2.1}$$

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:

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

(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:

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

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

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

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

$$\frac{\delta \mathcal{L}_\mathcal{N}}{\delta \bar{w}} = \bar{g}_N + \bar{\bar{H}}_N \Delta \bar{w} = \bar{0} \tag{2.5}$$

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

$$\Delta\hat{\bar{w}} = -\bar{\bar{H}}_N^{-1} \cdot \bar{g}_N \overbrace{=}^{Ljung} -\bar{\bar{H}}_0^{-1} \cdot \bar{g}_{\bar{w}}\left(\bar{w}_0\right) \tag{2.6}$$

(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 $\Delta\hat{\bar{w}}$, we can find the mean of the distribution:

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

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

2.1 Covariance of $\bar{w}$ - distribution

Why do we do this???

$$\left\langle \delta\bar{w}\delta\bar{w}^T \right\rangle_N = \left\langle \bar{\bar{H}}^{-1}\bar{g}\bar{g}^T\bar{\bar{H}}^{-1} \right\rangle \overbrace{=}^{Ljung} \bar{\bar{H}}_0^{-1}\left\langle \bar{g}\bar{g}^T \right\rangle \bar{\bar{H}}_0^{-1} + R' \tag{2.7}$$

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

$$\left\langle \bar{g}\bar{g}^T \right\rangle_N = \frac{1}{N^2}\sum_{n,n'=1}^{N}\left\langle \left.\frac{\delta\ell_n}{\delta_n\bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta\ell_{n'}}{\delta_n\bar{w}}\right|_{\bar{w}_0} \right\rangle \tag{2.8}$$

$$= \frac{1}{N^2}\left(\underbrace{\sum_{n\neq n'}\left\langle \left.\frac{\delta\ell_n}{\delta\bar{w}}\right|_{\bar{w}_0} \right\rangle \cdot \left\langle \left.\frac{\delta\ell_{n'}}{\delta\bar{w}^T}\right|_{\bar{w}_0} \right\rangle}_{0} + \sum_{n=1}^{N}\left\langle \left.\frac{\delta\ell_n}{\delta\bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta\ell_n}{\delta\bar{w}^T}\right|_{\bar{w}_0} \right\rangle \right) \tag{2.9}$$

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

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

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

Tekst test

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

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

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

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

$$\tag{2.14}$$

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

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

We now introduce