Lecture 18

Bayesian neural networks

So far, we have used ML techiques to find weights and biases. Let's now consider a Bayesian framework:

Regression

$$p(\underset{\substack{\uparrow\\ \text{single}\\ \text{output}\\ \text{(for simplicity)}}}{t}\,|\,\underset{\substack{\uparrow\\ \text{input}}}{\vec{x}}, \underset{\substack{\nwarrow\\ \text{weights/biases}}}{\vec{w}}, \overset{\substack{\text{precision}\\ \swarrow}}{\beta}) = \mathcal{N}(t\,|\,\underbrace{y(\vec{x},\vec{w})}_{\text{neural network model}}, \overset{\sigma^2}{\beta^{-1}})$$

Choose a prior:

$$p(\vec{w}|\alpha) = \mathcal{N}(\vec{w}|\vec{0}, \alpha^{-1}\mathbb{I})$$

N observations: $\vec{x}_1 \dots \vec{x}_N$ $\overset{\text{continuous vars}}{\swarrow}$

Target values: $\mathcal{D} = \{t_1 \dots t_N\}$

Then $\mathcal{L} = p(\mathcal{D}|\vec{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(\vec{x}_n,\vec{w}),\beta^{-1})$,

and the posterior $\overset{\swarrow \text{ wrt } \vec{w} \text{ only}}{\text{is}}$ given by

$$p(\vec{w}|\mathcal{D},\alpha,\beta) \sim \underbrace{\overbrace{p(\vec{w}|\alpha)}^{\text{gaussian}}\,p(\mathcal{D}|\vec{w},\beta)}_{\text{non-Gaussian since}}$$

$y(\vec{x},\vec{w})$ is a non-linear f'n of $\vec{w}$

$\Downarrow$

need to use Laplace approx'n

Consider

$$\log p(\vec{w}|\mathcal{D},\alpha,\beta) = -\frac{\alpha}{2}\vec{w}^T\vec{w} - \underbrace{\frac{\beta}{2}\sum_{n=1}^{N}(y(\vec{x}_n,\vec{w})-t_n)^2}_{\beta E} + $$
$$+ \text{const}(\vec{w})$$

↳ can find $\vec{w}_{MAP}$ which maximizes this (assume α & β are fixed for now) by error backpropagation + conjugate grad. Note that $\vec{w}_{MAP}$ is a <u>local</u> max in general.

Next, compute

$$\underset{\substack{\nearrow \\ \text{needed for Laplace appr'n}}}{A_{ij}} = -\frac{\partial^2}{\partial w_i \partial w_j}\log p(\vec{w}|\mathcal{D},\alpha,\beta) =$$

$$= \alpha\,\delta_{ij} + \beta\,\underbrace{H_{ij}}_{=\frac{\partial^2 E}{\partial w_i \partial w_j}}, \quad \underbrace{\text{Hessian}}_{\text{already discussed how to compute it efficiently}}$$

Recall that under Laplace,

$$f(\vec{z}) \Rightarrow q(\vec{z}) = \mathcal{N}(\vec{z}\,|\,\vec{z}_0, A^{-1})$$

So, the posterior becomes gaussian:

$$p(\vec{w}|\mathcal{D},\alpha,\beta) \Rightarrow q(\vec{w}|\mathcal{D}) = \mathcal{N}(\vec{w}\,|\,\vec{w}_{MAP}, A^{-1}).$$

The predictive distribution is

$$p(t|\vec{x},\mathcal{D}) = \int d\vec{w}\; \overbrace{p(t|\vec{x},\vec{w})}^{\text{non-gaussian}}\; \overbrace{q(\vec{w}|\mathcal{D})}^{\text{gaussian}}$$

$\uparrow$ explicit α, β dependence suppressed for brevity

If $q(\vec{w}|\mathcal{D})$ is sufficiently narrow,

$$y(\vec{x},\vec{w}) \simeq y(\vec{x},\vec{w}_{MAP}) + \vec{g}^{T}\cdot(\vec{w}-\vec{w}_{MAP}),$$

where

$$\vec{g} = \vec{\nabla}_{\vec{w}}\, y(\vec{x},\vec{w})\Big|_{\vec{w}=\vec{w}_{MAP}}$$

Then $p(t|\vec{x},\vec{w},\beta) \simeq \mathcal{N}(t|y(\vec{x},\vec{w}_{MAP}) +$

$+\,\vec{g}^{T}\cdot(\vec{w}-\vec{w}_{MAP}),\ \beta^{-1})$.

<u>Recall</u>:

$$\begin{cases} p(\vec{x}) = \mathcal{N}(\vec{x}|\vec{\mu},\Lambda^{-1}) \\ p(\vec{y}|\vec{x}) = \mathcal{N}(\vec{y}|A\vec{x}+\vec{b},\ L^{-1}) \end{cases}$$

$$\hookrightarrow\ p(\vec{y}) = \int d\vec{x}\; p(\vec{y}|\vec{x})\,p(\vec{x}) =$$

$$= \mathcal{N}(\vec{y}|A\vec{\mu}+\vec{b},\ L^{-1}+A\Lambda^{-1}A^{T})$$

Here,

$$\begin{cases} \vec{x} \to \vec{w} \\ p(\vec{x}) \to q(\vec{w}|\mathcal{D}) \to \begin{cases} \vec{\mu} = \vec{w}_{MAP}, \\ \Lambda^{-1} = A^{-1} \end{cases} \\ \vec{y} \to t \\ p(\vec{y}|\vec{x}) \to \begin{cases} \vec{b} \to y(\vec{x},\vec{w}_{MAP}) - \overset{\text{scalar}}{\vec{g}^{T}\cdot\vec{w}_{MAP}} \\ A \to \vec{g}^{T}, \quad L^{-1} \to \beta^{-1} \end{cases} \end{cases}$$

Consequently,

$$p(t|\vec{x},\mathcal{D},\underbrace{\alpha,\beta}_{\text{restored explicit dependence}}) = \mathcal{N}(t|y(\vec{x},\vec{w}_{MAP}),\sigma^2(\vec{x}))$$

$$\left\{\begin{aligned} A\vec{\mu}+\vec{b} &\to \vec{g}^T\cdot\vec{w}_{MAP} + y(\vec{x},\vec{w}_{MAP}) - \vec{g}^T\cdot\vec{w}_{MAP} = \\ &= y(\vec{x},\vec{w}_{MAP}) \\ L^{-1}+A\Lambda^{-1}A^T &\to \underbrace{\beta^{-1}}_{\substack{\text{uncertainty}\\ \text{due to noise}\\ \text{in target}\\ \text{variables}}} + \overbrace{\underbrace{\vec{g}^TA^{-1}\vec{g}}_{\sum_{i,j} g_i A^{-1}_{ij} g_j}}^{\text{uncertainty due to uncertainty in } \vec{w}} \equiv \underbrace{\sigma^2(\vec{x})}_{\substack{\vec{x}\text{-dependence}\\ \text{through}\\ \vec{g}=\vec{g}(\vec{x})}} \end{aligned}\right.$$

How to find α and β?

Consider
$$\underbrace{p(\mathcal{D}|\alpha,\beta)}_{\substack{\text{evidence for } \alpha,\beta\\ \text{[normalization factor}\\ \text{for posterior in weights]}}} = \int d\vec{w}\,\underbrace{p(\mathcal{D}|\vec{w},\beta)}_{\mathcal{L}}\,\underbrace{p(\vec{w}|\alpha)}_{\text{prior}}$$

Recall that under Laplace appr'n:

$$\int \underbrace{d\vec{z}}_{M \text{ dims}}\, f(\vec{z}) \simeq f(\vec{z}_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}}, \text{ where}$$

$$A = -\vec{\nabla}\vec{\nabla}\log f(\vec{z})\Big|_{\vec{z}=\vec{z}_0}$$

Here,
$$\log p(\mathcal{D}|\alpha,\beta) \simeq -\frac{\beta}{2}\sum_{n=1}^{N}(y(\vec{x}_n,\vec{w}_{MAP})-t_n)^2 - \frac{\alpha}{2}\vec{w}^T_{MAP}\cdot\vec{w}_{MAP} - \frac{N}{2}\log(2\pi\beta^{-1}) - \frac{W}{2}\log(2\pi\alpha^{-1}) \oplus$$
$$\oplus\ \frac{W}{2}\log(2\pi) - \frac{1}{2}\log|A| \ ⊜$$

$$\textcircled{\equiv} \quad -E(\vec{w}_{MAP}) - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log\beta + \frac{W}{2}\log\alpha - $$

$$\underline{\underline{-\frac{1}{2}\log|A|}} \, . \qquad W = \text{total \# prms in } \vec{w}$$

This is very similar to (3.82) & (3.86).

We need to maximize $\log p(\mathcal{D}|\alpha,\beta)$ to get point (ML) estimates $\hat{\alpha}$ & $\hat{\beta}$.

Define $$(\beta \underset{\substack{\uparrow \\ \text{Hessian} \\ \text{at } \vec{w}_{MAP}}}{H})\vec{u}_i = \lambda_i \vec{u}_i \qquad i=1,\dots,W$$

Then, as before, $$\frac{\partial}{\partial\alpha}\log|A| = \sum_i \frac{1}{\lambda_i+\alpha}$$

$$\frac{\partial}{\partial\alpha}\log p(\mathcal{D}|\alpha,\beta) \simeq \frac{W}{2\alpha} - \frac{1}{2}\vec{w}^T_{MAP}\cdot\vec{w}_{MAP} -$$

$$-\frac{1}{2}\sum_i \frac{1}{\lambda_i+\alpha} = 0, \text{ or}$$

$$\alpha\, \vec{w}^T_{MAP}\cdot\vec{w}_{MAP} = W - \alpha\sum_i \frac{1}{\lambda_i+\alpha} =$$

$$= \sum_i \frac{\lambda_i}{\lambda_i+\alpha} = \underline{\underline{\gamma}} \, .$$

$$\text{So, } \underbrace{\hat{\alpha} = \frac{\gamma}{\vec{w}^T_{MAP}\cdot\vec{w}_{MAP}}}_{} \, .$$

$$\text{Ignored that } \begin{cases} H = H(\alpha) \Rightarrow \lambda_i = \lambda_i(\alpha) \\ \vec{w}_{MAP} = \vec{w}_{MAP}(\alpha) \end{cases}$$

Similarly,

$$\frac{\partial}{\partial\beta}\log P(\mathcal{D}|\alpha,\beta) \simeq \frac{N}{2\beta} - \frac{1}{2}\sum_n \left(y(\vec{x}_n, \vec{w}_{MAP}) - t_n\right)^2 - \frac{1}{2}\frac{\gamma}{\beta} = 0, \text{ or}$$

Note that

$$\lambda_i \sim \beta \Rightarrow \frac{d\lambda_i}{d\beta} = \frac{\lambda_i}{\beta}.$$

Then $\displaystyle \frac{d}{d\beta}\log|A| = \sum_i \frac{d\lambda_i/d\beta}{\lambda_i+\alpha} = \frac{1}{\beta}\underbrace{\sum_i \frac{\lambda_i}{\lambda_i+\alpha}}_{\gamma}.$

$$\frac{N-\gamma}{\beta} = \sum_n (y - t_n)^2,$$

$$\hat{\beta}^{-1} = \frac{1}{N-\gamma}\sum_n (y-t_n)^2.$$

[Alternate between estimating $\vec{w}_{MAP}$ & $\hat{\alpha}, \hat{\beta}$.]

We can use $p(\mathcal{D}|\hat{\alpha},\hat{\beta})$ to compare e.g. two-layer networks with different M (# hidden units). Recall the $M!\,2^M$ symmetry factor $\Rightarrow$ compare $M!\,2^M p(\mathcal{D}|\hat{\alpha},\hat{\beta})$ when M is changed.

Classification

Consider a single sigmoid output in the $K=2$ classification problem:

$$\log \mathcal{L} = \log p(\mathcal{D}|\vec{w}) = \sum_n \left[t_n \log y_n + (1-t_n)\log(1-y_n) \right],$$

where $t \in (0,1)$ & $y_n = y(\vec{x}_n, \vec{w})$

Then we need to maximize the posterior:

$$\log P(\vec{w}|\mathcal{D},\alpha) = -\frac{\alpha}{2}\vec{w}^T\vec{w} + \log p(\mathcal{D}|\vec{w}) + \text{const}(\vec{w})$$

$\downarrow$

for a given α, find $\vec{w}_{MAP}$ by error back propagation + e.g. conjugate grad.

Next, find H at $\vec{w}_{MAP}$, use

$$A = \alpha \mathbb{I} + \underset{-\vec{\nabla}_{\vec{w}}\vec{\nabla}_{\vec{w}} \log P(\mathcal{D}|\vec{w})}{\underset{\shortparallel}{H}}$$ as before to get

$$P(\vec{w}|\mathcal{D},\alpha) \Rightarrow \underbrace{q(\vec{w}|\mathcal{D})}_{\text{gaussian approx'n to posterior}} = \mathcal{N}(\vec{w}|\vec{w}_{MAP}, A^{-1})$$

As before, under Laplace appr'n:

$$p(\mathcal{D}|\alpha) = \int d\vec{w}\, p(\mathcal{D}|\vec{w})\, p(\vec{w}|\alpha)$$

is given by

$$\log p(\mathcal{D}|\alpha) \simeq -\frac{\alpha}{2}\vec{w}^T_{MAP}\vec{w}_{MAP} + \sum_n \left[t_n \log y_n + (1-t_n)\log(1-y_n) \right] - \frac{W}{2}\log(2\pi\alpha^{-1}) \oplus$$

$\uparrow$ here, $y_n = y(\vec{x}_n, \vec{w}_{MAP})$

$$\oplus\ \frac{W}{2}\log(2\pi) - \frac{1}{2}\log|A| \quad , \quad \text{s.t.}$$

$$\frac{\partial}{\partial\alpha}\log P(\mathcal{D}|\alpha) \simeq -\frac{1}{2}\vec{w}^{T}_{MAP}\cdot\vec{w}_{MAP} + \frac{W}{2\alpha} -$$

$$-\frac{1}{2}\sum_{i}\frac{1}{\lambda_i+\alpha} = 0 \quad \text{gives}$$

$$\hat{\alpha} = \frac{\gamma}{\vec{w}^{T}_{MAP}\cdot\vec{w}_{MAP}} \quad \text{as before}$$

We need to alternate $\vec{w}_{MAP}$ & $\hat{\alpha}$ estimates.

Finally, we need the predictive distr'n:

$$p(\underset{\overset{\shortparallel}{1}}{t}|\vec{x},\mathcal{D}) = \int d\vec{w}\, \underbrace{p(t|\vec{x},\vec{w})}_{p(t=1|\vec{x},\vec{w})=\sigma(a(\vec{x},\vec{w}))}\ \underbrace{q(\vec{w}|\mathcal{D})}_{\text{gaussian}}.$$

Consider $\underbrace{a(\vec{x},\vec{w})}_{\text{output unit activation}} \simeq a(\vec{x},\vec{w}_{MAP}) + \vec{b}^{T}\cdot(\vec{w}-\vec{w}_{MAP})$, $\quad \vec{b} = \vec{\nabla} a(\vec{x},\vec{w})\Big|_{\vec{w}_{MAP}}$

(cannot expand $y(\vec{x},\vec{w})$ as in regression due to an extra sigmoid)

Use results of (4.5.2):

$$p(t=1|\vec{x},\mathcal{D}) \simeq \int d\vec{w}\, \sigma\big(\overbrace{a(\vec{x},\vec{w}_{MAP}) + \vec{b}^{T}\cdot(\vec{w}-\vec{w}_{MAP})}^{\text{linear in } \vec{w}}\big)\times$$

$$\times\, q(\vec{w}|\mathcal{D}) ⊜$$

$$\textcircled{=} \int d\tilde{a}\, \sigma(\tilde{a}) \underbrace{\left[\int d\vec{w}\, \overbrace{q(\vec{w}|D)}^{\text{gaussian}}\, \delta\left(\tilde{a} - a(\vec{x}, \vec{w}_{MAP}) - \vec{b}^T(\vec{x}) \cdot (\vec{w} - \vec{w}_{MAP})\right) \right]}_{= p(\tilde{a}|\vec{x}, D)}$$

, can be computed by expanding the δ-f'n in plane waves:

$$p(\tilde{a}|\vec{x}, D) = \mathcal{N}(\tilde{a} | a(\vec{x}, \vec{w}_{MAP}), \underbrace{\vec{b}^T A^{-1} \vec{b}}_{= \sigma^2_{\tilde{a}}(\vec{x})})$$

Finally,

$$p(t=1|\vec{x}, D) = \int d\tilde{a}\, \sigma(\tilde{a}) \mathcal{N}(\tilde{a} | \overbrace{a(\vec{x}, \vec{w}_{MAP})}^{\mu}, \overbrace{\sigma^2_{\tilde{a}}(\vec{x})}^{\sigma^2}) \textcircled{\approx}$$

and approximate $\sigma(\tilde{a})$ with $\Phi(\lambda\tilde{a})$, $\lambda = \sqrt{\frac{\pi}{8}}$ and use (4.153) to get:

$$\textcircled{\approx} \quad \sigma(\kappa(\sigma^2_{\tilde{a}})\, a(\vec{x}, \vec{w}_{MAP})), \text{ where}$$

$$\kappa(\sigma^2_{\tilde{a}}) = \frac{1}{\sqrt{1 + \frac{\pi \sigma^2_{\tilde{a}}}{8}}}$$

↗ (4.154)

double--check: (5.190) wrong?