Computing logistic regression predictions

In the previous note we approximated the logistic regression posterior with a Gaussian distribution. By comparing to the joint probability, we immediately obtained an approximation for the marginal likelihood $P(\mathcal{D})$ or $P(\mathcal{D}\,|\,\mathcal{M})$, which can be used to choose between alternative model settings $\mathcal{M}$.

Now we return to the question of how to make Bayesian predictions (all implicitly conditioned on a set of model choices $\mathcal{M}$):

$$\begin{aligned}P(y\,|\,\mathbf{x},\mathcal{D}) &= \int p(y,\mathbf{w}\,|\,\mathbf{x},\mathcal{D})\;\mathrm{d}\mathbf{w}\\ &= \int P(y\,|\,\mathbf{x},\mathbf{w})\,p(\mathbf{w}\,|\,\mathcal{D})\;\mathrm{d}\mathbf{w}.\end{aligned}$$

If we approximate the posterior with a Gaussian, $p(\mathbf{w}\,|\,\mathcal{D}) \approx \mathcal{N}(\mathbf{w};\,\mathbf{m},V)$, we still have an integral with no closed form solution:

$$\begin{aligned}P(y\!=\!1\,|\,\mathbf{x},\mathcal{D}) &\approx \int \sigma(\mathbf{w}^\top\mathbf{x})\,\mathcal{N}(\mathbf{w};\,\mathbf{m},V)\;\mathrm{d}\mathbf{w}\\ &= \mathbb{E}_{\mathcal{N}(\mathbf{w};\,\mathbf{m},V)}\left[\sigma(\mathbf{w}^\top\mathbf{x})\right].\end{aligned}$$

However, this expectation can be simplified. Only the inner product $a\!=\!\mathbf{w}^\top\mathbf{x}$ matters, so we can take the average over this scalar quantity instead. The linear combination a is a linear combination of Gaussian beliefs, so our beliefs about it are also Gaussian. By now you should be able to show that

$$p(a) = \mathcal{N}(a;\;\mathbf{m}^\top\mathbf{x},\;\mathbf{x}^\top V\mathbf{x}).$$

Therefore, the predictions given the approximate posterior, are given by a one-dimensional integral:

$$\begin{aligned}P(y\!=\!1\,|\,\mathbf{x},\mathcal{D}) &\approx \mathbb{E}_{\mathcal{N}(a;\;\mathbf{m}^\top\mathbf{x},\;\mathbf{x}^\top V\mathbf{x})}\left[\sigma(a)\right]\\ &= \int \sigma(a)\,\mathcal{N}(a;\;\mathbf{m}^\top\mathbf{x},\;\mathbf{x}^\top V\mathbf{x})\;\mathrm{d}a.\end{aligned}$$

One-dimensional integrals can computed numerically to high precision.

Murphy Section 8.4.4.2 reviews a further approximation (derivation non-examinable), which results in a closed form expression:

$$P(y\!=\!1\,|\,\mathbf{x},\mathcal{D}) \approx \sigma(\kappa\,\mathbf{m}^\top\mathbf{x}), \qquad \kappa = \frac{1}{\sqrt{1+\frac{\pi}{8}\mathbf{x}^\top V\mathbf{x}}}.$$

These predictions use the mean weights $\mathbf{m}$ under the Gaussian approximation. If we used the Laplace approximation, we're using the most probable or MAP weights. However, the activation is scaled down (with κ) when the activation is uncertain, so that predictions will be less confident far from the data (as they should be).