Expectation Propagation for a Probit Regression Model

Ondřej Dušek

May 28, 2013

1 The probit model

Suppose we have independent data points $\mathbf{x}_{i=1}^{n}$, each consisting of d features, thus building a matrix $X \in \mathbb{R}^{n \times d}$. Each data point $\mathbf{x}_i$ has a label $y_i \in \{-1, 1\}, i = 1 \dots n$, which gives a vector of labels $\mathbf{y}$.

We want to model this data using a *probit model*: $P(y_i|\mathbf{x}_i, \mathbf{w}) = \Phi(y_i \cdot \mathbf{w}^T \mathbf{x}_i)$. Φ denotes a standard Gaussian cumulative distribution function, i. e. $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \, \mathrm{d}t$.

In order to obtain a posterior estimate of the unknown parameters $\mathbf{w}$ ("weights vector") given our data $\{X, \mathbf{y}\}$, we use the standard Bayesian estimation scheme:

$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalization}} \tag{1}$$

Since we can choose our own prior on $\mathbf{w}$, we take the path of least resistance. We select a Gaussian prior on $\mathbf{w}$ with a zero mean and known variance v_0 and assume independence of w_j in the individual dimensions:

$$P(\mathbf{w}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{w}|\mathbf{0}, I \cdot v_0) \stackrel{\text{indep}}{=} \prod_{j=1}^{d} \mathcal{N}(w_j|0, v_0) \tag{2}$$

The form of likelihood is given by the probit model (remember that the data points are assumed to be independent, identically distributed):

$$P(\mathbf{y}|X, \mathbf{w}) \stackrel{\text{iid}}{=} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{n} \Phi(y_i \cdot \mathbf{w}^T \mathbf{x}_i) \tag{3}$$

The posterior then has the following form (where Z is a normalization constant):

$$P(\mathbf{w}|X, \mathbf{y}) = \frac{1}{Z} \cdot P(\mathbf{w}) \cdot \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) \tag{4}$$

Computing the posterior in this form is not tractable due to the product of probits in the likelihood. Therefore, we must approximate the posterior by a simpler distribution. We can use the Expectation Propagation (EP) algorithm to do that.

2 Expectation propagation algorithm

The EP algorithm assumes that we are given a joint distribution $P(X, \mathbf{y}, \mathbf{w})$ over observed data and unknown parameters, i.e. likelihood $\cdot$ prior, in the form of a product of factors:

$$P(X, \mathbf{y}, \mathbf{w}) = \prod_{i} f_i(\mathbf{w}) \tag{5}$$

The EP then tries to find an approximation $q(\mathbf{w})$ of the true posterior distribution $p(\mathbf{w}) \stackrel{\text{def}}{=} P(\mathbf{w}|X, \mathbf{y})$ by minimizing the Kullback-Leibler (KL) Divergence:

$$KL(p(\mathbf{w})||q(\mathbf{w})) = \int_{-\infty}^{\infty} p(\mathbf{w}) \log \left(\frac{p(\mathbf{w})}{q(\mathbf{w})} \right) \mathrm{d}\mathbf{w} \tag{6}$$

It does so by gradually refining one of the factors $\hat{f}_i(\mathbf{w}), i = 1 \dots n$ while keeping rest of $q(\mathbf{w})$ fixed. This is repeated until convergence (i.e. until the refined factors are undistinguishable from the original factors) and requires several passes over all factors in general.

The general flow of the algorithm looks like this:

1. Select a form of a distribution from the *exponential family* for your approximate posterior $q(\mathbf{w})$. It must be possible to express it as a product of approximate factors $\hat{f}_i(\mathbf{w})$, each approximating a factor $f_i(\mathbf{w})$ of the true posterior. We use the exponential family since it works nicely with KL divergence minimization (see below).
2. Initialize the approximate factors to some (arbitrary, but reasonable) values. You now have the first approximation of the posterior.
3. In several passes, select one factor $\hat{f}_i(\mathbf{w})$ to refine; keep the rest of the factors intact:
   (a) Take factor $\hat{f}_i(\mathbf{w})$ out of the current posterior approximation $q(\mathbf{w})$ to create a *cavity distribution* $q^{\setminus i}(\mathbf{w})$.
   (b) Now create a new approximation $\hat{f}_i^{\text{new}}(\mathbf{w})$ of the true factor $f_i(\mathbf{w})$ by minimizing:
   $$KL(f_i(\mathbf{w}) \cdot q^{\setminus i}(\mathbf{w})||\hat{f}_i(\mathbf{w}) \cdot q^{\setminus i}(\mathbf{w})) \tag{7}$$
   Note that we are not minimizing the distance to the true posterior, but to a distribution composed of the exact factor $f_i(\mathbf{w})$ and approximations of the rest, i.e. we are approaching the true factor in the context of our current approximation.
   The exponential family is very convenient here since we may use *moment matching*: we just compute the *sufficient statistics*[1] of the target distribution $f_i(\mathbf{w}) \cdot q^{\setminus i}(\mathbf{w})$ and use them for our approximation $\hat{f}_i(\mathbf{w}) \cdot q^{\setminus i}(\mathbf{w})$. If an approximation from the exponential family has the same sufficient statistics as the target distribution, it must have the lowest KL divergence.
   (c) Now replace your old posterior approximation with $q^{\text{new}}(\mathbf{w}) \propto \hat{f}_i^{\text{new}}(\mathbf{w}) \cdot q^{\setminus i}(\mathbf{w})$.
4. Repeat previous step until convergence.

3 Form of the approximation in the probit model

We now return to our probit model. We denote our posterior distribution on weights (4) as $p(\mathbf{w})$ and its individual factors as $f_i(\mathbf{w}), i = 0 \dots n$ (i.e. some functions of $\mathbf{w}$):

$$f_0(\mathbf{w}) \stackrel{\text{def}}{=} P(\mathbf{w}) \tag{8}$$

$$f_i(\mathbf{w}) \stackrel{\text{def}}{=} P(y_i|\mathbf{x}_i, \mathbf{w}) \quad i = 1 \dots n \tag{9}$$

Note that f_0 corresponds to the prior and $f_i, i = 1 \dots n$ correspond to the individual data points.

We now try to find an approximation $q(\mathbf{w})$ of $p(\mathbf{w})$. We choose the shape of $q(\mathbf{w})$ ourselves, the only requirement is that it has to be in the exponential family (see Section 2). A Gaussian with independent dimensions is the best way to keep things simple:

$$q(\mathbf{w}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{w}|\mathbf{m}, I \cdot \mathbf{v}) \stackrel{\text{indep}}{=} \prod_{j=1}^{d} \mathcal{N}(w_j|m_j, v_j) \tag{10}$$

[1] Sufficient statistics is a set of moments that uniquely define a distribution from the exponential family. For Gaussians, it is mean and variance.

Note that $\mathbf{m} = \{m_j\}_{j=1}^d$ and $\mathbf{v} = \{v_j\}_{j=1}^d$ denote means and variances in the individual dimensions.

We also want our $q(\mathbf{w})$ to be a product of factors of a similar form to (8,9), but simpler. We thus denote:

$$q(w) \stackrel{\text{def}}{=} \frac{1}{Z} \prod_{i=0}^{n} \hat{f}_i(w) \tag{11}$$

Where:

$$\hat{f}_0(\mathbf{w}) \stackrel{\text{def}}{=} f_0(\mathbf{w}) = P(\mathbf{w}) \tag{12}$$

$$\hat{f}_i(\mathbf{w}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{w}|\mathbf{m}_i, I \cdot \mathbf{v}_i) \cdot \mathbf{s}_i = \prod_{j=1}^{d} \mathcal{N}(\mathbf{w}_j|\mathbf{m}_{ij}, I \cdot \mathbf{v}_{ij}) \cdot s_{ij} \quad i = 1 \dots n \tag{13}$$

I.e. we use the exact prior (since it is a plain Gaussian) and choose the other factors $\hat{f}_i(\mathbf{w}), i = 1 \dots n$ as unnormalized Gaussians with independent dimensions. We know that the original factors $f_i(\mathbf{w}), i = 1 \dots n$ are not normalized with respect to $\mathbf{w}$ (since they are normalized with respect to y_i), but want them to have a simple form. We therefore use a Gaussian multiplied by a "de-normalization constant" s_{ij}.

We now aim to find $q(\mathbf{w})$ with such parameters $\mathbf{m}, \mathbf{v}$ that it is as close to $p(\mathbf{w})$ as possible. This is the task of the EP algorithm.

4 EP initialization step

We initialize our approximation $q(\mathbf{w})$ by setting $\hat{f}_0(\mathbf{w})$ to the prior and $\hat{f}_i(\mathbf{w})$ to uniform distributions.[2] The parameters of the approximate factors then look as follows:

$$m_{0j} := 0, \quad v_{0j} := v_0 \qquad j = 1 \dots d \tag{14}$$

$$m_{ij} := 0, \quad v_{ij} := \infty \qquad j = 1 \dots d, \quad i = 1 \dots n \tag{15}$$

Now our posterior approximation is in fact equal to our prior (if we view it as prior $\cdot \prod_{i=1}^n$ uniform).

5 Refining one factor

We select an approximate factor $\hat{f}_i(\mathbf{w})$ to be refined. The order of factors selected for refining is arbitrary and all factors should be refined multiple times.

5.1 Computing the cavity distribution

First, we compute the *cavity distribution* from our current posterior approximation $q(\mathbf{w})$ and the current approximate factor $\hat{f}_i(\mathbf{w})$:

$$q^{\setminus i}(\mathbf{w}) = \frac{q(\mathbf{w})}{\hat{f}_i(\mathbf{w})} \tag{16}$$

Since $q(\mathbf{w})$ and $\hat{f}_i(\mathbf{w})$ are both Gaussian from (10, 13), we can use the formulas for Gaussian identities to obtain an (unnormalized) Gaussian shape of $q^{\setminus i}(\mathbf{w})$:

$$q^{\setminus i}(\mathbf{w}) \propto \mathcal{N}(\mathbf{w}|\mathbf{m}^{\setminus i}, \mathbf{v}^{\setminus i}) \stackrel{\text{indep}}{=} \prod_{j=1}^{d} \mathcal{N}(w_j|m_j^{\setminus i}, v_j^{\setminus i}) \tag{17}$$

Where:

$$v_j^{\setminus i} = (v_j^{-1} - v_{ij}^{-1})^{-1} \tag{18}$$

$$m_j^{\setminus i} = v_j^{\setminus i}(v_j^{-1} m_j - v_{ij}^{-1} m_{ij}) \tag{19}$$

Note that m_{ij}, v_{ij} refer to the current approximation of $\hat{f}_i(\mathbf{w})$ and m_j, v_j refer to the current approximation of $q(\mathbf{w})$.

[2] Or as close to uniform distributions as we can get in practice since $\hat{f}_i(\mathbf{w})$ are assumed to be Gaussian.

5.2 Minimizing KL-divergence

Having fixed our cavity distribution, we want to minimize the KL divergence of our factor approximation in the context of the cavity distribution (7) to obtain a new, better approximation of the posterior, q^{new}. We have:

$$q^{\text{new}}(\mathbf{w}) = \arg \min_{q' \propto \hat{f}_i(w) q^{\setminus i}(\mathbf{w})} KL\left(\frac{1}{Z_i} f_i(\mathbf{w}) q^{\setminus i}(\mathbf{w}) \;||\; q'\right) \tag{20}$$

The KL-divergence for distributions in the exponential family is minimized by moment matching: setting the *sufficient statistics*, i.e. mean and variance in our case, equal to those of the distribution we want to approximate.

To do this, we will use the following clever formulas (from José's slide 15) for the moments of a Gaussian multiplied by some arbitrary factor. Given a distribution $r(\mathbf{x})$ in the following form:

$$r(\mathbf{x}) = \frac{1}{Z} t(\mathbf{x}) \mathcal{N}(\mathbf{x}|\mu, \Sigma) \text{ and } Z = \int t(\mathbf{x}) \mathcal{N}(x|\mu, \Sigma)\, \mathrm{d}\mathbf{x} \tag{21}$$

We can express its mean and variance as:

$$\mathbb{E}_r[\mathbf{x}] = \mu + \Sigma \cdot \frac{\partial \log Z}{\partial \mu} \tag{22}$$

$$\mathbb{E}_r[\mathbf{x}\mathbf{x}^T] - E_r[\mathbf{x}](E_r[\mathbf{x}])^T = \Sigma - \Sigma \cdot \left(\frac{\partial \log Z}{\partial \mu} \left(\frac{\partial \log Z}{\partial \mu} \right)^T - 2 \frac{\partial \log Z}{\partial \Sigma} \right) \cdot \Sigma \tag{23}$$

As the clever formulas are not clever enough to rid us of the normalizing constant, we must first compute Z_i:

$$Z_i = \int f_i(\mathbf{w}) q^{\setminus i}(\mathbf{w})\, \mathrm{d}\mathbf{w} = \int P(y_i|\mathbf{x}_i, \mathbf{w}) q^{\setminus i}(\mathbf{w})\, \mathrm{d}\mathbf{w} = \int \Phi(y_i \cdot \mathbf{w}^T \mathbf{x}_i) \prod_{i=1}^{d} \mathcal{N}(w_j | m_j^{\setminus i}, v_j^{\setminus i})\, \mathrm{d}\mathbf{w} \tag{24}$$

$$= \Phi\left(\frac{y_i \cdot \sum_{j=1}^{d} m_j^{\setminus i} x_{ij}}{\sqrt{\sum_{j=1}^{d} v_j^{\setminus i} x_{ij}^2 + 1}} \right) \tag{25}$$

If you know how we got the exact result, let me know. I don't. José just said it's relatively simple.

Now we can just fill in our values into (22, 23), using the value of Z_i computed in (25). We obtain the mean and the variance of the new approximate posterior $q^{\text{new}}(\mathbf{w})$:

$$m_j^{\text{new}} = m_j^{\setminus i} + v_j^{\setminus i} \cdot \frac{\partial \log Z_i}{\partial m_j^{\setminus i}} \tag{26}$$

$$v_j^{\text{new}} = v_j^{\setminus i} - \left(v_j^{\setminus i}\right)^2 \left(\left(\frac{\partial \log Z_i}{\partial m_j^{\setminus i}} \right)^2 - 2 \frac{\partial \log Z_i}{\partial v_j^{\setminus i}} \right) \tag{27}$$

5.3 Obtaining the new approximate factor

We now have the new approximate posterior $q^{\text{new}}(\mathbf{w})$ and need to obtain our new approximate factor $\hat{f}_i^{\text{new}}(\mathbf{w})$ for later use. We use an equation obtained from (20) by forcing Z_i as our new normalization constant:

$$q^{\text{new}}(\mathbf{w}) := \frac{1}{Z_i} \hat{f}_i(\mathbf{w}) q^{\setminus i}(\mathbf{w}) \tag{28}$$

$$\hat{f}_i(\mathbf{w}) = Z_i \frac{q^{\text{new}}(\mathbf{w})}{q^{\setminus i}(\mathbf{w})} \tag{29}$$

The parameters $m_{ij}^{\text{new}}, v_{ij}^{\text{new}}, s_{ij}^{\text{new}}$ are obtained from the Gaussian identities formulas:

$$v_{ij}^{\text{new}} = \left(v_j^{-1} - \left(v_j^{\backslash i} \right)^{-1} \right)^{-1} \tag{30}$$

$$m_{ij}^{\text{new}} = v_{ij}^{\text{new}} \cdot \left(m_j v_j^{-1} - m_j^{\backslash i} \left(v_j^{\backslash i} \right)^{-1} \right) \tag{31}$$

$$s_{ij}^{\text{new}} = Z_i \cdot C_j, \text{ where} \tag{32}$$

$$C_j = \sqrt{\frac{v_{ij}^{\text{new}} v_j^{\backslash i}}{(2\pi)^d v_j}} \exp\left(-\frac{1}{2} \left(m_j^2 v_j^{-1} - \left(m_j^{\backslash i} \right)^2 \left(v_j^{\backslash i} \right)^{-1} - \left(m_{ij}^{\text{new}} \right)^2 \left(v_{ij}^{\text{new}} \right)^{-1} \right) \right) \tag{33}$$

We can now use $\hat{f}_i^{\text{new}}(\mathbf{w})$ and $q^{\text{new}}(\mathbf{w})$ in the next iterations.