# Bayesian Linear Classifier

James Barrett
Institute of Mathematical and Molecular Biomedicine,
King's College London

November 24, 2013

**Abstract**

This document contains the mathematical theory behind the Bayesian Linear Classifier Matlab function. The likelihood function and it's partial derivatives are derived. Some numerical stability issues are also discussed.

## Model Definition

We observe covariates $\mathbf{x}_i \in \mathbb{R}^d$ (also called input variables) for each sample $i$ where $i = 1, \ldots, N$. We also observe a binary label $\sigma_i \in \{-1, +1\}$. We specify that the class membership probabilities are given by

$$p(\sigma_i = -1|\mathbf{x}_i, \mathbf{w}) = \frac{1}{2}\text{erfc}(\mathbf{w}\cdot\mathbf{x}_i + w_0) \quad \text{and} \quad p(\sigma_i = +1|\mathbf{x}_i, \mathbf{w}) = 1 - \frac{1}{2}\text{erfc}(\mathbf{w}\cdot\mathbf{x}_i + w_0) \tag{1}$$

respectively, where $\mathbf{w} \in \mathbb{R}^d$ is a vector of regression coefficients and $w_0$ is a bias term. We can include the bias term in the weight vector so that $\mathbf{w} = (w_0, w_1, \ldots, w_d) \in \mathbb{R}^{d+1}$. The complementary error function is defined as

$$\text{erfc}(h) = \frac{2}{\sqrt{\pi}}\int_h^\infty e^{-s^2}\text{d}s.$$

We will use Bayes' theorem to infer the most probable weights

$$p(\mathbf{w}|D, \alpha, \beta) = \frac{p(D|\mathbf{w})p(\mathbf{w}|\alpha, \beta)}{\int p(D|\mathbf{w}')p(\mathbf{w}'|\alpha, \beta)\text{d}\mathbf{w}'}, \tag{2}$$

where the data $D = \{(\sigma_1, \mathbf{x}_1), \ldots, (\sigma_N, \mathbf{x}_n)\}$. The evidence term factorises over samples

$$p(D|\mathbf{w}) = \prod_{i=1}^N p(\sigma_i|\mathbf{x}_i, \mathbf{w}). \tag{3}$$

We will assume a Gaussian prior for the weights

$$p(\mathbf{w}|\alpha) = \frac{e^{-\frac{1}{2\alpha^2}\mathbf{w}^2}}{(2\pi)^{N/2}\alpha^N}, \qquad p(w_0|\beta) = \frac{e^{-\frac{1}{2\beta^2}w_0^2}}{(2\pi)^{1/2}\beta}. \tag{4}$$

The hyperparameters $\alpha$ and $\beta$ control the variance of the weight components.

## Inferring the weights

We will numerically minimise the negative log of the posterior (2) using a gradient based optimisation algorithm:

$$\begin{aligned}\mathcal{L}_w(\mathbf{w}) &= -\frac{1}{N}\log p(\mathbf{w}|D,\alpha,\beta)\\ &= -\frac{1}{N}\sum_{\sigma_i=-1}\log\left\{\frac{1}{2}\mathrm{erfc}(\mathbf{w}\cdot\mathbf{x}_i+w_0)\right\} - \frac{1}{N}\sum_{\sigma_i=+1}\log\left\{1-\frac{1}{2}\mathrm{erfc}(\mathbf{w}\cdot\mathbf{x}_i+w_0)\right\}\\ &\qquad + \frac{1}{2N\alpha^2}\mathbf{w}^2 + \frac{1}{2}\log 2\pi + \log\alpha + \frac{1}{2N\beta^2}w_0^2 + \frac{1}{2N}\log 2\pi + \frac{1}{N}\log\beta. \qquad (5)\end{aligned}$$

The hyperparameters are fixed in the current implementation.

## Error bars

We will approximate the posterior distribution over $\mathbf{w}$ with a Gaussian centered on $\mathbf{w}^* = \min_{\mathbf{w}} \mathcal{L}_w(\mathbf{w})$. The curvature matrix can then be used to estimate the variance of the weights. Specifically,

$$p(\mathbf{w}|D,\alpha,\beta) = e^{-N\mathcal{L}_w(\mathbf{w}^*)-\frac{N}{2}(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{A}(\mathbf{w}-\mathbf{w}^*)}$$

where $\mathbf{A}$ is the matrix of second order partial derivatives of $\mathcal{L}_w(\mathbf{w})$ given below. We can interpret $(N\mathbf{A})^{-1}$ as a covariance matrix. Inferred regression weights are then given by

$$w_\mu^* \pm \sqrt{(N\mathbf{A})^{-1}_{\mu\mu}} \quad \text{for } \mu = 0,\ldots,d$$

## Partial Derivatives

Define $h_i = \mathbf{w}\cdot\mathbf{x}_i + w_0$. For the $-1$ class the first order derivatives for $\mu = 1,\ldots,d$ are

$$\frac{\partial}{\partial w_\mu}\log(\tfrac{1}{2}\mathrm{erfc}(h)) = -\frac{e^{-h^2}}{\frac{1}{2}\mathrm{erfc}(h)}x_{i\mu}.$$

We note that due to symmetry

$$\frac{\mathrm{d}}{\mathrm{d}h}\log\left\{1-\frac{1}{2}\mathrm{erfc}(h)\right\} = -\frac{\mathrm{d}}{\mathrm{d}h}\log\left\{\frac{1}{2}\mathrm{erfc}(-h)\right\}.$$

Combining these results we may write

$$\frac{\partial}{\partial w_\mu}\mathcal{L}_w(\mathbf{w}) = \frac{1}{N}\sum_{\sigma_i=-1}\frac{2}{\sqrt{\pi}}\frac{e^{-h_i^2}}{\mathrm{erfc}(h_i)}x_{i\mu} - \frac{1}{N}\sum_{\sigma_i=+1}\frac{2}{\sqrt{\pi}}\frac{e^{-h_i^2}}{\mathrm{erfc}(-h_i)}x_{i\mu} + \frac{w_\mu}{\alpha^2 N}. \qquad (6)$$

The derivative of the bias terms is

$$\frac{\partial}{\partial w_0}\mathcal{L}_w(\mathbf{w}) = \frac{1}{N}\sum_{\sigma_i=-1}\frac{2}{\sqrt{\pi}}\frac{e^{-h_i^2}}{\mathrm{erfc}(h_i)} - \frac{1}{N}\sum_{\sigma_i=+1}\frac{2}{\sqrt{\pi}}\frac{e^{-h_i^2}}{\mathrm{erfc}(-h_i)} + \frac{w_0}{\beta^2 N}$$

Second order partial derivatives for the $-1$ class the first order derivatives for $\mu = 1, \ldots, d$ are

$$\frac{\partial^2}{\partial w_\nu \partial w_\mu} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} x_{i\mu} = \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right)^2 - 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right) x_{i\nu} x_{i\mu}.$$

We observe that

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2} \log \left\{ 1 - \frac{1}{2} \operatorname{erfc}(h) \right\} = \frac{\mathrm{d}^2}{\mathrm{d}h^2} \log \left\{ \frac{1}{2} \operatorname{erfc}(-h) \right\}.$$

Second order partials are given by

$$\begin{aligned} \frac{\partial^2}{\partial w_\nu \partial w_\mu} \mathcal{L}_w(\mathbf{w}) &= \frac{1}{N} \sum_{\sigma_i = -1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right)^2 - 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right) \right] x_{i\nu} x_{i\mu} \\ &\quad + \frac{1}{N} \sum_{\sigma_i = +1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right)^2 + 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right) \right] x_{i\nu} x_{i\mu} + \frac{\delta_{\mu\nu}}{\alpha^2 N} \\ &= \mathbf{A}_{\mu\nu}. \end{aligned}$$

Furthermore

$$\begin{aligned} \frac{\partial^2}{\partial w_0^2} \mathcal{L}_w(\mathbf{w}) &= \frac{1}{N} \sum_{\sigma_i = -1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right)^2 - 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right) \right] \\ &\quad + \frac{1}{N} \sum_{\sigma_i = +1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right)^2 + 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right) \right] + \frac{1}{\beta^2 N} \end{aligned}$$

and finally,

$$\begin{aligned} \frac{\partial^2}{\partial w_\mu \partial w_0} \mathcal{L}_w(\mathbf{w}) &= \frac{1}{N} \sum_{\sigma_i = -1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right)^2 - 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(h_i)} \right) \right] x_{i\mu} \\ &\quad + \frac{1}{N} \sum_{\sigma_i = +1} \left[ \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right)^2 + 2h_i \left( \frac{2}{\sqrt{\pi}} \frac{e^{-h_i^2}}{\operatorname{erfc}(-h_i)} \right) \right] x_{i\mu}. \end{aligned}$$

## Numerical Instability

If $h \gg 0$ then $e^{-h^2} \approx 0$ and $\operatorname{erfc}(h) \approx 0$ and the numerical value of their ratio in the derivatives above will become inaccurate as the limit of machine precision is reached. In this regime we will use the asymptotic expansion (Menzel, 1960) of $\operatorname{erfc}(h)$ which is given by

$$\operatorname{erfc}(h) = \frac{e^{-h^2}}{h\sqrt{\pi}} \left[ 1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots \right] \qquad \text{for } h \gg 0.$$

This means that for the $\sigma_i = -1$ class when $h \gg 0$ (a cutoff of $h \geq 20$ is suitable for matlab)

$$\log\left(\tfrac{1}{2}\mathrm{erfc}(h)\right) = -h^2 - \log h - \log 2\sqrt{\pi} + \log\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right]$$
$$\frac{e^{-h^2}}{\mathrm{erfc}(h)} = h\sqrt{\pi}\left[1 - \frac{1}{2h^2} + \frac{2}{(2h^2)^2} - \frac{8}{(2h^2)^3} + \cdots\right]^{-1}.$$

For the $\sigma_i = +1$ class when $h \leq -20$ define $g = -h$ and use

$$\log\left(1 - \tfrac{1}{2}\mathrm{erfc}(h)\right) = \log\left(\tfrac{1}{2}\mathrm{erfc}(g)\right) = -g^2 - \log g - \log 2\sqrt{\pi} + \log\left[1 - \frac{1}{2g^2} + \frac{2}{(2g^2)^2} - \cdots\right]$$
$$\frac{e^{-g^2}}{\mathrm{erfc}(g)} = g\sqrt{\pi}\left[1 - \frac{1}{2g^2} + \frac{2}{(2g^2)^2} - \frac{8}{(2g^2)^3} + \cdots\right]^{-1}.$$

# References

Donald H. Menzel. *Fundamental Formulas of Physics*, volume one. Dover Publications, Inc. New York, 1960.