Logistic regression

Sep 2022

Murphy chap8, Bishop chap4

Maximum entropy property

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} P(y_i = k|x_i)\log P(y_i = k|x_i).$$

The fixed-mean constraints,

$$\sum_{i=1}^{N} P(y_i = k|x_i)x_{ij} = \sum_{i=1}^{N} \mu_k(y_i)x_{ij},\ \forall k = 1, ..., K,\ j = 1, ..., p.$$

The normalization contraint,

$$\sum_{k=1}^{K} P(y_i = k|x_i) = 1,\ i = 1, ..., N.$$

Maximum entropy property

6.3. **Binary logistic regression.** For binary classification, we encode the class labels using $\{0, 1\}$. Note $P(y = 1|x) = 1 - P(y = 0|x)$. We may write

$$\ln P(y = 0|x) = \beta_0^T x - \ln Z,$$

$$\ln(1 - P(y = 0|x)) = \beta_1^T x - \ln Z.$$

Since we have incorporated the normalization condition, the coefficients β_0^T, β_1^T must be compatible in some sense. One way to include the compatibility is to write the model as (by substracting the second equation from the first),

$$\ln \frac{P(y = 0|x)}{1 - P(y = 0|x)} = \beta^T x, \ \beta \in \mathbb{R}^p.$$

The LHS as a function of $P(y = 0|x)$ is called the logit or log odds function

$$\text{logit}(p) = \ln \frac{p}{1 - p},$$

which is the inverse of the logistic function

$$\sigma(t) = \frac{1}{1 + e^{-t}}.$$

Therefore

$$P(y = 0|x) = \sigma\left(\beta^T x\right) = \frac{1}{1 + e^{-\beta^T x}}.$$

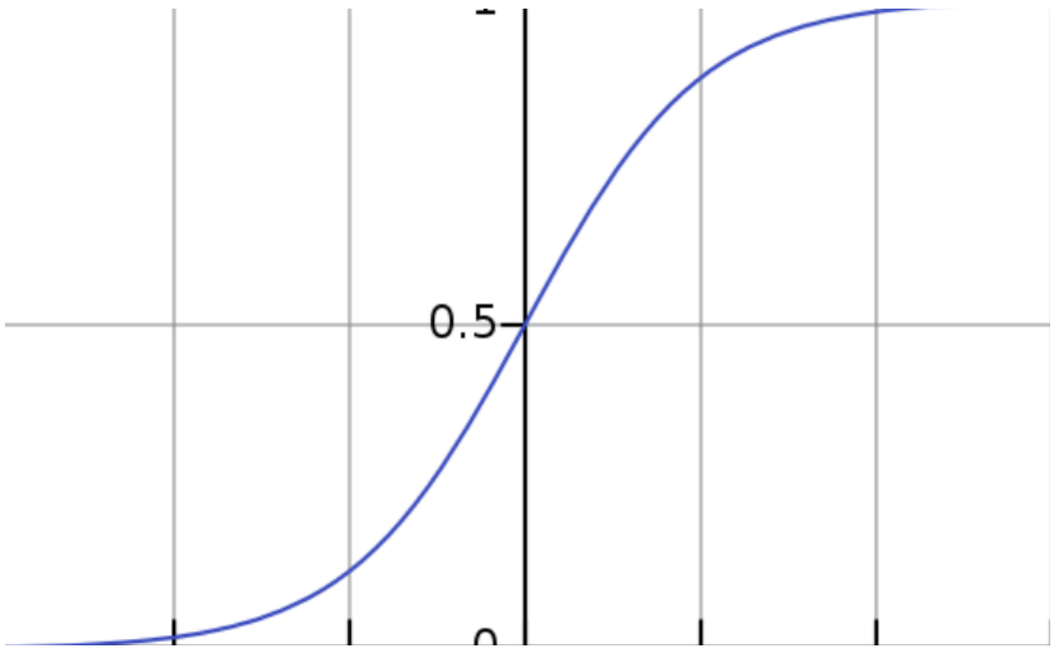
Logistic function

For later reference: the logistic function is

$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{1}{2} + \frac{1}{2}\tanh(\frac{t}{2})$$

$$\tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} = \frac{e^{2t}-1}{e^{2t}+1}$$

$$\frac{\partial \sigma}{\partial t} = \sigma(t)(1-\sigma(t))$$



Model selection

Entropy, AIC, BIC?

AIC:

Historically various 'information criteria' have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term to compensate for the over-fitting of more complex models. For example, the *Akaike information criterion*, or AIC (Akaike, 1974), chooses the model for which the quantity

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.73}$$

is largest. Here $p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}})$ is the best-fit log likelihood, and M is the number of adjustable parameters in the model. A variant of this quantity, called the *Bayesian*

Laplace Approximation

Here we introduce a simple, but widely used, framework called the Laplace approximation, that aims to find a Gaussian approximation to a probability density defined over a set of continuous variables. Consider first the case of a single continuous variable z, and suppose the distribution $p(z)$ is defined by

$$p(z) = \frac{1}{Z} f(z) \tag{4.125}$$

where $Z = \int f(z)\,\mathrm{d}z$ is the normalization coefficient. We shall suppose that the value of Z is unknown. In the Laplace method the goal is to find a Gaussian approximation $q(z)$ which is centred on a mode of the distribution $p(z)$. The first step is to find a mode of $p(z)$, in other words a point z_0 such that $p'(z_0) = 0$, or equivalently

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0. \tag{4.126}$$

Laplace Approximation

z_0 so that

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2 \tag{4.127}$$

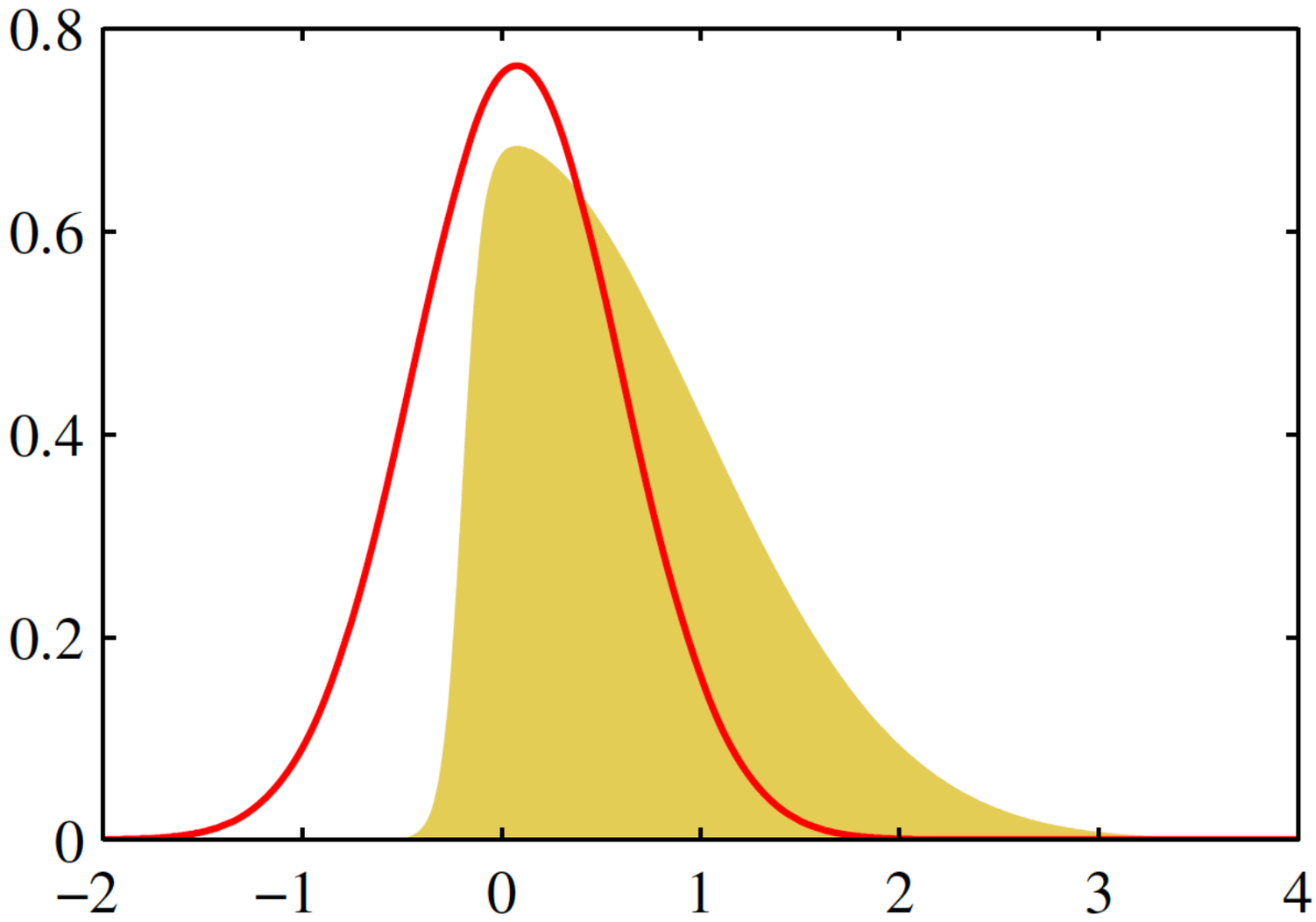
$$A = - \left. \frac{d^2}{dz^2} \ln f(z) \right|_{z=z_0} . \tag{4.128}$$

$$f(z) \simeq f(z_0) \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\} . \tag{4.129}$$

After normalization， the approximation q(x) of p(x) is:

$$q(z) = \left(\frac{A}{2\pi} \right)^{1/2} \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\} . \tag{4.130}$$

Laplace Approximation



$$p(z) \propto \exp(-z^2/2)\sigma(20z + 4)$$

where $\sigma(z)$ is the logistic sigmoid function defined by $\sigma(z) = (1 + e^{-z})^{-1}$

Laplace Approximation

We can extend the Laplace method to approximate a distribution $p(\mathbf{z}) = f(\mathbf{z})/Z$ defined over an M-dimensional space $\mathbf{z}$. At a stationary point $\mathbf{z}_0$ the gradient $\nabla f(\mathbf{z})$ will vanish. Expanding around this stationary point we have

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathrm{T}} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \tag{4.131}$$

where the $M \times M$ Hessian matrix $\mathbf{A}$ is defined by

$$\mathbf{A} = - \nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0} \tag{4.132}$$

and ∇ is the gradient operator. Taking the exponential of both sides we obtain

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0) \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathrm{T}} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\}. \tag{4.133}$$

Laplace Approximation

From (4.133):

$$
\begin{aligned}
Z &= \int f(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\simeq f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}\,\mathrm{d}\mathbf{z} \\
&= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \qquad (4.135)
\end{aligned}
$$

Use (4.135) to approximate model evidence !

BIC

Consider a data set $\mathcal{D}$ and a set of models $\{\mathcal{M}_i\}$ having parameters $\{\boldsymbol{\theta}_i\}$. For each model we define a likelihood function $p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)$. If we introduce a prior $p(\boldsymbol{\theta}_i|\mathcal{M}_i)$ over the parameters, then we are interested in computing the model evidence $p(\mathcal{D}|\mathcal{M}_i)$ for the various models. From now on we omit the conditioning on $\mathcal{M}_i$ to keep the notation uncluttered. From Bayes' theorem the model evidence is given by

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}. \tag{4.136}$$

Identifying $f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $Z = p(\mathcal{D})$, and applying the result (4.135), we obtain

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \underbrace{\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|}_{\text{Occam factor}} \tag{4.137}$$

where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is the value of $\boldsymbol{\theta}$ at the mode of the posterior distribution, and $\mathbf{A}$ is the *Hessian* matrix of second derivatives of the negative log posterior

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}). \tag{4.138}$$

BIC

If we assume that the Gaussian prior distribution over parameters is broad, and that the Hessian has full rank, then we can approximate (4.137) very roughly using

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2} M \ln N \tag{4.139}$$

where N is the number of data points, M is the number of parameters in $\boldsymbol{\theta}$ and we have omitted additive constants. This is known as the *Bayesian Information Criterion* (BIC) or the *Schwarz criterion* (Schwarz, 1978). Note that, compared to AIC given by (1.73), this penalizes model complexity more heavily.

Bayesian logistic regression

4.5. Bayesian Logistic Regression

We now turn to a Bayesian treatment of logistic regression.

begin with a Gaussian prior, which we write in the general form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{4.140}$$

where $\mathbf{m}_0$ and $\mathbf{S}_0$ are fixed hyperparameters. The posterior distribution over $\mathbf{w}$ is given by

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w}) \tag{4.141}$$

Bayesian logistic regression

where $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$. Taking the log of both sides, and substituting for the prior distribution using (4.140), and for the likelihood function using (4.89), we obtain

$$\begin{aligned} \ln p(\mathbf{w}|\mathbf{t}) &= -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ &\quad + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const} \end{aligned} \tag{4.142}$$

where $y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)$. To obtain a Gaussian approximation to the posterior distribution, we first maximize the posterior distribution to give the MAP (maximum posterior) solution $\mathbf{w}_{\mathrm{MAP}}$, which defines the mean of the Gaussian. The covariance is then given by the inverse of the matrix of second derivatives of the negative log likelihood, which takes the form

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}. \tag{4.143}$$

next page

Bayesian logistic regression

$$\frac{\partial(tlny + (1-t)ln(1-y))}{\partial w}$$
$$= t\frac{y(1-y)}{y}\phi + (1-t)\frac{(-y)(1-y)}{1-y}\phi$$
$$= t\phi - ty\phi + ty\phi - y\phi$$
$$= (t-y)\phi$$

$$\frac{\partial^2(tlny + (1-t)ln(1-y))}{\partial w^2}$$
$$= -y(1-y)\phi\phi^T$$

Bayesian logistic regression

The Gaussian approximation to the posterior distribution therefore takes the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N). \tag{4.144}$$

Bayesian logistic regression

Having obtained a Gaussian approximation to the posterior distribution, there remains the task of marginalizing with respect to this distribution in order to make predictions.

4.5.2 Predictive distribution

The predictive distribution for class $\mathcal{C}_1$, given a new feature vector $\boldsymbol{\phi}(\mathbf{x})$, is

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{t})\, \mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) q(\mathbf{w})\, \mathrm{d}\mathbf{w} \tag{4.145}$$

Denoting $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$, we have

$$\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) \sigma(a)\, \mathrm{d}a \tag{4.146}$$

where $\delta(\cdot)$ is the Dirac delta function. From this we obtain

$$\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) q(\mathbf{w})\, \mathrm{d}\mathbf{w} = \int \sigma(a) p(a)\, \mathrm{d}a \tag{4.147}$$

Bayesian logistic regression

$$\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})\sigma(a)\,\mathrm{d}a \tag{4.146}$$

where $\delta(\cdot)$ is the Dirac delta function. From this we obtain

$$\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} = \int \sigma(a)p(a)\,\mathrm{d}a \tag{4.147}$$

where

$$p(a) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{4.148}$$

p(a) still a Gaussian distribution !

Bayesian logistic regression

$$a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi},$$

$$\mu_a = \mathbb{E}[a] = \int p(a)a\,\mathrm{d}a = \int q(\mathbf{w})\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\,\mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi} \tag{4.149}$$

where we have used the result (4.144) for the variational posterior distribution $q(\mathbf{w})$. Similarly

$$\begin{aligned}\sigma_a^2 &= \mathrm{var}[a] = \int p(a)\left\{a^2 - \mathbb{E}[a]^2\right\}\,\mathrm{d}a \\ &= \int q(\mathbf{w})\left\{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})^2 - (\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi})^2\right\}\,\mathrm{d}\mathbf{w} = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}.\end{aligned} \tag{4.150}$$

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a)\,\mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,\mathrm{d}a. \tag{4.151}$$

See 8.4.4.2 for the approximation of the integral

Monte Carlo Approximation

focusses on prediction. The posterior predictive distribution has the form

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \tag{8.59}$$

Unfortunately this integral is intractable.

The simplest approximation is the plug-in approximation, which, in the binary case, takes the form

$$p(y = 1|\mathbf{x}, \mathcal{D}) \approx p(y = 1|\mathbf{x}, \mathbb{E}[\mathbf{w}]) \tag{8.60}$$

where $\mathbb{E}[\mathbf{w}]$ is the posterior mean. In this context, $\mathbb{E}[\mathbf{w}]$ is called the **Bayes point**. Of course, such a plug-in estimate underestimates the uncertainty. We discuss some better approximations below.

Monte Carlo Approximation

Monte Carlo approximation

A better approach is to use a Monte Carlo approximation, as follows:

$$p(y=1|\mathbf{x},\mathcal{D}) \quad \approx \quad \frac{1}{S}\sum_{s=1}^{S}\text{sigm}((\mathbf{w}^s)^T\mathbf{x}) \tag{8.61}$$

where $\mathbf{w}^s \sim p(\mathbf{w}|\mathcal{D})$ are samples from the posterior. (This technique can be trivially extended to the multi-class case.) If we have approximated the posterior using Monte Carlo, we can reuse these samples for prediction. If we made a Gaussian approximation to the posterior, we can draw *independent* samples from the Gaussian using standard methods.