Formulas from Logistic Regression Document

Discriminative Linear Models – Logistic Regression

$$\log \frac{P(C = h_1|x)}{P(C = h_0|x)} = \log \frac{f_{X|C}(x|h_1)}{f_{X|C}(x|h_0)} + \log \frac{\pi}{1-\pi} = w^T x + b$$

Logistic Regression Model

$$P(C = h_1|x, w, b) = e^{(w^T x + b)} P(C = h_0|x, w, b)$$

$$P(C = h_1|x, w, b) = \frac{e^{(w^T x + b)}}{1 + e^{(w^T x + b)}} = \frac{1}{1 + e^{-(w^T x + b)}} = \sigma(w^T x + b)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

Likelihood Estimation

$$P(C_1 = c_1, \ldots, C_n = c_n|x_1, \ldots, x_n, w, b) = \prod_{i=1}^{n} P(C_i = c_i|x_i, w, b)$$

$$y_i = P(C_i = 1|x_i, w, b) = \sigma(w^T x_i + b)$$

$$P(C_i = 0|x_i, w, b) = 1 - y_i = \sigma(-w^T x_i - b)$$

Log-Likelihood Function

$$L(w, b) = \prod_{i=1}^{n} y_i^{c_i} (1 - y_i)^{(1-c_i)}$$

$$\ell(w, b) = \sum_{i=1}^{n} [c_i \log y_i + (1 - c_i) \log(1 - y_i)]$$

Maximum Likelihood Estimation

$$w^*, b^* = \arg\max_{w,b} \ell(w, b)$$

Minimizing the negative log-likelihood:

$$J(w, b) = -\ell(w, b) = \sum_{i=1}^{n} -[c_i \log y_i + (1 - c_i) \log(1 - y_i)]$$

Binary Cross-Entropy

$$H(c_i, y_i) = -[c_i \log y_i + (1 - c_i) \log(1 - y_i)]$$

Logistic Loss Function

The objective function can thus be rewritten as:

$$J(w, b) = \sum_{i=1}^{n} H(c_i, y_i)$$

$$= \sum_{i=1}^{n} \log\left(1 + e^{-z_i(w^T x_i + b)}\right)$$

$$= \sum_{i=1}^{n} l(z_i(w^T x_i + b))$$

where

$$l(x) = \log(1 + e^{-x})$$

is the logistic loss function.

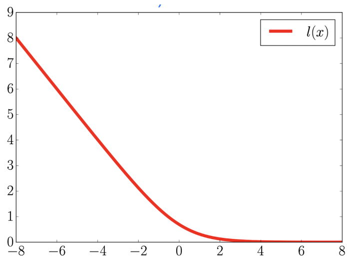

Figure 1: Plot of the logistic loss function $l(x)$

$$\log \frac{P(h_1|x_i)}{P(h_0|x_i)} = w^T x_i + b = s_i$$

$$s_i < t$$

$$s_i > t$$

Since $s_i = w^T x_i + b$, decision rules are linear hyperplanes orthogonal to the vector w. Moreover, s_i is related to the distance of the sample x_i from the separating surface. **The cost we pay for each sample is** $l(z_i s_i)$

Regularized Logistic Regression

$$R(w, b) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-z_i(w^T x_i + b)})$$

The weight vector w does not explicitly appear in the log-likelihood formulation because it is already embedded within the sigmoid function, which models the probability of the positive class as $P(y = 1|x) = \sigma(w^T x + b)$. The objective function, defined as the negative log-likelihood (binary cross-entropy), inherently depends on w through this probability. However, regularization terms, such as L1 or L2 penalties, are explicitly expressed in terms of w because they are additional constraints imposed to control model complexity and prevent overfitting, independently of the probabilistic framework.

Prior Weighted Logistic Regression

$$R(w) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{\pi_T}{n_T}\sum_{i|z_i=1} l(z_i s_i) + \frac{1-\pi_T}{n_F}\sum_{i|z_i=-1} l(z_i s_i)$$

Multiclass Logistic Regression (Softmax)

$$P(C = k|x) = \frac{e^{w_k^T x + b_k}}{\sum_{j=1}^{K} e^{w_j^T x + b_j}}$$

For a given sample:

$$y_{ik} = \frac{e^{w_k^T x_i + b_k}}{\sum_j e^{w_j^T x_i + b_j}}$$

Log-likelihood:

$$\ell(W, b) = \sum_{i=1}^{n} \log P(C_i = c_i | X_i = x_i, W, b)$$

Cross-Entropy for Multiclass:

$$H(z_i, y_i) = -\sum_{k=1}^{K} z_{ik} \log y_{ik}$$

z_{ik} is the one hot encoded representation!

The ML solution is again the solution that minimizes the (average) cross-entropy:

$$\arg\max_{W,b} \ell(W, b) = \arg\min_{W,b} \sum_{i=1}^{n} H(z_i, y_i)$$

Or, in terms of loss function (depends on xi, ci, W, b):

$$J(W, b) = -\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log y_{ik} = \sum_{i=1}^{n} l(x_i, c_i, W, b)$$

Regularized Multiclass Logistic Regression:

$$R(W, b) = \Omega(W) + \frac{1}{n} J(W, b)$$

L2 Regularization:

$$\Omega(w_1, \ldots, w_N) = \frac{1}{2} \sum_{i} \|w_i\|^2$$

Expanded feature space

Remember that, for binary Logistic Regression (LR), we assumed linear separation surfaces:

$$\log \frac{P(C = h_1 \mid \mathbf{x})}{P(C = h_0 \mid \mathbf{x})} = \mathbf{w}^T \mathbf{x} + b$$

which has the same form as the Gaussian classifier with tied covariances.
For a Gaussian classifier with non-tied covariances, we have:

$$\log \frac{P(C = h_1 \mid \mathbf{x})}{P(C = h_0 \mid \mathbf{x})} = \mathbf{x}^T A \mathbf{x} + \mathbf{b}^T \mathbf{x} + c = s(\mathbf{x}, A, \mathbf{b}, c)$$

The expression:

$$s(\mathbf{x}, A, \mathbf{b}, c) = \mathbf{x}^T A \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$$

is quadratic in $\mathbf{x}$, however, it is linear in A and $\mathbf{b}$.
If we define:

$$\phi(\mathbf{x}) = \begin{bmatrix} \mathrm{vec}(\mathbf{x}\mathbf{x}^T) \\ \mathbf{x} \end{bmatrix}$$

and

$$\mathbf{w} = \begin{bmatrix} \mathrm{vec}(A) \\ \mathbf{b} \end{bmatrix}$$

then the class log-posterior ratio can be expressed as:

$$s(\mathbf{x}, \mathbf{w}, c) = \mathbf{w}^T \phi(\mathbf{x}) + c$$