LOGISTIC REGRESSION (binary)

Assume that we have features described by a random vector $X \in \mathbb{R}^M$ and a random variable Y with range in $\{0,1\}$, $Y \sim Be(p)$.

- X represents features
- Y represents the label of classification.

We are interested in finding

$$\mathbb{P}(Y=1), \quad \mathbb{P}(Y=0)$$

The variable Y is <u>dependent</u> from the variable X.
What we want to reconstruct is thus

$$\mathbb{P}(Y=1 \mid X=x) =: p(x)$$

This means that we want to reconstruct the probability of classifying with "1" when we observe some "features" x.

<u>Remark</u>: By the law of total probability

$$\sum_x \mathbb{P}(Y=1 \mid X=x)\, \mathbb{P}(X=x) = p.$$

the one of $Be(p)$

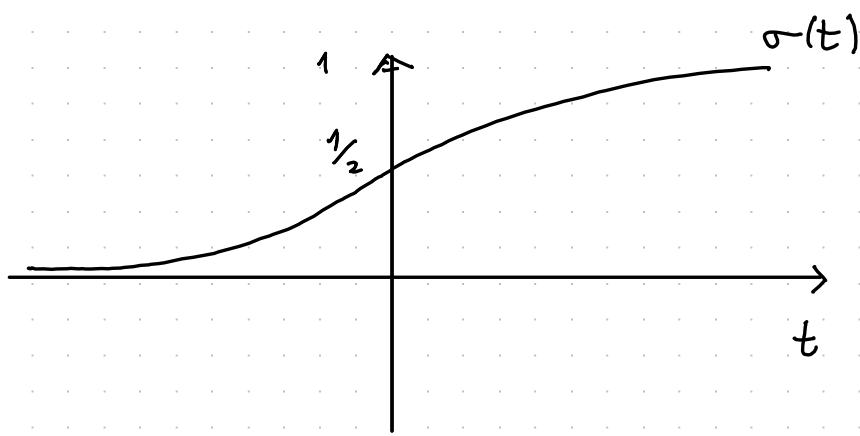
The logistic model makes the following ansatz:

$$\mathbb{P}(Y=1 \mid X=x) = p(x) \approx \sigma(xw+b) =: q(x; w, b)$$

where

- $x \in \mathbb{R}^{1\times M}$ is the observed feature
- $w \in \mathbb{R}^{M\times 1}$ are weights
- $b \in \mathbb{R}$ is a bias
- σ is the sigmoid / logistic function

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



Assume that we have observed data $\{(X_i, Y_i)\}_{i=0,\dots,N-1}$, where

- $x_0, \dots, x_{N-1} \in \mathbb{R}^{1\times M}$ are realizations of a random sample drawn from X
- $y_0, \dots, y_{N-1} \in \{0,1\}$ are the corresponding labels.

We are making an ansatz on the model:

$$\mathbb{P}(Y=1 \mid X=x) = \sigma(xw+b) =: q(x; w, b)$$

where w, b are the parameters.

We want to find the MLE of w, b and compute its realization on the observed data.

The $-\log-$likelihood is:

$$-\log\left(\mathbb{P}\left(\bigcap_{i=0}^{N-1}\{X_i=x_i, Y_i=y_i\}\right)\right)=$$

$$=-\log\left(\prod_{i=0}^{N-1}\mathbb{P}(\{X_i=x_i, Y_i=y_i\})\right)=$$

$$=-\sum_{i=0}^{N-1}\log\left(\mathbb{P}(Y_i=y_i\,|\,X_i=x_i)\,\mathbb{P}(X_i=x_i)\right)$$

$$=-\sum_{i=0}^{N-1}\log\left(\mathbb{P}(Y_i=y_i\,|\,X_i=x_i)\right)+$$

$$\underbrace{-\sum_{i=0}^{N-1}\log\mathbb{P}(X_i=x_i)}$$

This quantity is independent of the parameters W, b. It is not interesting for the optimization problem

We study only the first term.

$$-\sum_{i=0}^{N-1} \log\Big(\mathbb{P}(Y = y_i \mid X = x_i)\Big) =$$

$$= -\sum_{i=0}^{N-1} \log\left(q(x_i; w,b)^{y_i} (1-q(x_i;w,b))^{1-y_i} \right) =$$

$$= -\sum_{i=0}^{N-1} y_i \log q(x_i;w,b) + (1-y_i)\log(1-q(x_i;w,b))$$

The part of the negative log-likelihood depending on the parameters w, b is:

$$L(w,b; \{(x_i,y_i)\}_{i=0,\dots,N-1}) =$$

$$= -\sum_{i=0}^{N-1} y_i \log q(x_i;w,b) + (1-y_i)\log(1-q(x_i;w,b))$$

To minimize this, we need numerical methods (we will see them later).

CROSS-ENTROPY

Cross-entropy is a tool in information theory.
Assume that Ω is the sample space (discrete) and $\mathbb{P}$ and $\mathbb{Q}$ are two probability measures on Ω.
Let us set:

$$p(\omega) := \mathbb{P}(\{\omega\}) \quad , \quad q(\omega) := \mathbb{Q}(\{\omega\})$$

The <u>cross-entropy</u> is defined by

$$H(p,q) = \mathbb{E}_{\mathbb{P}}(-\log q) =$$

$$= -\sum_{\omega \in \Omega} p(\omega) \log q(\omega)$$

in this formula $-\log q : \Omega \to \mathbb{R}$ is treated as a r.v.

Let us understand how to interpret this.
First of all,

$$H(p) = H(p,p) = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega)$$

This is the entropy of the probability distribution $p(\omega)$.

The entropy is the expected amount of information carried by events sampled from the distribution.

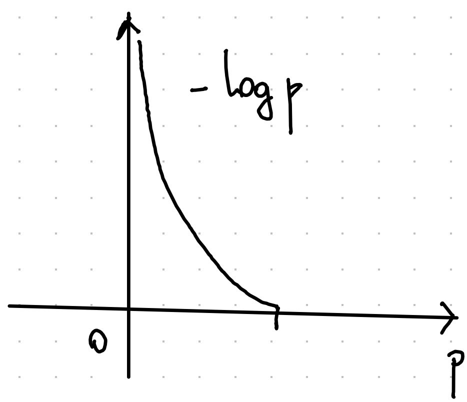
The idea is the following:

- very likely events carry a little amount of information
- very unlikely events carry a lot of information

A function that models this is

$$-\log p(\omega)$$

for a discrete event x.



In some sense, this measures how much we are "surprised" of observing the discrete event ω.

Let us go back to cross entropy:

$$H(p,q) = -\sum_{\omega\in\Omega} p(\omega)\log q(\omega) =$$

$$= -\sum_{\omega\in\Omega} p(\omega)\left(\log\frac{q(\omega)}{p(\omega)} + \log p(\omega)\right) =$$

$$= -\sum_{\omega\in\Omega} p(\omega)\log\frac{q(\omega)}{p(\omega)} + H(p)$$

$$= \underbrace{D_{KL}(p\|q)}_{} + H(p)$$

This is called <u>Kullback-Leibler divergence</u>

We use the interpretation of "surprise" to explain the meaning of the Kullback-Leibler divergence.

First of all, we assume that $\mathbb{P}$ is the probability distribution underlying the described phenomenon.

<u>Remark</u> : $D_{KL}(p \| q) \geq 0$. This is known as Gibb's inequality.
A proof is the following:
log is a concave function.
By Jensen's inequality

$$\sum_{\omega \in \Omega} p(\omega) \log \frac{q(\omega)}{p(\omega)} \leq \log \Big(\underbrace{\sum_{\omega \in \Omega} q(\omega)}_{=1} \Big) = 0 .$$

Equality holds true if and only if $\frac{q(\omega)}{p(\omega)}$ is constant.

- If $\frac{q(\omega)}{p(\omega)} \approx 1$, this means $q(\omega) \approx p(\omega)$. Using $\mathbb{Q}$ to estimate the probability of ω instead of the "true" probability $\mathbb{P}$ is a good choice. We pay a low cost.

- If $\frac{q(\omega)}{p(\omega)}$ far from 1, then we are making a mistake. With $-\log \frac{q(\omega)}{p(\omega)}$ we measure how much we are surprised of seeing the event ω drawn from $\mathbb{Q}$ relatively to how much we are surprised if the probability distribution is $\mathbb{P}$

Then

$$D_{KL}(p \| q) = \mathbb{E}[-\log q + \log p]$$

is the expected excess in surprise.

This can be thought as follows:

- there is an underlying distribution for the data: $\mathbb{P}$
- We use a model $\mathbb{Q}$
- We measure the mistake we are making by using $\mathbb{P}$ instead of $\mathbb{Q}$ by $D_{KL}(p \| q)$.

Minimizing the cross-entropy with respect to q is equivalent to minimizing the D_{KL}.

Cross-entropy for binary classification

Assume that we have

- X random vector for the "explaining" features
- Y random variable explained by X, $Y \sim Be(p)$

Here we have the following situation: The whole vector (X, Y) is distributed according to some law. If we are working in the discrete setting, there is a joint probability mass function

$$\tilde{p}(x, y) = \mathbb{P}(X = x, Y = y)$$

About the joint probability mass function, we know that:

$$\tilde{p}(x,y) = \mathbb{P}(X=x, Y=y) =$$
$$= \mathbb{P}(Y=y \mid X=x)\,\mathbb{P}(X=x)$$

$$\tilde{p}(x,1) = p(x)\,\mathbb{P}(X=x)$$
$$\tilde{p}(x,0) = (1-p(x))\,\mathbb{P}(X=x)$$

Instead of $\tilde{p}$, the logistic model uses $\tilde{q}$ for which $p(x)$ is replaced by

$$q(x;w,b) := \sigma(wx+b)$$

Assuming that the distribution of X is the same:

$$\tilde{q}(x,1;w,b) = q(x)\,\mathbb{P}(X=x)$$

$$\tilde{q}(x,0;w,b) = (1-q(x))\,\mathbb{P}(X=X)$$

The cross-entropy is

$$H(\tilde{p}, \tilde{q}) = - \mathbb{E}_{\tilde{p}}(\log \tilde{q})$$

We are thinking at $-\log(\tilde{q})$ as a random variable, in the following sense:

$$\omega \in \Omega \longmapsto -\log(\tilde{q}(X(\omega), Y(\omega)))$$

Let us write more explicitly the expression of the cross-entropy:

$$H(\tilde{p}, \tilde{q}) = - \sum_x \sum_y \tilde{p}(x,y) \log \tilde{q}(x,y; w,b) =$$

$$= \{ Y \sim Be(p) \} =$$

$$= - \sum_x \left(\tilde{p}(x,1) \log \tilde{q}(x,1; w,b) + \tilde{p}(x,0) \log \tilde{q}(x,0; w,b) \right)$$

<u>Lemma</u>: Let (X,Y) be distributed with $\tilde{p}$.

$$-\mathbb{E}_{\tilde{p}}\Big[Y\log(q(X;w,b)) + (1-Y)\log(1-q(X;w,b))\Big] =$$

$$= H(\tilde{p},\tilde{q})$$

Proof:

$$-\sum_x\sum_y \tilde{p}(x,y)\Big(y\log(q(x)) + (1-y)\log(1-q(x))\Big)$$

$$= -\sum_x \tilde{p}(x,1)\log q(x) + \tilde{p}(x,0)\log(1-q(x))$$

$\square$

We cannot compute explicitly the cross-entropy, but we have data at our disposal $\{(x_i,y_i)\}_{i=0,\dots,N-1}$. These are realizations of a random sample $\{(X_i,Y_i)\}_{i=0,\dots,N-1}$, i.e., i.i.d. random vectors distributed with $\tilde{p}$. We consider the random variables:

$$-\left(Y_i \log q(X_i; w, b) + (1 - Y_i) \log\left(1 - q(X_i; w, b)\right)\right)$$

and their empirical average:

$$-\frac{1}{N}\sum_{i=0}^{N-1}\left(Y_i \log q(X_i; w, b) + (1 - Y_i) \log\left(1 - q(X_i; w, b)\right)\right)$$

By the law of large numbers, for $N \to +\infty$ the empirical average converges to the mean of the population, which, by the Lemma, is precisely $H(\tilde{p}, \tilde{q})$.

(In fact, it is also an unbiased estimator, since its expectation is $H(\tilde{p}, \tilde{q})$).

This allows us to use

$$-\frac{1}{N}\sum_{i=0}^{N-1}\left(y_i \log (x_i; w, b) + (1 - y_i) \log\left(1 - q(x_i; w, b)\right)\right)$$

as an estimate of the cross-entropy.

Minimizing the empirical cross-entropy is therefore equivalent to minimizing the $-\log$-likelihood of the data.

LOGISTIC REGRESSION (multiclass)

In this situation we have

- a random vector $X : \Omega \to \mathbb{R}^{1 \times M}$ (the features)
- a random label $Y : \Omega \to \{0, 1, \dots, K-1\}$

Again, (X, Y) is distributed according to a probability distribution with probability mass function
$\tilde{p} : \mathbb{R}^M \times \{0, 1, \dots, K-1\} \longrightarrow [0, 1]$,
$\tilde{p}(x, y)$.

We are interested in studying

$$\mathbb{P}(Y = k \mid X = x)$$

The logistic regression model is based on the ansatz:

$$\mathbb{P}(Y = k \mid X = x) \approx q_k(x; \underbrace{W, b}_{\substack{\text{parameters} \\ \text{that we discuss} \\ \text{later}}})$$

Let us now understand the structure of $q_k(x; W, b)$.

In the binary logistic regression, we had to define only the probability of falling in class 1, given by

$$\sigma(\underbrace{xW + b}_{\text{logit}})$$

Now we have classes $0, 1, \dots, K-1$.

We make the ansatz that the logit for class $k = 0, \dots, K-1$ is of the form

$$x w_k + b_k$$

where

- $x \in \mathbb{R}^{1 \times M}$ are features
- $w_k \in \mathbb{R}^{M \times 1}$ are weights, column of $W \in \mathbb{R}^{M \times K}$
- $b_k \in \mathbb{R}$ are biases, entry of $b \in \mathbb{R}^K$

Then we transform $xW_k + b_k$ in a number in $[0,1]$. This must be consistent with the fact that the probabilities sum to one.

$$\sigma_k(xW_k + b_k) = c\, e^{xW_k + b_k}$$

$$\sum_{h=0}^{K-1} c\, e^{xW_h + b_h} = 1 \Rightarrow c = \frac{1}{\sum_{h=0}^{K-1} e^{xW_h + b_h}}$$

It follows that

$$\sigma_k(xW + b) = \frac{e^{xW_k + b_k}}{\sum_{h=0}^{K-1} e^{xW_h + b_h}}$$

We have defined a function $\sigma : \mathbb{R}^K \longrightarrow [0,1]^K$ given by

$$\sigma(a)_k = \sigma_k(a) = \frac{e^{a_k}}{\sum_{h=0}^{K-1} e^{a_h}}$$

This function is called <u>softmax</u> (in fact, softargmax)

As before, we have some observed data $\{(x_i, y_i)\}_{i=0,\dots,N-1}$.

We can compute $-\log$-likelihood of the data:

$$-\log\left(\mathbb{P}\left(\bigcap_{i=0}^{N-1}\{X_i = x_i, Y_i = y_i\}\right)\right) =$$

$$= -\log\left(\prod_{i=0}^{N-1}\mathbb{P}(X_i = x_i, Y_i = y_i)\right) =$$

$$= -\sum_{i=0}^{N-1}\log\left(\mathbb{P}(Y_i = y_i \mid X_i = x_i)\,\mathbb{P}(X_i = x_i)\right)$$

$$= -\sum_{i=0}^{N-1}\log\left(\mathbb{P}(Y_i = y_i \mid X_i = x_i)\right) +$$

$$\underbrace{-\sum_{i=0}^{N-1}\log\left(\mathbb{P}(X_i = x_i)\right)}_{\text{this is independent of the parameters}}$$

Let us study only the first term:

$$-\sum_{i=0}^{N-1} \log\left(\mathbb{P}(Y_i = y_i \mid X_i = x_i)\right)$$

To express the probability more explicitly, it is convenient to use a one-hot encoding for the variable Y.

We define the matrix $\tilde{y} \in \mathbb{R}^{N \times K}$ such that the row $\tilde{y}_{i\cdot} \in \mathbb{R}^{1 \times K}$ is given by $\tilde{y}_{i\cdot} = e_k$ if $y_i = e_k$.

$$-\sum_{i=0}^{N-1} \log\left(\mathbb{P}(Y_i = y_i \mid X_i = x_i)\right) =$$

$$= -\sum_{i=0}^{N-1} \log\left(\mathbb{P}(\tilde{Y}_i = \tilde{y}_{i\cdot} \mid X_i = x_i)\right) =$$

$$= -\sum_{i=0}^{N-1} \log\left(\prod_{k=0}^{K-1} \sigma_k(xW + b)^{\tilde{y}_{ik}}\right)$$

$$= -\sum_{i=0}^{N-1} \sum_{k=0}^{K-1} \tilde{y}_{ik} \log \sigma_k(xW + b)$$

Recall that $w_0 = 0$, $b_0 = 0$, hence the 0 class is not affecting the optimization.

Hence multiclass logistic regression has the objective to find

$$W \in \mathbb{R}^{M \times K}, \quad b \in \mathbb{R}^K$$

such that

$$(W, b) \in \operatorname{argmin} L\left(W, b; \{(x_i, y_i)\}_{i=0,\dots,N-1}\right)$$

where

$$L\left(W, b; \{(x_i, y_i)\}_{i=0,\dots,N-1}\right) = -\sum_{i=0}^{N-1} \sum_{k=1}^{K-1} \underbrace{\tilde{y}_{ik}}_{\text{one-hot encoding}} \log \underbrace{\sigma_k(xW + b)}_{\text{softmax}}$$

Let us interpret this using the cross-entropy.

The data $\{(x_i, y_i)\}_{i=0,\dots,N-1}$ is the realization of a random sample $(X_0, Y_0), \dots, (X_{N-1}, Y_{N-1})$ extracted from a probability distribution with joint probability mass function $\tilde{p}(x,y)$.

The random variables Y_i have range $\{0, 1, \dots, K-1\}$.

Instead of $\tilde{p}(x,y)$, we use a model which induces a joint probability distribution $\tilde{q}(x,y;W,b)$

$$\tilde{q}(x,k;W,b) = \sigma_k(xW + b)\,\mathbb{P}(X = x)$$

The cross-entropy is

$$H(\tilde{p}, \tilde{q}) = -\mathbb{E}_{\tilde{p}}(\log \tilde{q}) =$$

$$= -\sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k) \log \tilde{q}(x,k;W,b)$$

Let us use the notation $\tilde{Y}$ to denote the one-hot encoding of Y.

By $\log\sigma(xW+b)$ we mean the vector $(\log\sigma_k(xW+b))_{k=0,\dots,K-1}$.

<u>Lemma</u>: We have that

$$\mathbb{E}_{\tilde{p}}\Big(-\tilde{Y}\cdot\underbrace{\log\sigma(xW+b)}_{\text{interpreted as a vector}}\Big) = H(\tilde{p},\tilde{q}) - \overbrace{H(p_X)}^{\text{(entropy of }X\text{)}}$$

<u>Proof</u>:

$$\mathbb{E}_{\tilde{p}}\big(-\tilde{Y}\cdot\log\sigma(Wx+b)\big) =$$

$$= -\sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k)\,\tilde{k}\cdot\log\sigma(Wx+b) =$$

$$= -\sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k)\log\sigma_k(Wx+b)$$

$$= -\sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k)\log\frac{\tilde{q}(x,k;W,b)}{\mathbb{P}(X=x)} =$$

...

$$H(\tilde{p},\tilde{q}) = \underbrace{- \sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k) \log \tilde{q}(x,k;W,b)}_{H(\tilde{p},\tilde{q})} +$$

$$+ \underbrace{\sum_x \sum_{k=0}^{K-1} \tilde{p}(x,k) \log \mathbb{P}(X=x)}_{\sum_x \mathbb{P}(X=x) \log \mathbb{P}(X=x) =}$$

$$= - H(p_X)$$ entropy of the distribution X $\square$

By the law of large numbers,

$$-\frac{1}{N} \sum_{i=0}^{N-1} \tilde{Y}_i \cdot \log \sigma(W X_i + b)$$

converges to $H(\tilde{p},\tilde{q}) - H(p_X)$.

The realization on the dataset is

$$-\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} \tilde{y}_{ik} \log \sigma_k(W x + b).$$

We conclude that minimizing the cross-entropy is equivalent to minimizing the log-likelihood.